Queuing Modelling and Performance Analysis of Content Transfer in Information Centric Networks

Submitted by Han Xu, to the University of Exeter

as a thesis for the degree of

Doctor of Philosophy in Computer Science

in August 2022

I certify that all material in this thesis which is not my own work has been identified and that any material that has previously been submitted and approved for the award of a degree by this or any other University has been acknowledged.

Abstract

With the rapid development of multimedia services and wireless technology, new generation of network traffic like short-form video and live streaming have put tremendous pressure on the current network infrastructure. To meet the high bandwidth and low latency needs of this new generation of traffic, the focus of Internet architecture has moved from host-centric end-to-end communication to requester-driven content retrieval. This shift has motivated the development of Information-Centric Networking (ICN), a promising new paradigm for the future Internet. ICN aims to improve information retrieval on the Internet by identifying and routing data using unified names. In-network caching and the use of a pending interest table (PIT) are two key features of ICN that are designed to efficiently handle bulk data dissemination and retrieval, as well as reduce bandwidth consumption.

Performance analysis has been and continues to be key research interests of ICN. This thesis starts with the evaluation of content delivery delays in ICN. The main component of delay is composed of propagation delay, transmission delay,processing delay and queueing delay. To characterize the main components of content delivery delay, queueing network theory has been exploited to coordinate with cache miss rate in modelling the content delivery time in ICN. Moreover, different topologies and network conditions have been taken into account to evaluate the performance of content transfer in ICN.

ICN is intrinsically compatible with wireless networks. To evaluate the performance

of content transfer in wireless networks, an analytical model to evaluate the mean service time based on consumer and provider mobility has been proposed. The accuracy of the analytical model is validated through extensive simulation experiments. Finally, the analytical model is used to evaluate the impact of key metrics, such as the cache size, content size and content popularity on the performance of PIT and content transfer in ICN.

Pending interest table (PIT) is one of the essential components of the ICN forwarding plane, which is responsible for stateful routing in ICN. It also aggregates the same interests to alleviate request flooding and network congestion. The aggregation feature of PIT improves performance of content delivery in ICN. Thus, having an analytical model to characterize the impact of PIT on content delivery time could allow for a more precise evaluation of content transfer performance. In parallel, if the size of the PIT is not properly determined, the interest drop rate may be too high, resulting in a reduction in quality of service for consumers as their requests have to be retransmitted. Furthermore, PIT is a costly resource as it requires to operate at wirespeed in the forwarding plane. Therefore, in order to ensure that interests drop rate less than the requirement, an analytical model of PIT occupancy has been developed to determine the minimum PIT size.

In this thesis, the proposed analytical models are used to efficiently and accurately evaluate the performance of ICN content transfer and investigate the key component of ICN forwarding plane. Leveraging the insights discovered by these analytical models, the minimal PIT size and proper interest timeout can be determined to enhance the performance of ICN. To widen the outcomes achieved in the thesis, several interesting yet challenging research directions are pointed out.

Acknowledgements

First and foremost, I would like to express my sincere gratitude towards my supervisor Prof. Geyong Min for the inspiration and continuous support of my PhD study. I have gained immensely from his integral view on research, his critical thinking, and most importantly, from his hard working attitude. Without his guidance and constant feedback, this PhD would not have been achievable. Apart from his tremendous academic support, he has provided me with a great deal of guidance and assistance in my daily activities. I am really glad to be associated with a great person like Prof. Geyong Min in my life.

It is also my pleasure to express my sincere thanks to my second supervisor, Dr. Jia Hu for the dedicated support he has provided to my research and daily life. We have numerous meetings together to discuss the progress of my research, and I really appreciate his dedication to polishing my papers. Without his guidance, mentoring and knowledge, this study would not have been completed.

I would like to thank Dr. Haozhe Wang for her generous support in the direction of my research. In addition, I want to thank my fellow research colleagues and labmates, Dr. Yulei Wu, Dr. Chunbo Luo, Dr. Wang Miao, Dr. Chengqiang Huang, Dr. Yuan Zuo, Dr. Xiangle Cheng, Dr. Zhengxin Yu, Dr. Jin Wang, Dr. Zeyi Chen, Yujia Zhu, Dr. Fangming Zhong, Dr. Lejun Chen, for always standing by my side and sharing a great relationship as compassionate friends. All of you have enriched my life and been a great strength to me all throughout my PhD pursuit.

Last but not the least, I would like to thank my family, my mother and father, from the bottom of my heart for all their love and encouragement and the sacrifices that they have made on my behalf. They have the kindness and the patience to tolerate my absence from the many responsibilities to complete my PhD study.

Table of Contents

List of Figures

List of Tables

List of Publications

- Xu, H, Wang, H, Hu, J, Min, G, "Analytical Modelling of Content Transfer in Information Centric Networks", in *Proceedings of 2021 IEEE 24th International Conference on Computational Science and Engineering (CSE)*, pp. 64-71, 2021. (Outstanding Paper Award)
- Xu, H, Wang, H, Hu, J, Yu, Z, "Modeling Short-form Video Transfer in Information Centric Network", in *2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS)*, pp. 224-231, 2021.
- Xu, H, Wang, H, Hu, J, Min, G, "Modelling Distribution of Content Delivery Time and Minimal Pending Interest table size in ICN" (To be Submitted)

Chapter 1

Introduction

The proliferation of multimedia services such as short-form video, ultrahigh-definition (UHD) multimedia services and streaming media that demand high bandwidth and ultra-low latency put tremendous pressure on the current communication networks. Instead of focusing on guaranteeing the location of the host, these services are more interested in how to provide the high quality of services. In parallel, users are typically more concerned with how fast and reliable the content can be accessed [3]. Consequently, the Internet is shifting from a host-centric to a content-oriented model. The evolution of the Internet has led to the addition of numerous patches and overlays to the current Internet protocol stack [4], including P2P networks, proxy servers, CDNs, multi-homing, multicast, and mobile IP, among others. However, despite addressing the challenges posed by new applications, the host-centric structure of the Internet has resulted in decreased performance for end-users. This is due to its limited scaling ability, rendering it inadequate for addressing the issues of scalable content distribution, Quality-of-Experience (QoE), and mobility. As a result, the complexity of the Internet has increased, and its ability to provide optimal performance to end-users has diminished. It is necessary to consider alternative approaches to address these limitations and improve the overall performance of the Internet.

To tackle the aforementioned shortcoming and meet the requirements of the future In-

ternet, Information-Centric Network (ICN), an emerging paradigm for the future network has been proposed by research communities [2] to relieve the network pressure arisen from multimedia content transfer. ICN is a receiver-driven networking model, where end-users only need to emit their interests for a given content, then the entire network is in charge of routing the requests based on the content names only. The requests are forwarded towards the best content containers and delivering the contents through the reverse paths to the end-users. It decouples content from location to achieve a naturally content-centric architecture. [5]

The following section of this chapter is structured as follows: In Section 1.1, the motivations and challenges of this research are discussed. Section 1.2 presents the research objectives and significant contributions of this thesis. Finally, the overall structure of this thesis is outlined in Section 1.3.

1.1 Motivations and Challenges

ICN offers a distinct advantage in terms of scalable and highly efficient information retrieval, facilitated by its two key components: the Pending Interest Table (PIT) and in-network caching. In-network caching enables requests to be served by intermediate routers, reducing the load on the original server. The PIT allows for efficient aggregation of requests, mitigating the risk of request flooding. However, despite these advantages, there remain several open research issues that need to be addressed in order to fully realize the potential of ICN. These include performance modeling and resource allocation, which require further investigation. For example

- As ICN is still in its research phase, it is meaningful to have a comprehensive understanding of the factors that impact its transfer performance prior to its widespread adoption in commercial systems. While current evaluations of ICN content transfer

primarily focus on cache performance, specifically considering propagation delay, the total delay experienced by content transfers is comprised of not only propagation delay but also transmission delay and queueing delay. The transmission delay and queueing delay are influenced by the traffic load at intermediate routers and the size of the content being transferred, which presents a challenge in analytically characterizing the traffic load at arbitrary routers. Thus, it is necessary to consider these additional factors in order to accurately assess the transfer performance of ICN.

- Multimedia applications have emerged as major contributors to Internet traffic [6], owing to their widespread usage in the current Internet landscape. The dynamic nature of multimedia applications leads to variations in the generated data streams over time. The traffic generated by multimedia services is known to exhibit bursty behavior [7], which is also reflected in the content request process of these services in ICN due to its receiver-driven nature. However, the constant arrival rate or Poisson process models [8] commonly employed in existing studies on ICN cache performance evaluation are insufficient in capturing the burstiness of multimedia traffic. In light of this, the traffic pattern becomes a critical aspect to consider, and the development of a new analytical model that accurately captures the bursty nature of content requests is essential for a comprehensive evaluation of the performance of content transfer in ICN.

- PIT is a critical component in stateful routing for ICN systems. It enables efficient content transfer by blocking redundant requests and aggregating requests for the same content. However, due to the impact of multiple network metrics, it is challenging to analytically characterize the performance of the PIT in ICN. Despite this difficulty, the evaluation of the PIT performance is crucial for optimizing the content transfer performance in ICN systems. Therefore, it is necessary to develop a comprehensive

and accurate analytical model to evaluate the performance of the PIT in ICN. This will provide valuable insights into the behavior and efficiency of the PIT and enable the optimization of content transfer in ICN systems.

- As the escalation of mobile data traffic continues, it is anticipated that the number of mobile hosts and the volume of their traffic will surpass that of wired counterparts. This increasing trend in mobile traffic highlights the significance of evaluating the performance of content transfer in wireless networks. With the introduction of multimedia, users have the opportunity to serve as content providers, leading to the consideration of both the mobility of consumers and providers in the design of wireless networks. However, research regarding the evaluation of content transfer in ICN wireless networks taking into account PIT influence has yet to be conducted. Hence, the development of an analytical model to examine the performance of content transfer in ICN wireless networks is deemed necessary.

- The function of PIT, in ICN networks, is to track the interfaces of incoming requests, thereby allowing the requested data to be returned to the consumer, and to aggregate requests for the same data in order to prevent request flooding. To ensure the proper operation of the PIT in ICN routers, it is necessary for the PIT to operate at wirespeed in the forwarding plane. However, this requirement makes the PIT a costly resource. For cost-saving purposes, allocating a small size to the PIT may seem reasonable. However, ICN routers with insufficient PIT sizes may result in unacceptably high drop rates for incoming interests, leading to adverse effects on content delivery time due to the need for retransmission of content and increased delay experienced by consumers. To guarantee that the interest drop rate remains below the required level, having an analytical model that determines the minimum size of the PIT is imperative.

- In ICN, the consumer is responsible for retransmitting the interest packet in the event that the corresponding chunk packet is not received within the specified retransmission timeout (RTO) interval. The design of the RTO, however, poses a significant challenge due to the difficulty in accurately evaluating the distribution of the Round Trip Time (RTT). This is because the delay in the network can vary significantly as a result of in-network caching and the use of the Pending Interest Table (PIT), making it difficult to precisely estimate the RTT distribution. Thus, the task of accurately determining the RTT distribution remains a challenge in the field of ICN.

1.2 Research Aims and Objectives

The research work reported in this thesis is focused on the analysis and optimisation of the performance of ICN serving multimedia services under realistic network environments. The main objectives of the research are:

- To develop a new analytical model for content delivery within ICN environments by employing the principles of queueing network theory. The importance of various factors, such as content requests and service popularity, in determining the efficiency and effectiveness of content delivery is emphasized and taken into consideration in the proposed model.

- To develop a new analytical model for characterizing PIT performance and investigate content transfer in ICN under bursty content requests.

- To develop an analytical model to investigate the performance of content transfer in wireless networks of ICN.

- To develop an analytical model of PIT occupancy for determining the minimum PIT size and develop an analytical model of Round Trip Time (RTT) distribution to

accurately predict the retransmission timeout.

To achieve the above objectives, this research develops new analytical models and resource allocation strategies. The effectiveness and accuracy of the proposed models and algorithms are demonstrated by extensive experimental results with various network conditions. The original contributions of this research are summarised as follows:

- A novel analytical model is presented for the evaluation of content transfer performance in ICN under diverse topologies. The model incorporates queuing delays to provide a more precise representation of the content delivery time in ICN. A comprehensive analysis has been performed to examine the content transfer performance of arbitrary network topologies under a Zipf-like content distribution and the application of a Least Recently Used (LRU) caching strategy. The results of this analysis provide valuable insights into the content transfer performance of ICN.

- A comprehensive model for the assessment of the impact of PIT aggregation on content transfer has been proposed. In order to accurately represent the bursty nature of multimedia services, the proposed model incorporates the Markov modulated Poisson process (MMPP) as a method of characterizing the time-varying content request process. The PIT miss ratio, derived within the framework of the model, serves as a crucial metric in the evaluation of content transfer performance. Additionally, the model has been utilized to investigate the effect of chunk transmission window size on the content transfer performance. The results of this study provide valuable insights into the optimization of content transfer in ICN networks.

- An analytical model is presented for chunk transfer in the wireless networks of ICN. The model takes into consideration both consumer and provider mobility, making

it distinct from traditional TCP/IP networks that primarily focus on consumer mobility. The aim is to investigate the feasibility of incorporating mobility features into the architecture of ICN networks, particularly those operating in wireless environments. The results obtained from the proposed model have implications for the design of future ICN networks and can contribute to the understanding of the impact of mobility on the mean service time.

- The Occupancy of PIT has been analyzed through the creation of a mathematical model with the aim of determining the minimum required size of PIT. The model has been utilized to study the distribution of mean chunk service time, which has implications for the design of the retransmission timeout in ICN.

1.3 Outline of the Thesis

The rest of this thesis is organised as follows.

- Chapter 2 introduces the background of ICN including the key features, architectures and mobility. A detailed literature review on the performance modelling is then presented.

- Chapter 3 exploits the queueing network theory to develop a new analytical model for content delivery in ICN considering various content requests and service popularity. The analytical model is used to investigate the impact of new consumers' behaviours on cache miss rate and content delivery time.

- Chapter 4 proposes a novel analytical model of PIT performance to investigate content transfer in ICN under bursty multimedia content requests. In parallel, an analytical model for data transfer in wireless networks is also proposed in this chapter

to investigate the impact of consumer and provider mobility on mean service time in ICN.

- Chapter 5 develops an analytical model of PIT occupancy to determine the minimum PIT size. In parallel, an analytical model of RTT distribution is proposed to accurately predict the retransmission timeout.

- Finally, Chapter 6 concludes the thesis and presents the future research work.

Chapter 2

Background and Literature Review

Since the Internet was designed in 60s-70s, it has played an increasingly important role in the lives of mankind. From its inception, the Internet operated on top of the protocol stack of TCP/IP with the intention of connecting a small number of machines. The Internet information exchange is realized through the establishment of communication channels. This host-centric Internet has perfectly matched the early Internet usage. However, over the past few years, the proliferation of multimedia services has led to the host-oriented communication model, which was designed for special data transmission in the early age of Internet, added more and more overlay and became too heavy. Consequently, ICN is proposed to solve these problems. As a permanent clean-slate approach for the next generation Internet, ICN has attracted much attention from network researchers in the passed few years. Although ICN enters now into the main stream of networking research, it is still in its early stage. Many projects have been carried on in order to propose a concrete ICN solution to deploy it in reality. This chapter gives a general background knowledge and presents an in-depth review of the related research on wired and wireless networks of ICN.

The rest of this chapter is organized as follows. The background knowledge including the ICN architectures with important features is presented in Section 2.1.1. One of

the most promising ICN architectures, Content-Centric Networking (CCN) is introduced in Section 2.1.2. The characteristic of content is detailed in Section 2.1.3. A detailed literature review on modelling of ICN performance is then presented in Section 2.2.

2.1 Information-Centric Networking (ICN)

In order to overcome the mismatch between the current usage patterns and the original design of the Internet paradigm. ICN is proposed to shift the current host-centric communication model to a content-oriented model. To realize this, ICN uses the name of content for routing, which decouples content from location to achieve a naturally content-centric architecture. In this section, the fundamental features of ICN common to most proposed designs are investigated.

2.1.1 Fearture of ICN

There have been several ICN architectures proposed by different research groups. Despite the differences in the specifics of each design, these proposed architectures share several overarching objectives and common components. These common components include location-independent naming, in-network caching, name-based routing, and content security. These key features are central to the realization of a content-centric communication model and ensure that the fundamental principles of ICN are consistently upheld across different architectural designs. It is worth noting that while the specific details of these ICN architectures may differ, they all aim to address the limitations of the current host-centric communication model and offer a more effective solution for the demands of the modern digital landscape.

Naming

The Internet has undergone a significant transformation from an academic network to a worldwide infrastructure for the widespread dissemination of information. Currently, there are over 4 billion connected devices and more than 80% of the traffic consists of video [6], leading to a shift in the Internet paradigm from a host-based to a name-based architecture. In ICN, content naming, used for routing, is location-independent and persistent, unlike IP addresses that are associated with a specific geographical location. Various naming schemes have been proposed to achieve efficient and scalable name look-up for billions of contents. These schemes can be broadly classified into three categories: hierarchical name, flat name, and Attribute-Value-Based Name [9]. Each scheme has its advantages and disadvantages in terms of security and scalability, thus, naming remains an open research issue.

The hierarchical name scheme, similar to IP addresses and Uniform Resource Locator (URL), aggregates prefixes and performs the longest or shortest match. This design enhances the scalability of the routing system, as seen in some proposed ICN architectures such as Content-Centric Networking (CCN) [1]. In Fig. 2.1(a), a content name prefix in a hierarchical format is presented as an illustration. This depiction conforms to a convention commonly used in the organization of information in a systematic manner.

As for the flat name, it is primarily derived from hash algorithms applied to content. There is neither a semantic nor a structure behind a flat name. The generated name is not human-friendly and cannot be assigned to dynamic content that has not yet been published. In addition, flat naming has problems with scalability since it does not support routing aggregation. Fig. 2.1(b) shows a flat name for the hierarchical name in Fig. 2.1(a). This result has been obtained by using the MD2 hashing algorithm.

For the Attribute-Value-Based Name, there are several attributes associated with the

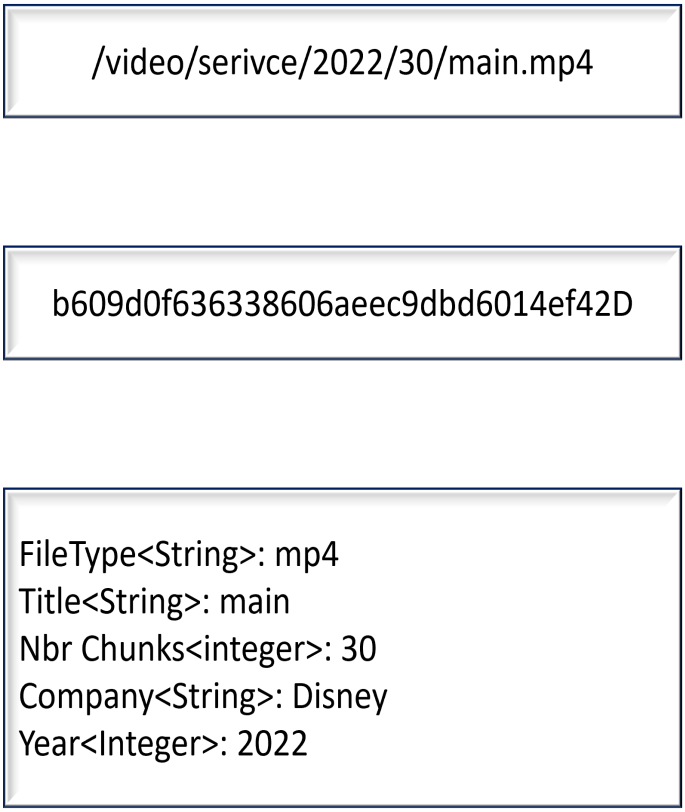


Figure 2.1: ICN naming schemes for the same content

content; each attribute has a name, a type, and a set of possible values. Collectively, they represent a unique content and its properties. This kind of naming scheme is benefit for the searching process by using known content keywords, which means it is quite possible to find many contents against one search request. Hence, it is hard to ensure naming uniqueness and find unique content in a short period. Fig. 2.1(c) illustrates attribute naming scheme's example.

In-network Caching

An important feature that differentiates ICN from the current Internet is the ability to cache data within the network. The routers of the ICN have the capability of temporarily storing copies of the data packets that traverse them. The name of content delivered in ICN is transparent to the router based on the unique names. Therefore, requests can be satisfied by any node holding a copy in its cache within the network, which reduces the

transfer delay. In this way, instead of leveraging any external storage resource such as Content Delivery Networks (CDN) and Peer-to-Peer (P2P) systems, commonly adopted in the current Internet architecture [10], ICN architectures naturally offer a reliable, scalable and application-independent content delivery system. The feature of ubiquitous in-network caching significantly improves content retrieving efficiency and alleviates traffic congestion.

Pending Interest Table

PIT is an essential component for stateful routing in ICN. When the router receives an interest, it first checks its local cache and the PIT to determine whether it holds the corresponding chunk of data. If the requested chunk is not available locally and the PIT does not have a record of the interest, the router forwards the interest to the Forwarding Information Base (FIB) for further processing.

The PIT, in parallel, updates its record by appending the name of the requested chunk and the incoming face of the interest, in a manner similar to the breadcrumb mechanism. This helps in aggregating subsequent interests that request the same chunk and reduces the flooding of interests. The PIT records the incoming interface of each new interest and, if the corresponding chunk is not yet available, it waits for its return.

In the process of being delivered to the provider, the interest may encounter a router that holds the required chunk. In such cases, the router's cache will return the chunk through the reverse path of interest forwarding. Upon arrival of a new chunk at a router, the router checks the chunk name against its PIT. If there is a match, the chunk is transmitted through the interfaces associated with the matched entry.

Routing and Delivery

There are two main approaches to realizing interest routing in ICN: naming resolution-based routing and direct name-based routing [11]. In naming resolution-based routing, one or several centralized servers are utilized. These servers collect information about the publication of content and maintain a global view of the contents within the network. When an ICN router receives an interest, it sends the request to the server, which then resolves the chunk name to identify the closest node that stores a copy of the chunk.

In contrast, direct name-based routing is performed directly by the ICN routers. Each router maintains a local routing FIB that stores information about the publication of content. In this case, the router addresses the interest and forwards the request to one or multiple potential data sources, following its own forwarding strategies.

The chunk retrieval methods in ICN heavily rely on the PIT, which records the path of incoming interests and facilitates the delivery of chunks back to consumers through the reverse path. By leveraging the PIT, the network is able to efficiently route interests and deliver chunks to the intended recipients, thereby improving the overall performance and scalability of the network.

Mobility

The decoupling of content from location and stateless nature of the publish/subscribe paradigm of ICN is expected to facilitate the support of mobility. Consumer mobility is generally easier to handle as it only requires consumers to re-issue pending interests upon mobility.

The provider mobility requires updating the name resolution system with the new location of the mobile provider. When a move of provider takes place, the interest has a high probability to be served by intermediate node. If the interest is not served by the

intermediate cache from consumer to provider, it will have to wait for the notification of the mobile provider's reconnection to the global network.

Security

Traditional host-centric Internet security relies on the protection of the communication channel. As an example, Transport Layer Security (TLS) is used to establish end-to-end connectivity, and synchronized three-way authentication is performed by an authenticated server. Contrarily, ICN adopts a content-based security concept by applying security mechanisms to the content itself and using selfcertification. [12]. Although such a technique can improve the security of a network, knowing the semantics of the name as well as using plain text may cause different issues. Leshov et al. [13] proposed a naming scheme for the NDN architecture in order to improve the security of content naming in sensitive applications. The main idea is that the requester uses a symmetric key to encrypt a part of the chunk name in the Interest packet. This allows a secure information exchange (the Interest and Data packets) only between trusted nodes that are in the path while hiding the real content name. In parallel, intermediate nodes utilize a three-way handshake mechanism to transfer the key for decrypting the real chunk name.

2.1.2 Content Centric Network

Content-Centric Networking (CCN) [1], designed by Palo Alto Research Centre (PARC), is one of the most well-known ICN paradigms. CCN proposes a clean-slate Internet architecture considering the naming structure and forwarding functionalities. CCN also elaborates the mobility, security and content dissemination issues. The design of CCN mainly focuses on the network layer and transport layer of the Open Systems Interconnection model (OSI). CCN introducing the naming of content to replace the IP address for content dissemination. The research work described in this thesis is mainly focused on the

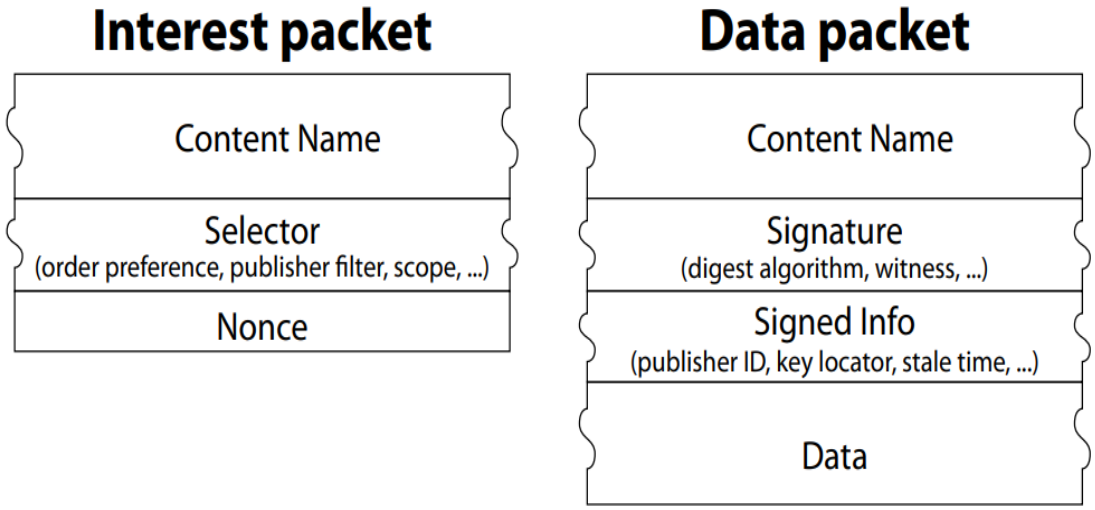


Figure 2.2: CCN packet types. (Source: [1])

architecture proposed in CCN, but the results can broadly apply in the context of general ICN architectures and other caching-related work. In this section, the CCN architecture is described.

The forwarding process of CCN is driven by types of packets : Interest and Data. The structure of the two types of packets is shown in Fig. 2.2. Interest packets sent by consumers to request the named contents. Every Interest packet contains a hierarchical structure content name, similar to a URL, with the prefix indicating the global and organizational routing information, and a suffix illustrating the details of version and segmentation. Data packet, which also refers to as chunk packet, contains a unit of data of the request content, is transmitted in response to the matching interest.

The Content Store (CS) is the cache of the router. In general, the purpose of CS is to cache chunks that are likely to be requested in the near future. When new interest arrives, the CS will check whether a corresponding chunk of interest is stored. If it is, the CCN router will deliver back that chunk from the CS.

The PIT is a new structure proposed by CCN. PIT carries out two functions: one is recording the incoming interface of the interest if the interest is not served by the cache of

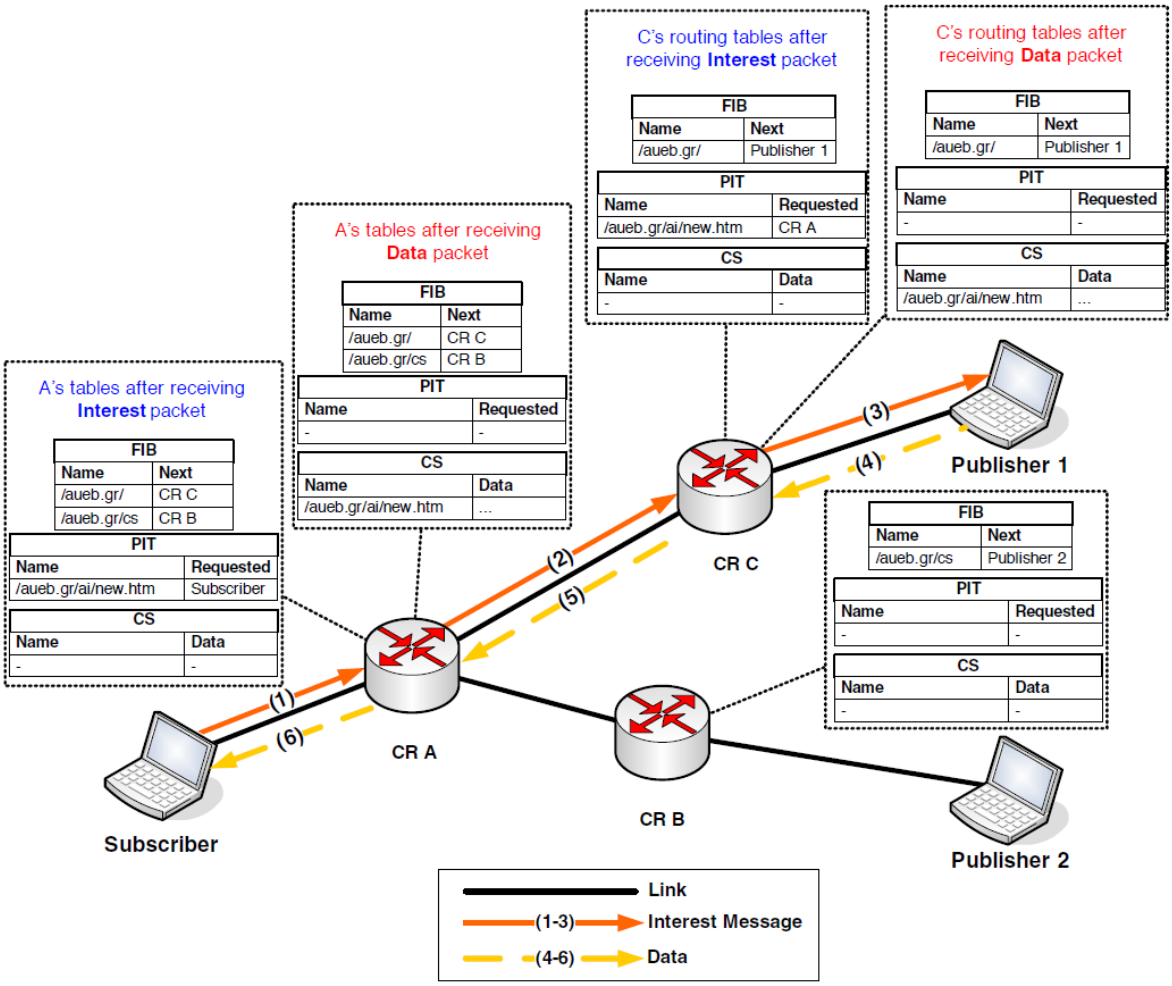


Figure 2.3: The key components of CCN router. (Source: [2])

router, and, similarly to the breadcrumb, assists in returning the chunk back through the reverse path of the interest. Another function is aggregating the same interests before the return of the chunk.

In CCN, content names replace the IP addresses for routing. The FIB in CCN maps the names of content to the output interface(s) that should be used to forward Interest towards appropriated router(s). This router may have the right chunk or knowledge about how to propagate the interest to potential data sources. As shown in Fig. 2.3, content store, PIT and FIB are the three main parts of the CCN router.

The simple working flow of CCN is illustrated in Fig. 2.4. A unique identifier (name) must be assigned to all content in ICN, which is used by consumers for retrieval. An

interest with a unique name is generated by consumers to request content. Producers provide the corresponding Data in response. As soon as an interest is received by a router, the router looks for the content in its own cache called the Content Store (CS). In case of a hit, the router replies with the corresponding data. If a miss occurs, the router looks up the PIT which stores the incoming interface of interest that has not yet been served. When a request for the same content already exists in the PIT, the entry in the PIT is updated to record the interface of the new interest, but the new request is not forwarded by the router. The router adds an entry into the PIT for the interest if a match cannot be found in the PIT. After that, the router refers to the FIB, which maps each name prefix to the next hop router, and finally forwards the interest to the next hop router.

For data packet routing, when a packet arrives at the CCN router via the reverse path of the interest, it will first be delivered to the PIT for checking which interface(s) it should be forwarded to. The data packet will be dropped if there is no match in the PIT. Otherwise, the data packet will be sent to the content store. In accordance with the caching replacement policy, the data packet has the possibility of being stored in the content store. Following this, the data packet will be sent to the next router via the interface recorded in the PIT.

2.1.3 Content characteristics

In this section, the key characteristics of content that are important for content transfer performance of ICN are discussed.

Types of content: The Cisco Visual Networking Index published in 2019 classifies Internet traffic and forecasts global demand for the period 2017-2022 [6]. Cisco forecast that in 2022, nearly 98% of traffic will be content retrieval, classified as web, email and data, gaming and Internet video. Moreover, Internet video streaming and downloads are

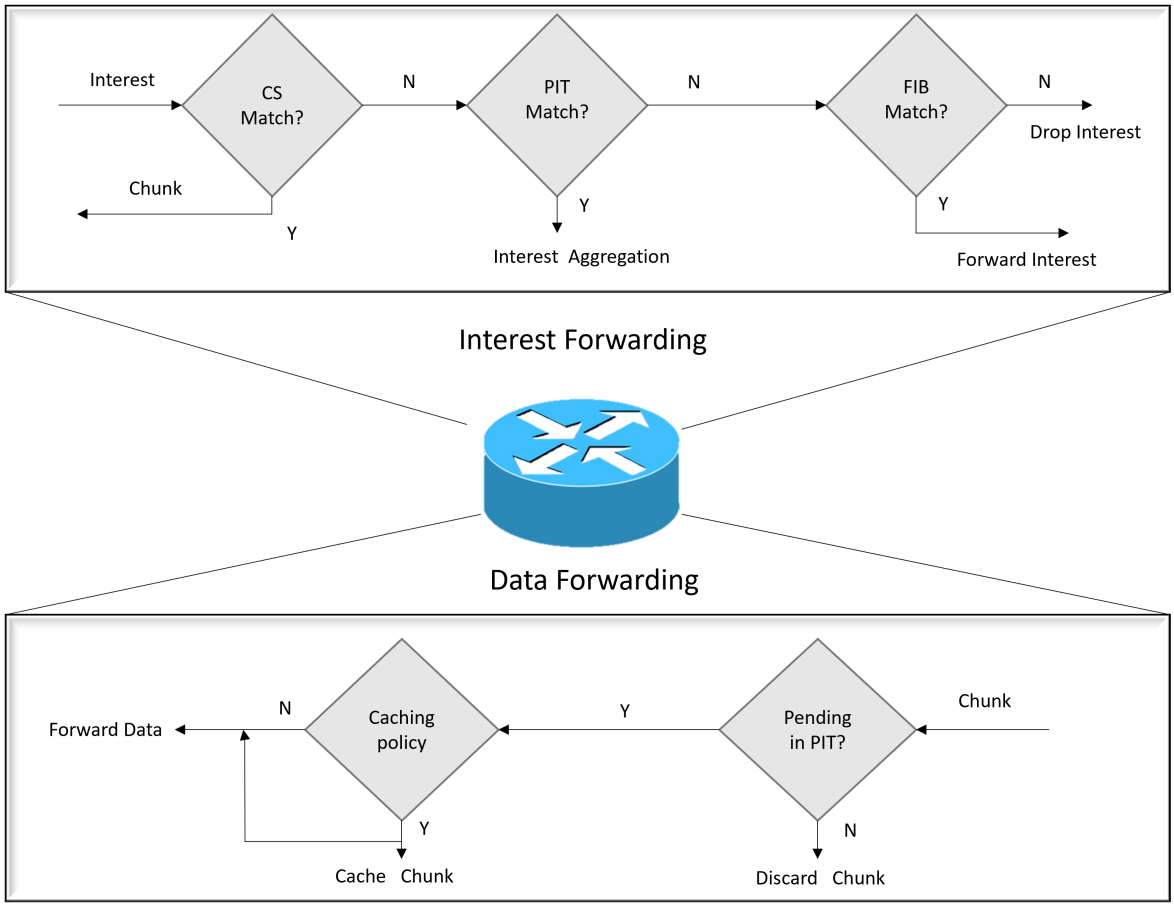


Figure 2.4: The forwarding flow of Interest/data

beginning to take a larger share of bandwidth and will grow to more than 82 percent of all consumer Internet traffic in 2022. Internet video can be further divided into short-form Internet video (for example, YouTube, video on Facebook and Twitter), long-form Internet video (for example, Hulu), live Internet video, Over-the-top video (for example, Netflix, HBO, Amazon Video), online video purchases and rentals, webcam viewing, and web-based video monitoring (excludes P2P video file downloads).

Catalog and content size refers to the number of individual contents in a network. According to [14], there are at least 4.33 billion indexed web pages in the world. Due to the fact that ICN identifies contents by their global unified names at a chunk level, the number of names object in ICN is estimated at 10^{23} orders of magnitude [15].

Performance and efficiency of caching policies may be affected by deployment and scalability concerns [16]. User Generated Content (UGC) is dominated by YouTube. A

recent study by Zhou et al. estimated that there are currently 5×10^8 YouTube videos of mean size 10 MB [17].

Popularity distribution Distributions of content popularity play a crucial role in determining the performance of content transfer. [18] highlights that the popularity of Youtube content can be described by Zipf-like laws: the request rate $q(n)$ for the n^{th} most popular page is proportional to $\frac{1}{n^{\alpha}}$. Many caching strategies are based on content popularity distribution, such as LRU, LFU and more sophisticated variants derived from them.

2.2 Literature Review

In recent years, a considerable amount of research has been dedicated to the field of ICN. This literature survey focuses on the studies addressing the modeling of content transfer in ICN. These models aim to analyze the behavior of content delivery and identify the key factors that influence the efficiency of the transfer process.

Similarly, the issue of PIT occupancy has been widely studied in the context of ICN. The PIT is a key component in ICN and its efficient management is crucial for ensuring the performance and scalability of the network. Studies have aimed to address the issue of PIT occupancy by proposing various solutions, such as optimization algorithms, efficient data structures, and proactive cache management. Furthermore, the contribution of this research will be discussed.

2.2.1 Content transfer performance modelling of ICN

In-network caching plays a crucial role in evaluating the performance of ICN and as a result, a substantial body of research has been devoted to studying the behavior and features of in-network caching. [19–31]. The research conducted on ICN cache have provided valuable insights into the intrinsic behavior and performance of these networks.

In particular, these studies have shed light on the impact of in-network caching on the efficiency of content transfer, making it an indispensable factor in the evaluation of the performance of content transfer in ICN.

In the realm of transport in ICN, numerous proposals have been put forth, including HR-ICP [32], ICTP [33] , ECP [34] , CCTCP [35], and HIS [36]. However, there is limited research on transport models in ICN, and these protocols often lack consideration of the impact of in-network caching on congestion control algorithms. This is a significant omission, as in-network caching is a hallmark feature of ICN and has a major influence on content delivery performance.

Despite this, much work has been done on caching in ICN, focusing mainly on cache deployment and sharing mechanisms, cache decision policies, and cache replacement algorithms. However, there is a scarcity of studies on the analytical modeling of cache performance in ICN. Furthermore, there is limited research on the interaction between caching and transport in ICN, particularly with regards to modeling.

In the research of content transfer performance. The main component of delay is composed of propagation delay, transmission delay, queueing delay. The cache performance is a key factor in evaluating the propagation delay. Gallo et al. [37] derived a mathematical model to estimate the cache hit rate based on random cache replacement policy (RND). Gallo's work serve as the foundation for the modelling of ICN performance.

The model proposed by Gallo enables the prediction of cache hit rates, which is essential for evaluating the performance of content delivery in ICN. As Gallo's work builds a stepping stone in performance evaluation, many work has been done to model various aspects of ICN [38–44], which also includes the content transfer performance [8, 45–48]. In [48], author not only analyze the miss probability, but also derive expressions for the average content delivery time. However, they only consider the propagation delay, which short

of considering the impact of transmission and queueing delay on content delivery time. In [45], author both consider the propagation delay and transmission delay. However, in the process of calculating the content delivery time, they did not accumulate the cache miss rate. In parallel, they model the content delivery time in simple topology and did not consider a realistic topology, thus ignored the correlations among ICN nodes. In contrast to the linear and tree topologies, the realistic ICN topology of cache networks should be represented by arbitrary graphs [43] due to the mobility and on-path caching features of ICN. In light of this, to fill up the gap. We design a comprehensive model which takes into account the effects of cache hits and queueing delays on the performance of content delivery under arbitrary topology.

2.2.2 PIT and content transfer performance modelling of ICN in wired and wireless network

Rapid developments in mobile technologies and multimedia services promote the development of ICN. In the past decade, it has become a research hotspot in the field of future Internet architecture [49]. Several features such as routing by name, network caching, and natural mobility make ICN ideal for the transfer and distribution of multimedia content. In recent research, more and more researchers [22, 50–55] believe that ICN can be served as a promising alternative to solve the dilemma that the current TCP/IP network can not cope with the traffic of live-streaming and video content transfer.

There have been extensive studies on caching mechanisms at the application level, primarily in relation to web applications. Analysis of current traffic patterns may shed further light on the popularity characteristics of information today and the potential benefits that can be gained from the widespread use of caching. Recently, a study has shown that the popularity of web information has changed over the past few years, affecting

application-level caching performance [56]. In ICN, caching has been brought to network level. Cache space management therefore becomes crucial for the ICN. In today's service-oriented network, different types of multimedia traffic will compete for the same caching space, therefore the simplified traffic models such as Poisson is not enough to accurately quantify the characteristics of content requests in ICN. Most of the existing in-network cache studies consider the simplified traffic models such as constant arrival rate and Poisson process [45, 57, 58], which fail to capture the bursty nature of content requests in ICN.

Among the existing research work on performance of ICN, only a few have considered the transport model of ICN [8, 48]. In [48], author provides the analytical model of the cache miss rate, which help to characterize the propagation of the content transfer. In [45], the author attempts to take queueing delay into account to provide a more accurate estimate of content transfer performance. PIT, however, is an equally important component of ICN content forwarding that yet to receive proper consideration in current research. During the time between the issuance of an interest and the receipt of the corresponding chunk, the PIT is capable of aggregating identical interests, thereby reducing the incidence of interest flooding and shortening the delivery time. To address the gap in understanding the role of the PIT and its impact on content delivery, Through a thorough examination of the aggregation function of the PIT, this study presents an analytical method for evaluating the performance of the PIT in content delivery time.

Due to the decoupling of time and space between request resolution and data transfer, the publish/subscribe communication model used in the context of ICN fundamentally realizes the seamless mobility of mobile nodes (MNs) [59, 60]. In light of this, the mobility issue has drawn increasing attention [2, 61, 62]. Several approaches have recently been proposed to address the issue of mobility in ICN from the perspectives of consumers

and providers. The approaches to enable consumer mobility in ICN mainly include architectural mobility support [2, 61, 62], fast request retransmission recovery [63–65] and proactive caching [66–68]. Indirection [69–71] and routingbased [72] techniques are considered to support publisher mobility in ICN. However, ICN research currently lacks a model to assess the impact of consumer and provider mobility on content transfer performance. Therefore, we are motivated to develop an analytical model to investigate the performance of content transfer in wireless network.

In this work, instead of the simple traffic assumption, we take into account the bursty traffic arisen by multimedia services, heterogeneous content popularity distribution and PIT aggregation to investigate content transfer performance in wired and wireless networks of ICN. To the best of our knowledge, this is the first attempt to evaluate the impact of PIT on content transfer performance.

2.2.3 PIT Occupancy modelling of ICN

In the literature, the sizing of router buffers for TCP/IP networks has been extensively explored [73, 74]. In ICN routers, the PIT can be viewed as a buffer that stores information about pending interests. Upon receipt of a data packet, the PIT is checked. It is essential that these operations be performed at wire-speed. Because of the trade-off between the cost of implementing PIT at wire-speed and the required performance value, finding the minimum size of PIT is an important issue. Authors in [75–79] discuss the problem of wire-speed implementation of PIT.

Perino et al. [80] systematically evaluate the usability of the existing router components for the support of Content Centric Networks. This study examines the three main components of ICN routers, namely, the FIB, the CS, and the PIT. Taking into account parameters such as interest arrival and data response rate, they present a primitive model

for different metrics such as PIT hit probability and PIT miss probability. By evaluating these metrics, the authors determine the minimum size required for the PIT in the worst case scenario. Contrary to our work, the authors do not model the occupancy of the PIT analytically.

Abu et al. [81] develop a Markov model for PIT occupancy with interest timeouts and retransmissions. The authors assume that interest arrivals follow a Poisson distribution and interest service time follows an exponential distribution. Using these assumptions, they model the interest drop probability based on a two-dimensional continuous time Markov chain. While their model takes into account various factors associated with ICN, such as caching in-network and interest drop probabilities, their work only considers that interest arrivals follow a Poisson distribution.

In ICN, a consumer retransmits the Interest packet if the corresponding Data packet is not received within a retransmission timeout (RTO) interval [65]. Like the calculation method of RTO in IP networks, it relies on the predicted Round Trip Time (RTT), which is the time interval between the transmission of the Interest packet and arrival of the corresponded content [65]. However, because of in-network caching and PIT aggregation, the transmission delay can vary significantly, which makes it harder to exactly estimate RTT [63]. As of the present, the current work only examines the average round-trip time, which is not sufficient to accurately design the RTO. For the purpose of filling this gap, we evaluate the distribution of mean service time of chunks to support the design of RTO in ICN.

Although authors in [81] address the problem of minimum PIT size by analytically modeling PIT occupancy taking interest drop probability into account, our proposed model models PIT occupancy as a queue for more accurately evaluating the PIT miss rate. In parallel, our developed model can also be extended to evaluate the distribution of mean

service time of chunks.

2.3 Summary

In this chapter, the background knowledge of ICN, including the key features and existing architectures have been investigated. Furthermore, the important components and work flow of the most well-known ICN paradigm, CCN, has been presented. More specifically, the key characteristics of content that are important for content transfer performance of ICN are discussed. At last, a comprehensive literature review of state-of-the-art content transfer, PIT miss rate and PIT occupancy modelling has been presented. Building upon the previous research efforts, this work propose a model for analyzing the delay of file transfers in ICN under various levels of popularity. To further understand the impact of network load on transfer efficiency, the queueing network model has been employed. Our aim is to provide a comprehensive analysis of the interplay between network load and file transfer delay in ICN. The study presents a thorough examination of the impact of the PIT on transfer efficiency in ICN systems. Additionally, the effect of consumer and provider mobility on file transfers in wireless ICN networks has been analyzed. Given the costly nature of PIT, the study optimizes its size to facilitate the effective deployment of ICN networks. The results of the study offer valuable insights into the design of retransmission mechanisms in ICN networks, particularly with regard to the delay distribution of file transfers under varying levels of popularity. The findings provide a deeper understanding of the complex interplay between the various factors that influence transfer efficiency in ICN systems.

Chapter 3

Analytical Modelling of Content Transfer in Information Centric Networks

3.1 Introduction

With the rapid development of multimedia services, the Internet has invaded all aspects of human life. The novel services such as Augmented Reality (AR), Virtual Reality (VR) and ultra-high-definition (UHD) put tremendous pressure on the bandwidth and latency of the network architecture. Therefore, latency becomes a crucial factor in evaluating the performance of content transfer. In particular, the high transfer efficiency is a key advantage of ICN over the traditional TCP/IP architecture. The main component of delay composed of propagation delay, transmission delay, queueing delay and processing delay. The processing delay, which is characterized by the time required for intermediate routers to process the data packet, encompasses tasks such as header checksum calculations, the updating of Time-to-Live (TTL) value, and the determination of the forwarding path. The processing delay is dependent on the speed of the router's processor. With high-speed

routers typically exhibiting processing delays of microseconds or less, in this study, the processing delay has been excluded. However, it should be noted that, if deemed to be indispensable, the processing delay could be incorporated into the service time of the queueing network in our model.

In-network caching is an integral part of ICN [21] that contributes to the efficient acquisition of content, reducing the network load, and improving the Quality of Experience (QoE) of users. Transport performance is affected by caching dynamics. There may be a situation in which routers lack enough computing resources to handle traffic at wirespeed as a result of the increasing load of Internet traffic. A high level of traffic leads to chunks joining the queue of the waiting list at the router, awaiting the sending of prior chunks. In light of this, queueing delay is also a significant factor of delay that should be quantified.

A unified modeling framework is necessary to evaluate data transfer performance as well as to guide the design of optimized CCN protocols. The weakness of the existing transfer performance studies of ICN is short of consideration of the impact of queueing delay and transmission delay. To fill the gap, this chapter aims to investigate the performance issues of content transfer in ICN. To this end, a new analytical model is developed as a cost-effective performance tool to investigate the content transfer of ICN, especially the cache miss rate and average waiting time at routers. Two scenarios are considered in the analytical model, namely the model for the tree network and the model for the new consumer at the intermediate router. The developed analytical model adopts the Markov Modulated Rate Process (MMRP) as the input process. The Least-Recently-Used (LRU) replacement policy is taken into account because it has been successfully implemented in many caching systems [44]. Extensive simulation experiments are conducted to validate the accuracy of the analytical model. In addition, the analytical model is used to evaluate the impact of key metrics, such as cache size, content size, and content popularity, on the

performance of content transfer in ICN.

The remainder of this chapter is organized as follows. Section 3.2 presents the system parameters and the model of the input process, and then derives the analytical model for content transfer of ICN. Section 3.3 validates the analytical model and carries out the performance analysis. Finally, Section 3.4 concludes the chapter.

3.2 Analytical Model of Content Delivery

In this section, an analytical model is proposed to investigate the performance of content transfer in ICN under single consumer and network of consumers scenarios.

We mentioned the main component of delay is composed of propagation delay, transmission delay, queueing delay and processing delay. Processing delays are determined by the processing efficiency of ICN routers, and currently there is no concrete standard for this measurement. Additionally, the simulation does not conduct the processing time in parallel. Therefore, our model omitted the processing time, however it could be used in a scenario where processing time is considered by adjusting our queueing network model.

To derive the expression of content transfer, the system parameters will be introduced first.

3.2.1 System Parameters

The system parameters used to derive the analytical model are introduced in this subsection, followed by the process of chunk request arrived represented. The system investigates

1) A total of M different contents equally partitioned in K classes. Each class is considered as one kind of services that contains $m = M/K$ different contents. Futhermore, Contents belonging to class are requested with the identical probability q_k following

a Zipf distribution, where $q_k = d/k^{\alpha}$ with parameter $d \geq 1$. The Zipf distribution has been extensively utilized to quantify the popularity of Internet content [44, 82, 83], and [84] has conducted a comprehensive study and discovered that the popularity of video content on two major platforms, BiliBili and YouTube, exhibits a Zipf distribution. This finding highlights the importance and wide-spread application of the Zipf distribution in characterizing content popularity on digital media platforms.

2) The content is segmented into chunks and the chunk size is set at L bytes. The distribution of content size follows a geometrically distributed model with an average of δ. The CS of each CCN router runs the LRU cache replacement policy, with a cache size of router i is c_i chunks. The LRU policy with low complexity has been used in [85–87]. Moreover, the caching operations of LRU can be implemented at line speed, which is one of the requirements of CCN.

The arrival chunk request rate for a given content is modelled by Markov Modulated Rate Process (MMRP) [88], in which the emission of content requests of class k follows the Poisson distribution with intensity $\lambda_k = \lambda q_k$, where λ is the requests rate of all types of contents. Contents in the given class have the same possibility to be requested. The process of content transfer in ICN starts from sending the first requesting packet of a content called interest. Once the chunk request has been matched and the corresponding chunk has been returned to the consumer, the next chunk request is generated. The cycle repeats until the last chunk of the content is received or being terminated by the user. To facilitate reading, the notations of system parameters are summarised in Table 1.

3.2.2 Queueing network model

In this subsection, we present a queueing network model for content transfer in ICN and derive an closed-form expression for the average waiting time of a chunk in the queueing

Table 3.1: SYSTEM PARAMETERS INVESTIGATED IN CHAPTER 3

Parameter	Meaning
N	Number of ICN routers in the network
M	Number of different contents served in the network
m	Number of different contents with same popularity
K	Number of different popularity classes
δ	Average content size in number of chunks
α	Zipf exponent
c_i	Cache size of node i in number of chunks
N_i	The size of queue in number of chunks at node i
$B(i)$	Link bandwidth from node i to node $i-1$
λ,$\lambda(i)$	Content request rate at at node 1,i
q_k,$q_k(i)$	Popularity distribution for class k at node 1, i
$p_k(i)$	Miss probability for class k at node i
$T_N(i)$	Queueing delay for one chunk experienced at node i
VRTT_k	Virtual round trip delay of calss k

network.

To illustrate our model from an intuitive perspective, we consider our model in a linear topology as shown in Fig. 3.1. The network comprised of N nodes, numbered from 1 to N, forming a linear path from the consumer to the repository. A request of a chunk, named interest, yielded from the consumer node is transmitted hop by hop to the repository or the cache that contains a temporary copy of the data chunk along the path. Once the request is satisfied, the chunk flows down to the consumer following the reverse path of the request. In this process, buffer resource management, i.e. queues of packets organised on the router's interfaces, is one of the essential components of content transfer. Our queueing network system is composed of several separate router interface queues, the size of each interface queue depends on the amount of traffic that is transmitted to the interface per unit of time. Compared to the chunk size, the interest packet is quite small. Therefore, in our queueing network model, we focus on the return traffic of chunks within the interface queues. The process by which a chunk is sent back to the consumer moves along a tandem queue, where the chunk coming out of each queue is fed into the next queue in the chain. As a request not only can be satisfied at the repository but also can be hit at the cache of

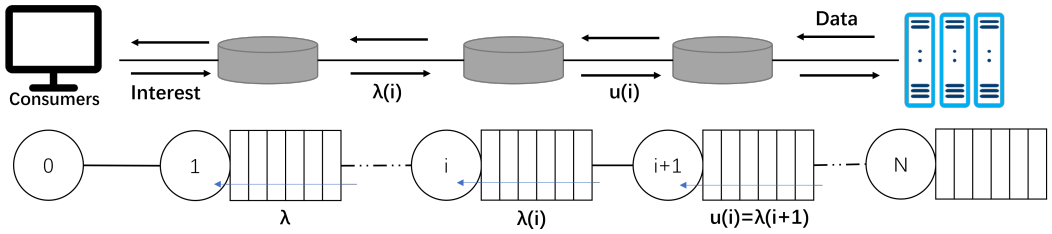


Figure 3.1: Basic network topology

any on-path router, the number of queues that the data passed is determined by where the request is satisfied. Considering the chunks output from interface queue at node i only fed into node $i-1$, Our queueing network model can analyse through the following variables.

We define $\lambda(i)$ to indicate the arrival rate of interests at node i. Among those interests that arrive at node i, some of them are hit at the cache of node i, while the remaining interests may hit at the subsequent nodes or finally be satisfied at the repository. Based on the feature of PIT, ICN inherently route the content on the interest?s reverse path through recording the interface of the coming interest. For those interests that have reached node i, all of their corresponding chunks will be passed back to the node i. Consequently, the chunk arrival rate at different interfaces of node i is equal to the arrival interest rate from each interface. In the linear topology, interests only come from one interface. Thus, the chunk arriving rate at node i is equal to the output interest rate at node $i-1$. We have, $\lambda(i) = \sigma(i-1)$, where $\sigma(i-1)$ denotes the cache miss rate at node $i-1$. The expression for calculating $\lambda(i)$ will be given in the following subsection.

Let $\mu(i)$ denote the service rate at node i, which is determined by the chunk size and link bandwidth, as $\mu(i) = B(i)/L$. The utilisation of the interface queue from node i to $i-1$ is $\rho(i) = \sigma(i-1)/\mu(i)$. Since the chunk size in our model is fixed, the queue of each interface could be modelled as a M/D/1/N queue with a finite capacity N. We extend the waiting time of a M/D/1/N queue [89], and derive the average service time for one chunk

experienced at node i as

$$T_N(i) = \frac{1}{\sigma(i-1)} \frac{1+\rho(i)b_{N_i-1}}{b_{N_i-1}} \frac{N_i + N_i\rho(i)b_{N_i-1} - \sum_{n=0}^{N_i-1} b_n}{1+\rho(i)b_{N_i-1}}. \tag{3.1}$$

where N_i represents the capacity of the queueing system at node i. Letting $H_k(t) = ((\lambda t)^k/k!)e^{-\lambda t}$, the coefficients b_n could be derived as [89]:

$$b_n = \sum_{k=0}^{n} H_k((k-n)T), \qquad \forall n \geq 0. \tag{3.2}$$

where T denotes the average service time in the system, which is $T = 1/\mu(i)$ in our model. As ICN inherently routes the chunk on the reverse path of the interest by recording the interface of the coming interest, the chunk arrival rate at interface of node i is equal to the interest arrival rate at interface of node i. Therefore, the chunk arrvial rate at node i is represented by $\lambda(i)$ chunks per second, with $\lambda(N)$ being the interest arrival rate at the provider. The average waiting time of a chunk in the queueing network T_{avg} can be obtained by:

$$T_{avg} = \sum_{1}^{N} \frac{\lambda(i)T_N(i)}{\lambda}. \tag{3.3}$$

where λ represents the arrival of content request at node 1. The average delivery time for chunks in the queueing network system is primarily determined by the content request rate $\lambda(i)$ at each node, and we will quantify the content request rate through cache miss rate in the next subsection.

By taking processing delay into consideration, the waiting time of a content in the router increases. The service rate should incorporate processing time in our model in order to evaluate the impact of processing delay on the performance of content trans-

fer. Therefore, in this scenario, service time is composed of both transmission delay and processing delay.

3.2.3 Single consumer model

In this subsection, we derive an expression for the average delivery time for content items in popularity class k under single consumer.

Cache miss rate characterization

Assuming the request arrival process is modelled through MMRP with class popularity follows Zipf distribution ($q_k = d/k^{\alpha}$) and cache implements the LRU replacement policy, the cache miss rate at chunk level can be expressed as $p_k = e^{-\frac{\lambda}{m} q_k g c_i^{\alpha}}$, where $1/g = \lambda d \delta^{\alpha} m^{\alpha-1} \Gamma(1-\frac{1}{\alpha})^{\alpha}$ and $g(i)/g = \sigma(i-1)/\lambda(i)$.

The proof and accuracy of the property can be found in [58]. The following is a brief description of the approximation methodology.

Let R^i_{jk} denotes the number of requests for chunk i of content j in class k in an open interval (u,t), and $R^i_{jk} \sim Poisson(\lambda q_k/m)$. In parallel, $B^i_{jk} = 1\{R^i_{jk}(u,t) > 0\}$ is the Bernoulli variable, which means that at least one chunk i in class k is requested in the open interval (u,t). Based on the probability mass function of Poisson distribution, the probability can be written as

$$\mathbb{P}(B^i_{jk}(u,t) = 1) = 1 - e^{-\frac{\lambda q_k}{m}(t-u)}. \tag{3.4}$$

$S(u,t)$ indicates the number of different chunks that are requested in the interval (u,t) over all popularities.

$$S(u,t) = \sum_{k=1}^{K} \sum_{j=1}^{m_k} \sum_{i=1}^{\delta} B^i_{jk}(u,t). \tag{3.5}$$

To derive the mean arrival rate of new chunk requests between the interval $(0,t)$, the number of different chunks that are requested in the interval (0,t), $\lim_{t\to\infty}\mathbb{E}[S(0,t)]$, should be first addressed. The lower bound of $\lim_{t\to\infty}\mathbb{E}[S(0,t)]$ is computed as $K\to\infty$

$$\begin{aligned}
\mathbb{E}[S(0,t)] &= \sum_{k=1}^{\infty}\sum_{j=1}^{m_k}\mathbb{E}[\sum_{i=1}^{\delta}B^i_{jk}(0,t)] \\
&= m\delta\sum_{k=1}^{\infty}\int_k^{k+1}(1-e^{-\frac{\lambda q_k}{m}t})du \\
&\geq m\delta\sum_{k=1}^{\infty}\int_k^{k+1}(1-e^{-\frac{\lambda q_u}{m}t})du \\
&= m\delta\int_0^{k+1}(1-e^{-\frac{\lambda}{m}\frac{d}{u^\alpha}t})du \\
&= (\frac{\lambda dt}{m})^{\frac{1}{\alpha}}\frac{m\delta}{\alpha}\int_0^{\frac{\lambda dt}{m}}v^{1-\frac{1}{\alpha}}e^v dv \\
&\sim \Gamma(1-\frac{1}{\alpha})m\delta(\frac{\lambda dt}{m})^{\frac{1}{\alpha}}, \qquad t\to\infty.
\end{aligned} \tag{3.6}$$

Similarly, the upper bound can be derived as,

$$\begin{aligned}
\mathbb{E}[S(0,t)] &= m\delta\sum_{k=1}^{\infty}\int_k^{k+1}(1-e^{-\frac{\lambda q_k}{m}t}du) \\
&\leq m\delta(1-e^{-\frac{\lambda}{m}dt})+m\delta\sum_{k=1}^{\infty}\int_k^{k+1}(1-e^{-\frac{\lambda q_u}{m}t})du \\
&= m\delta(1-e^{-\frac{\lambda}{m}dt})+m\delta\int_0^{k+1}(1-e^{-\frac{\lambda}{m}\frac{d}{u^\alpha}t})du \\
&= m\delta(1-e^{-\frac{\lambda}{m}dt})+(\frac{\lambda dt}{m})^{\frac{1}{\alpha}}\frac{m\delta}{\alpha}\int_0^{\frac{\lambda dt}{m}}v^{1-\frac{1}{\alpha}}e^v dv \\
&\sim m\delta+\Gamma(1-\frac{1}{\alpha})m\delta(\frac{\lambda dt}{m})^{\frac{1}{\alpha}}, \qquad t\to\infty.
\end{aligned} \tag{3.7}$$

Assuming that c_i is the cache size in number of chunks, then the interval T_c between two consecutive misses occurs after more than c_i different requests have been sent to the router within the interval. thus, T_c can be written as

$$
\begin{aligned}
T_c &= \frac{c_i}{\mathbb{E}[S(0,t)]/t} \\
&\approx \frac{c_i^{\alpha}}{\mathbb{E}[S(0,t)]^{\alpha}} t, \qquad t \to \infty.
\end{aligned}
\tag{3.8}
$$

As a consequence, both the upper and lower bounds of the mean arrival rate of chunk requests, $\frac{1}{g}$, coincide with

$$
\begin{aligned}
\frac{1}{g} &= \lim_{t\to\infty} \frac{\mathbb{E}[S(0,t)]^{\alpha}}{t} \\
&= \lambda d \delta^{\alpha} m^{\alpha-1} \Gamma(1 - \frac{1}{\alpha})^{\alpha}.
\end{aligned}
\tag{3.9}
$$

After g is derived, the minimal interval of two consecutive misses occur can be determined. A cache miss occurs when more than C different chunks are requested after the previous request for a chunk. Considering that the number of arrival chunks is greater than the cache size, C, the given chunk has been removed from the cache prior to the arrival of the new request. Thus, the minimal interval of two consecutive misses can be written as

$$
T_c = g c_i^{\alpha}. \tag{3.10}
$$

Let p_k and q_k indicate the miss probability and probability of requests for class k at the first cache node respectively. Here, we use $p_k(i)$ and $q_k(i)$ to represent the miss probability and popularity distribution at node i. Having $p_k(1) \equiv p_k$ and $q_k(1) \equiv q_k$, the

popularity distribution can be written as

$$q_k(i) = \frac{q_k \prod_{j=1}^{i-1} p_k(j)}{\sum_{l=1}^{K} q_l \prod_{j=1}^{i-1} p_l(j)}. \tag{3.11}$$

As we described in Sec.III.B, the input content request rate of each node is equal to the intensity of miss request at the front node, and thus $\lambda(i)$ is given by

$$\begin{aligned} \lambda(i) &= \lambda(i-1)\textstyle\sum_{k=1}^{K} p_k(i-1)q_k(i-1) \\ &= \lambda(i-2)\textstyle\sum_{k=1}^{K} p_k(i-1)p_k(i-2)q_k(i-2) \\ &= \lambda\textstyle\sum_{k=1}^{K} q_k \prod_{j=1}^{i-1} p_k(j), \qquad \forall i > 1. \end{aligned} \tag{3.12}$$

As a result, the expression of the miss probability at node i is given by

$$\begin{aligned} p_k(i) &= e^{-\frac{\lambda(i)}{m} q_k(i) g(i) c_i^{\alpha}} \\ &= e^{-\frac{\lambda q_k \prod_{j=1}^{i-1} p_k(j)}{m} \frac{\sigma(i-1)}{\lambda(i)} g(\frac{c_i}{c_1})^{\alpha} c^{\alpha}} \\ &= p_k(1)^{(\frac{c_i}{c_1})^{\alpha} \prod_{j=1}^{i-1} p_k(j)}, \qquad \forall i > 1. \end{aligned} \tag{3.13}$$

Average delivery time

After obtaining the miss probability for class k at each node, we analyse the average content delivery time.

We define VRTT_k to indicate the average duration between the emission of an interest in class k and the reception of the corresponding chunk. It is important to note that virtual round trip times are not constant in practice since hit/miss probabilities may change over time. However, we examine the system in steady state under a stationary content request process where the average hit/miss probabilities and VRTT_k have converged into a constant value. VRTT_k is formed of the weighted sum of propagation delay

and transmission of each node, where the value of the weight is cache miss probabilities $p_k(i-1)$ at the previous node. Therefore, the virtual round trip time can be obtained by $\mathrm{VRTT}_k = \sum_{i=1}^{N}(T_N(i)+2\theta_i)p_k(i-1)$, where θ_i indicates the link delay from node $i-1$ to node i and $p_k(0)=1$ denotes that all the requests dispatched by consumer will be sent to the connected router. Note that VRTT_k is in term of chunk packet. Accordingly, one can easily infer content level delivery time $E(T_k) = \delta \mathrm{VRTT}_k$.

Thus, $\forall i \geq 1$, the average delivery time for content items in popularity class k is given by:

$$E(T_k) = \delta \sum_{i=1}^{N}(T_N(i)+2\theta_i)\prod_{j=1}^{i}\mathrm{P}_k(j-1). \tag{3.14}$$

where $T_N(i)$ can be calculated by Eq. (4.16)

3.2.4 Network of consumers

In the previous subsection, we derive the expression for the average delivery time for content items in popularity class k under single consumer. However, as shown in Fig. 3.2, the fact that multiple consumers in the networks in realistic scenarios makes the performance of content delivery hard to estimate. To tackle this problem, we present an analytical model of the average delivery time for content under network of consumers.

The node in which the previous request arrives and the new requester accesses is called a junction node. Having junction nodes is a key feature that differentiates network of consumers model from single consumer model. Recall that the average delivery time is generated based on cache miss rate and Eq. (3.13) is also applicable for router before the junction node. Therefore, we characterize the cache miss rate at the junction node before giving the expression of average content delivery time.

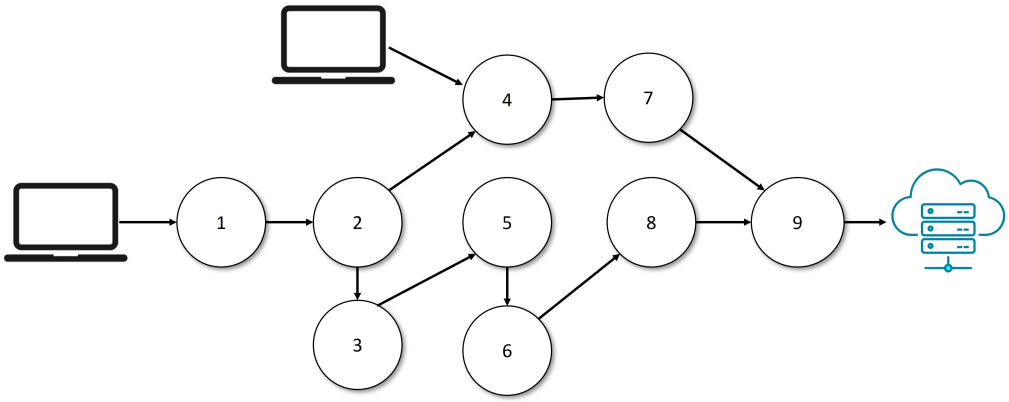


Figure 3.2: Network of consumers topology

Cache miss rate characterization

Given another consumer with similar request preference connect to network at the x-th router, the cache miss rate $p_k(i)$ for class k at node i (where $i > 1$) is given by

$$p_k(i) = \begin{cases} p_k(1)^{(\frac{c_i}{c_1})^{\alpha} \prod_{j=1}^{i-1} p_k(j)} & i < x \\ p_k(1)^{(\frac{c_i}{c_1})^{\alpha}(1+\prod_{j=1}^{i-1} p_k(j))} & i = x. \\ p_k(1)^{(\frac{c_i}{c_1})^{\alpha}(1+\prod_{j=1}^{i-1} p_k(j))\prod_{l=x}^{i-1} p_k(l)} & i > x \end{cases} \tag{3.15}$$

Proof. The content request rate at the junction node $\lambda(x) = \lambda + \sigma(x-1)$ is composed of the total content request rate from the new consumer λ and the intensity of request missed at node $x-1$, which could be obtained through the expression (3.12). The popularity distribution for class k at node x, $q_k(x)$ is determined by the Probability of requests from new consumer and requests missed from the previous node, can be written as

$$q_k(x) = \frac{q_k + q_k \prod_{j=1}^{i-1} p_k(j)}{1 + \sum_{l=1}^{K} q_l \prod_{j=1}^{i-1} p_l(j)}. \tag{3.16}$$

Based on the probability mass function of Poisson distribution, The probability that the chunk request miss at the router is $\mathbb{P}(X = 0) = \frac{\gamma^X e^{-\gamma}}{X!} = e^{-\gamma}$, where γ represents the

average request rate for the given chunk between the average interval of two consecutive misses and X indicates the frequency with which cache hits occur. Here we have both chunk request rate at junction node $\lambda(x)$ and popularity distribution for class k at junction node. $g(i)$ is the fixed under the linear topology. Thus, the miss probability at the junction node can be written as

$$p_k(x) = e^{-\frac{\lambda(x)}{m} q_k(x) g(x) c_x^{\alpha}} = p_k(1)^{(\frac{c_x}{c_1})^{\alpha}(1+\prod_{j=1}^{i-1} p_k(j))}. \tag{3.17}$$

Interests that did not hit on the junction node are directed to the next node, thus the miss probability for class k at node $x+1$ can be calculated using the interest arrival rate and popularity distribution at node $x+1$. It results

$$\begin{aligned} p_k(x+1) &= e^{-\frac{\lambda(x+1)}{m} q_k(x+1) g(x+1) c_{x+1}^{\alpha}} \\ &= e^{-\frac{\lambda(x)\sum_{k=1}^{K} p_k(x) q_k(x)}{m} \frac{p_k(x) q_k(x)}{\sum_{k=1}^{K} p_k(x) q_k(x)} \frac{\sigma(x)}{\lambda(x+1)} g c_x^{\alpha} (\frac{c_{x+1}}{c_x})^{\alpha}} \\ &= e^{-\frac{\lambda(x) q_k(x)}{m} g c_x^{\alpha} p_k(x) (\frac{c_{x+1}}{c_x})^{\alpha}} \\ &= p_k(1)^{(\frac{c_{x+1}}{c_1})^{\alpha}(1+\prod_{j=1}^{x-1} p_k(j)) p_k(x)}. \end{aligned} \tag{3.18}$$

where $g(x+1) = g\sigma(i-1)/\lambda(i) = g$ when all the missed interests have been sent to the next node. Based on this, The miss probability at node i (where $i > x$) can be derived by iteratively calculating the miss probability of previous nodes. So we have

$$p_k(i) = p_k(1)^{(\frac{c_i}{c_1})^{\alpha}(1+\prod_{j=1}^{i-1} p_k(j)) \prod_{l=x}^{i-1} p_k(l)}. \tag{3.19}$$

Here we derive the expression for cache miss rate when there are two consumers. For the scenario that has network of consumers, the miss probability can be computed by iterating Eq. (3.15) when $i = x$ and $i > x$.

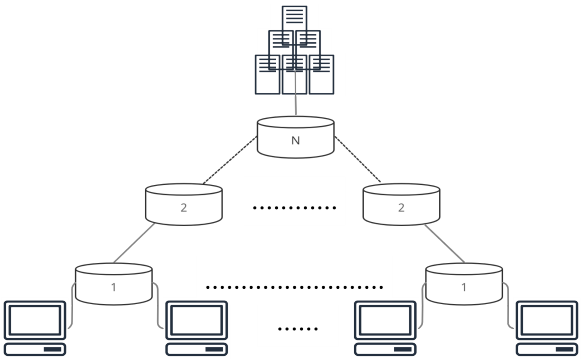


Figure 3.3: Tree network topology

Similarly, our model can be extend to the tree topology in Fig. 3.3. The tree topology is a reasonable topology for ICN, because with the repository at the root, the shortest path is always the path linked to the parent. The only difference between the linear and tree topology is the intensity of input rate at upper level caches, $\lambda(i) = 2\sigma(i-1)$. Let us consider the case $i = x+1$. Note that $g(x+1)/g = \sigma(x)/\lambda(x+1) = \frac{1}{2}$. So we have

$$\begin{aligned} p_k(x+1) &= e^{-\frac{\lambda(x+1)}{m} q_k(x+1) g(x+1) c_{x+1}^{\alpha}} \\ &= e^{-\frac{2\lambda(x)\sum_{k=1}^{K} p_k(x) q_k(x)}{m} \frac{p_k(x) q_k(x)}{\sum_{k=1}^{K} p_k(x) q_k(x)} \frac{\sigma(x)}{\lambda(x+1)} g c_x^{\alpha} (\frac{c_{x+1}}{c_x})^{\alpha}} \\ &= p_k(1)^{2\frac{\sigma(x)}{\lambda(x+1)} (\frac{c_{x+1}}{c_1})^{\alpha} (1+\prod_{j=1}^{x-1} p_k(j)) p_k(x)} \\ &= p_k(1)^{(\frac{c_{x+1}}{c_1})^{\alpha} (1+\prod_{j=1}^{x-1} p_k(j)) p_k(x)}. \end{aligned} \tag{3.20}$$

We could find that the result of Eq. (3.20) is equal to Eq. (3.18). Therefore, the expression of cache miss rate for tree topology follows the same calculations made for the linear topology.

Average delivery time

Since we get the miss probability for class k at each node under network of consumers, the output interest rate from the front node can be written as

$$\sigma(i) = \begin{cases} \lambda \sum_{k=1}^{K} q_k \prod_{j=1}^{i} p_k(j), & i < x \\ \lambda (\sum_{k=1}^{K} q_k \prod_{j=1}^{x-1} p_k(j) + 1) * & \\ \qquad \sum_{k=1}^{K} q_k(x) \prod_{j=x}^{i} p_k(j), & i \geq x. \end{cases} \tag{3.21}$$

By using $\sigma(i)$ as an input of Eq. (4.16), one can obtain the queueing delay at each node. The next step is to compute the average delivery time for content of class k for each consumer. We define $E_x(T_N(i))$ to indicate the average delivery time for new consumer connect to network at x-th router. The delay of content transfer is composed of propagation delay, transmission delay and queueing. The propagation delay can be derived directly from the miss probability and the link delay . The transmission delay is based on the content size and bandwidth . As for queueing delay, it require to acquire chunk arrival rate at first. As ICN inherently routes the chunk on the reverse path of the interest by recording the interface of the coming interest, the chunk arrival rate at interface of node i is equal to the interest arrival rate at interface of node i. Recall that in Sec.3.2.2 we define $\rho(i) = \sigma(i-1)/\mu(i)$, the purpose of using $\sigma(i-1)$ replace $\lambda(i)$ is to just compute those chunks that pass through this interface. As for the new consumer, chunks arrival rate at the junction is equal to λ. Therefore, the queueing delay for the new consumer at the junction can be obtained by $T_N(1)$, and thus the average delivery time of class k can be written as

$$E_x(T_k(x)) = \begin{cases} \delta \sum_{i=1}^{N} (T_N(i) + 2\theta_i) \prod_{j=1}^{i} \mathrm{P}_k(j-1), & x = 1 \\ \delta (T_N(1) + 2\theta_x + \sum_{i=x+1}^{N} (T_N(i) + 2\theta_i) \prod_{j=x}^{i} \mathrm{P}_k(j-1)), & x > 1. \end{cases} \tag{3.22}$$

Following our discussion so far, we derive the miss probability and the average de-

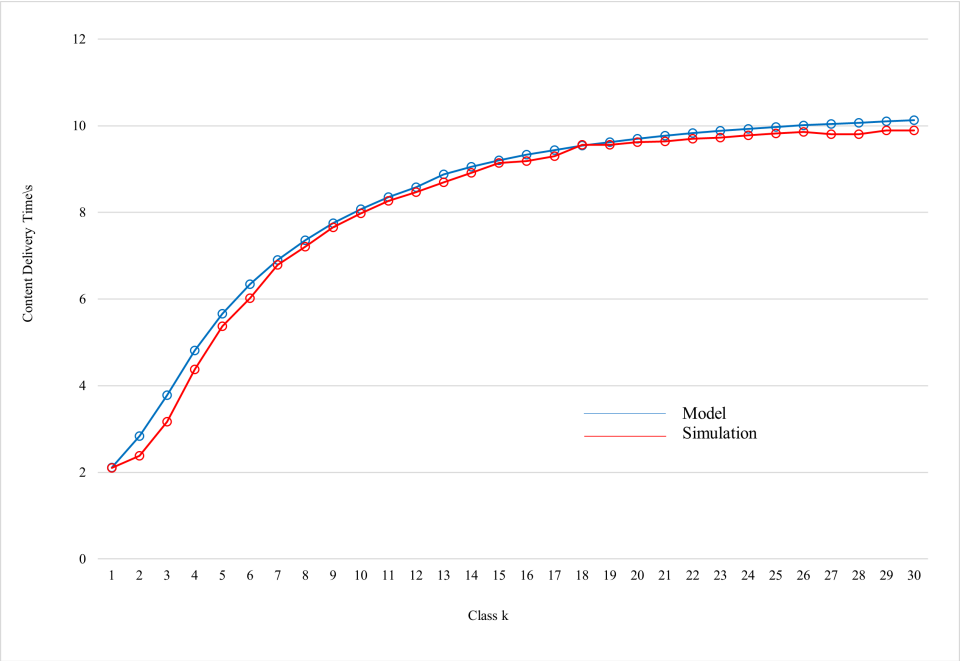


Figure 3.4: Average content delivery time as a function of the popularity class k

livery time for both single consumer and network of consumers, which is a fundamental performance metric for content delivery. We will validate our proposed model in the next section.

3.3 Validation of The Model

The effectiveness and accuracy of the developed analytical models are validated via a modular NDN simulator ndnSIM [90], developed under the ns-3 framework. This open-source simulator implements the CS, PIT and FIB data structures, and content retrieve operations of ICN. For all the simulation scenarios mentioned, we run the simulations 50 times and present our findings based on the mean value of the simulation results.

3.3.1 Single consumer

We consider a population of $M = 6 \times 10^3$ different contents, all the contents are evenly allocated in $K = 30$ classes, each class has $m = 200$ contents. Content popularity follows the Zipf distribution with the exponent parameter $\alpha = 2$. Jia et al. [84] indicates that the realistic Internet video content of YouTube follows Zipf distribution with exponent $\alpha = 2$.

We assume each chunk with the same size of 1024 Bytes, which has good bandwidth savings result in ICN scenario [91]. The size of content is geometrically distributed with average $\delta = 10^3$ chunks. We use the topology in Fig. 3.1 with $N = 5$, in the network, all the links have the same delay 1ms, and the same bandwidth capacity $100Mbps$. Every router is equipped with a cache with size $c = 10^6$ chunks (1GBs) and implement an LRU replacement policy.

Fig. 3.4 depicts the average content delivery time where the arrival of consumer's content request follows the Poisson process with intensity $\lambda_{content} = 10\ content/sec$ to compare the numerical results of model and simulation. It shows that the cache node located closest to the consumer is more likely to host the popular content, driving down the content delivery time.

3.3.2 Network of consumers

In this subsection, we verify the correctness of our model under network of consumers. Here we use the topology in Fig. 3.2 with $N = 5$. The junction node where the new consumer connects in is allocated with $x = 3$, The parameters are otherwise identical to those used in single consumer. Fig. 3.5 depicts the average content delivery time for consumers connect to network at node 1 and node 3. In comparing Fig. 3.4 and Fig. 3.5, we can see that when a network has new consumers connected, the delivery time for content with higher popularity becomes shorter than the situation where there is only one consumer. With new consumers connected to the network, the proportion of requests for high popularity content after the junction node increases, resulting in an increase in the possibility of high popularity content being cached after the junction node. Consequently, The content delivery time for these popular contents correspondingly becomes smaller.

The figure reveals that the analytical performance results closely match those ob-

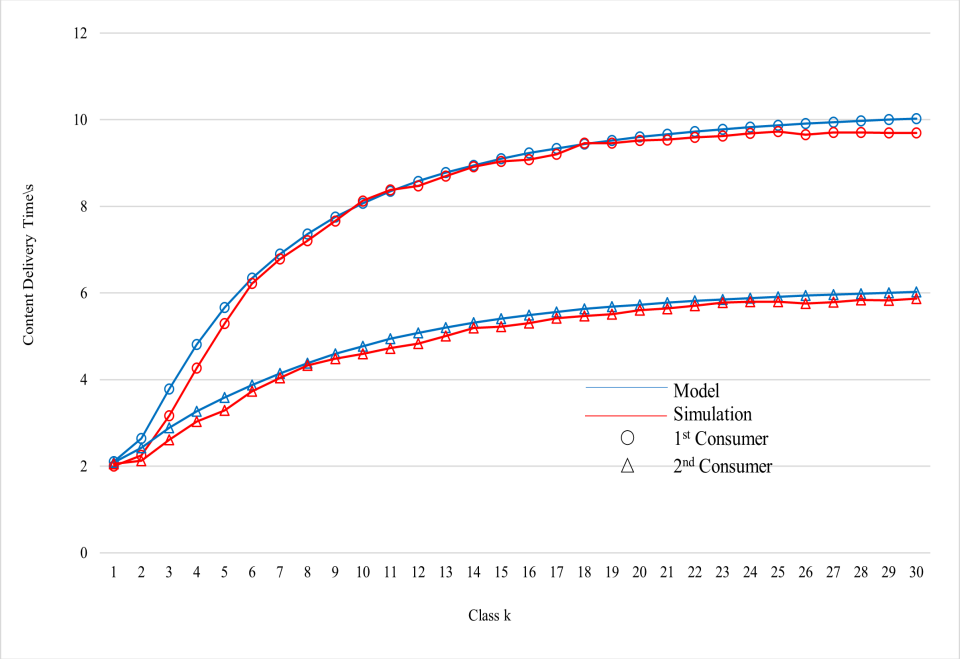


Figure 3.5: Average content delivery time as a function of the popularity class k under network of consumers

tained from the simulation experiments both in single consumer model and network of consumers model, validating the accuracy of the developed analytical model.

3.3.3 Performance analysis

In this subsection, the developed model is used as a cost-efficient tool to predict the impact of key metrics on the performance of content transfer for a specific service. According to Eq. (4.16), Eq. (3.15) and Eq. (3.22), the content delivery time is a function of content arrival rate λ, the cache size c and the Zipf exponent α. We will investigate the impact of these three parameters by changing their value. The parameters are otherwise identical to those used in the single consumer subsection.

As depicted in Fig. 3.6, the content delivery time grows with the amount of λ. With an increase of content arrival rate, the time that chunks wait in the queue increases, which in turn leads to an increase of delivery time.

The result of Zipf exponent α influence on the performance of content delivery is shown in Fig. 3.7. With the α decreases, content requests can be more diverse, which

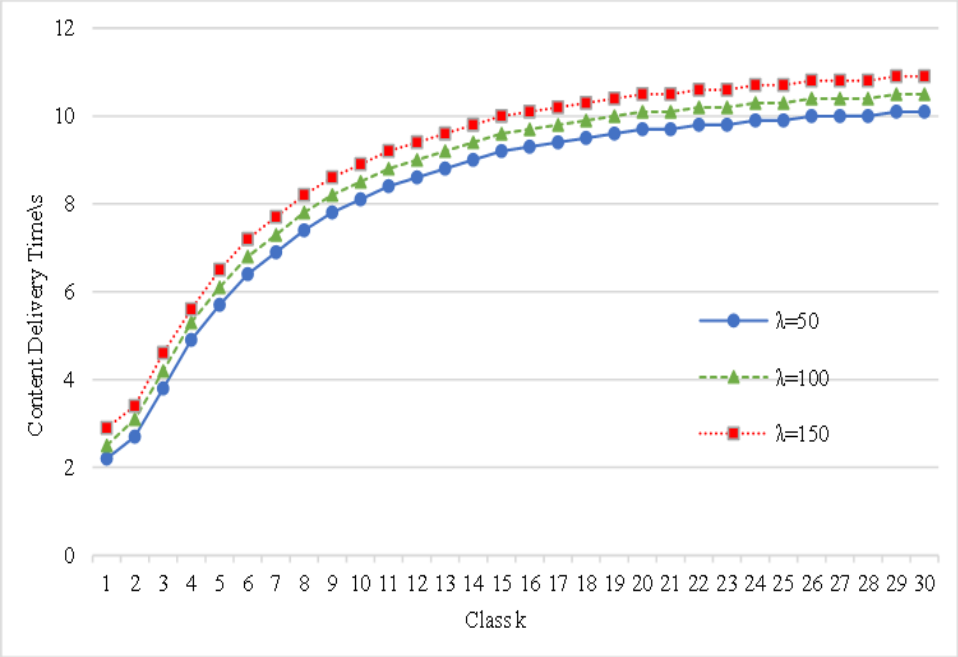


Figure 3.6: Content delivery time predicted by the model for different content arrive rate

results in chunks being removed from the cache more frequently. This drives the increase of cache miss rate and finally lead to the increasing of content delivery time.

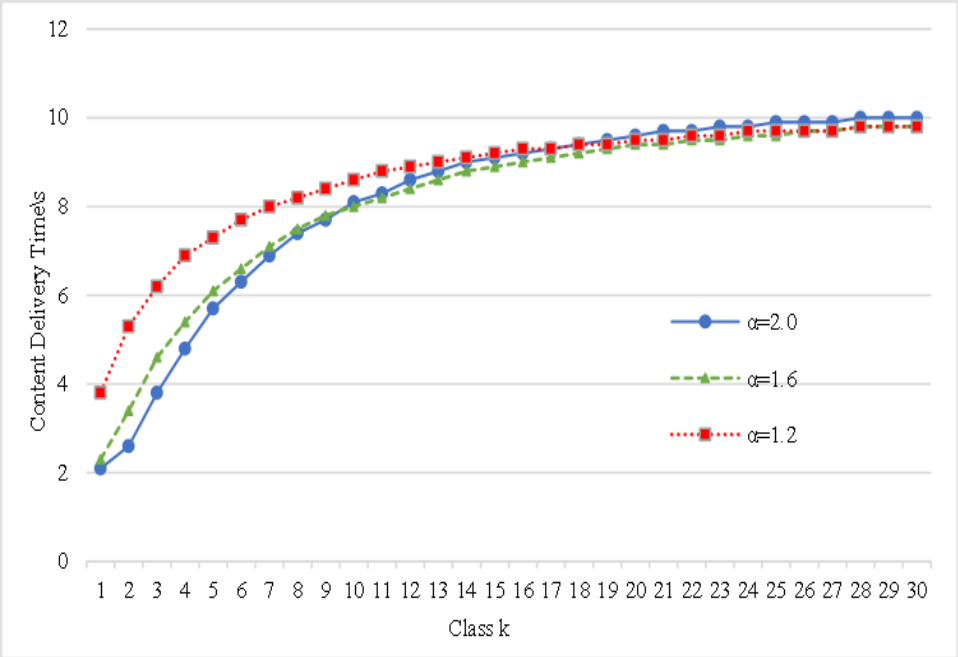


Figure 3.7: Content delivery time predicted by the model for different Zipf α

The cache size c of ICN router also play an important role to predict content delivery time. The result is depicted in Fig. 3.8. When the cache size grows, the number of contents stored close to the consumer increases, improving the cache hit rate, and finally,

reducing the delivery time of content.

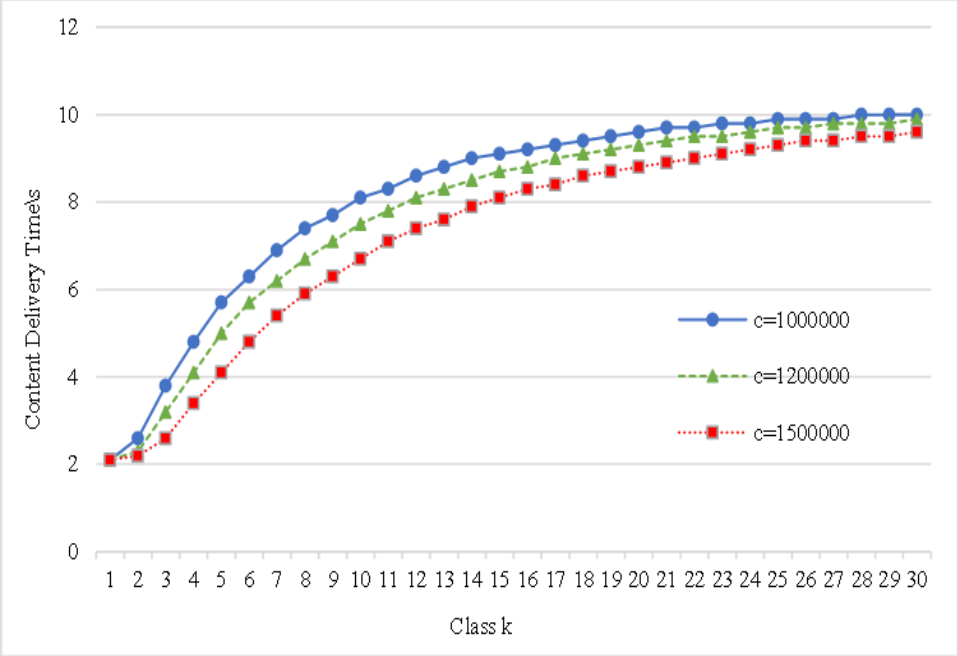


Figure 3.8: Content delivery time predicted by the model for different cache size

The figures demonstrate that the developed analytical model manages to predict the content delivery time for each content with different popularity in ICN.

3.4 Summary

In this chapter, a new analytical model based on the queueing network has been proposed to investigate the performance of content delivery in ICN. Building upon prior work, which estimated the impact of the propagation delay on the performance of content transfer, the present model includes consideration of both the transmission delay and queueing delay. These factors also have impact on network performance and result in a more accurate and practical assessment of content delivery time. In parallel, extensive experiments have been performed to evaluate the accuracy of our proposed model, and the experimental studies demonstrated that the performance results predicted by the analytical model closely match those obtained from the simulation. The proposed analytical model offers a valuable tool for predicting the impact of new consumers on transfer performance in

ICN systems and evaluating the performance of various cache replacement policies and services. This study provides a comprehensive analysis of the ICN system's behavior and highlights the importance of key metrics for optimizing performance. this study lay a groundwork for future research. In this next chapter, the content transfer in the wireless network of ICN will be investigated, and the effect of interests aggregation will be evaluated.

Chapter 4

Performance Analysis of Pending interest Table and Content Transfer in Wired and Wireless ICN

4.1 Introduction

A fundamental principle of ICN is to replace the current Internet's host-based IP address with an information-based naming scheme [76]. As a result of this principle, ICN is able to provide consumers with uniquely identifiable network content by naming-based routing rather than the routing of traditional TCP/IP that needs to establish the connection between consumer and provider [17]. Consequently, information-centric naming allows seamless mobility without needing to perform complex network management required in IP networks when a mobile node's physical and topological location changes. As reported in [92], global mobile data traffic will increase 11-fold from 2013 to 2018. By 2018, fixed devices will contribute to 39 percent of network traffic, while the rest of network traffic will be brought about by WiFi and mobile devices. As opposed to traditional content providers who have stationary servers. New generations of content producers, such as

live streamers and YouTubers, are more likely to work from an unfixed location. As a result, when studying mobile traffic, we should not only consider the mobility of the consumer but also consider the mobility of the provider.

Pending interest table (PIT) is one of the essential components of the ICN forwarding plane, which is responsible for stateful routing in ICN. It facilitates the routing of chunks back to the reverse path of interest. Furthermore, it aggregates the same interests to alleviate congestion and reduce the load on the network. With the PIT factored into the consideration, the analysis of the content transfer process would be closer to that of realistic network interaction.

A unified and accurate model for characterizing of PIT performance can be used to evaluate the content transfer efficiency in ICN. The existing studies, including our previous work as shown in chapter 3, are short of considering the impact of PIT on content transfer performance. In parallel, the weakness of the existing content transfer studies of ICN is also reflected in the assumption of traffic pattern. Most of the existing studies consider the simplified traffic models such as constant arrival rate and Poisson process [57, 58, 83, 93], which fail to capture the bursty nature of content requests in ICN. To fill in the gap, this chapter aims to characterize the PIT miss rate and investigate the performance issues of content transfer under bursty content requests. To this end, a new analytical model is developed as a cost-effective performance tool to investigate the content transfer of ICN under bursty content requests. Using the developed model, we examine the influence of chunk transmission window size on content transfer performance. As ICN is intrinsically compatible with wireless networks, we also developed an analytical model to evaluate the mean service time based on consumer and provider mobility. All the developed analytical model adopts the Markov modulated Poisson process (MMPP) to capture the bursty nature of the content requests. The Least-Recently-Used (LRU) re-

placement policy is taken into account because it has been applied successfully in many caching systems [94]. The accuracy of the analytical model is validated through extensive simulation experiments. Finally, the analytical model is used to evaluate the impact of key metrics, such as the cache size, content size and content popularity on the performance of PIT and content transfer in ICN.

The rest of this chapter is organized as follows. Section 4.2 presents the detail of characterizing the PIT miss rate. Then, an analytical model is developed to investigate the performance of content transfer under bursty content requests. After that, the impact of chunk transmission window size on content delivery time is evaluated. The analytical model for evaluating the mean service time based on consumer and provider mobility is also proposed. Section 4.3 validates the analytical model and carries out the performance analysis. Finally, Section 4.4 concludes the chapter

4.2 Analytical Model of Content Delivery

In this section, we model the content delivery time in ICN under bursty arrival. In order to do this, the system parameters are introduced. After that, the impact of cache hit and PIT miss on content transfer under bursty arrival will be represented. Then, we approximated a multistate MMPP with a 2-state MMPP. Finally, we model the content delivery time by incorporating both delays caused by queueing and propagation.

4.2.1 System Parameters

We summarize the parameters used to develop the analytical model in this section, followed by the explanation of how bursty content requests are modeled.

Notations: To facilitate reading, the notations of system parameters are summarised in Table 1. The detailed explanation of notations is that data chunks may be obtained from the repository or from any cache along the path that contains a temporary copy of

Table 4.1: SYSTEM PARAMETERS INVESTIGATED CHAPTER 4

Parameter	Meaning
W	The size of chunk transmission window
c_i	Cache size of node i in number of chunks
B_i	Link bandwidth from node $i-1$ to node i
θ_i	The propagation delay from node $i-1$ to node i
λ,$\lambda(i)$	Content request rate at at node 1,i
q_k,$q_k(i)$	Popularity distribution for class k at node 1, i
$p_k(i)$	Cache miss rate for class k at node i
$p_{pit,k})$	The PIT miss rate for content of class k at the first router
$P_k(i)$	Total Miss probability for class k at node i
$T_N(i)$	Queueing delay for one chunk experienced at node i
VRTT_k	Virtual round trip delay of calss k
$\text{RVRTT}_k(i)$	The residual virtual round trip time of class k at node i
$T_{k,p}$	The average waiting time of pending interest of class k before consumer moving
$T_{k,chunk}(n)$	The average service time for chunk of class k when producer be the n^{th} node
$W_{k,consumer}$	Mean service time for chunk of class k in wireless network in term of consumer mobility
$W_{k,provider}$	Mean service time for chunk of class k in wireless network in term of provider mobility

the data chunk. This results in chunks of the same content being retrieved from different locations, which makes the round trip times (RTTs) not identical, and therefore affects the performance of the delivery. To model the performance of delivery. The virtual round trip time of class k, VRTT_k, is defined as the average time between the dispatch of a chunk request and its reception in steady state. This variable functions similarly to the round trip time for TCP connections in IP networks. In this chapter, the notation that was introduced in Chapter 3 is consistently employed and retains the same definition throughout.

Bursty content request: This model assumes that consumers generate interests according to a Markov Modulated Poisson Process, which is a doubly stochastic process. MMPP has been widely employed in models of bursty content request since it provides a qualitative representation of time-dependent arrival rates while remaining computationally tractable. On the basis, A special case of the MMPP called Interrupted Poisson Process (IPP) is used to model the arrival interest for a specific type of service in this chapter. The IPP is defined as a 2-state MMPP with one arrival rate being zero. IPP_k with subscript k represents the request for a content in class k, and it is characterized by

an infinitesimal generator Q_k of the underlying Markov process as well as the coefficient matrix Λ_k. Q_k and Λ_k are given by

$$Q_k = \begin{bmatrix} -\sigma_{12} & \sigma_{12} \\ \sigma_{21} & -\sigma_{21} \end{bmatrix}, \quad \Lambda_k = \begin{bmatrix} \lambda_{rk} & 0 \\ 0 & 0 \end{bmatrix}. \tag{4.1}$$

where σ_{12} indicates the transition rate from state 1 to state 2, and σ_{21} donates the transition rate from state 2 to 1. The request arrival rate is λ_{rk} when the Markov chain is in state 1. We define the steady-state vector of MMPP as π, and $\pi_k = (\pi_1, \pi_2)$ donates the steady-state vector for IPP_k, according to the definition of MMPP in cookbook [95], The steady-state vector π_k need to satisfy $\pi Q = 0$ and $\pi e = 1$, where $e = (1, 1, ..., 1)^{\mathrm{T}}$ is the column vector. We also have the normalizing equation $\pi_1 + \pi_2 = 1$. On the basis, we have

$$\pi_1 = \frac{\sigma_{21}}{\sigma_{12} + \sigma_{21}}, \quad \pi_2 = \frac{\sigma_{12}}{\sigma_{12} + \sigma_{21}}. \tag{4.2}$$

then, the mean arrival rate of IPP_k in steady state, λ_k is given by

$$\lambda_k = \lambda_{rk} \times \pi_1 + 0 \times \pi_2 = \frac{\sigma_{21}\lambda_{rk}}{\sigma_{12} + \sigma_{21}}. \tag{4.3}$$

The superposition of K individual IPPs process is remains an MMPP. From the individual generators Q_k and coefficient matrices Λ_k, the computation of the generator Q and the rate matrix Λ are performed as follows

$$\begin{aligned} Q &= Q_1 \oplus Q_2 \oplus \cdots \oplus Q_K, \\ \Lambda &= \Lambda_1 \oplus \Lambda_2 \oplus \cdots \oplus \Lambda_K. \end{aligned} \tag{4.4}$$

where $\oplus$ represents the Kronecker-sum(defined in [96]. Since each IPPs is a 2-state MMPP, the composite Q and Λ of the superposed MMPP are $2^k \times 2^k$ matrices and can

be written as

$$
\begin{aligned}
Q &= \begin{bmatrix} -\sigma_1 & \sigma_{12} & \cdots & \sigma_{12^K} \\ \sigma_{21} & -\sigma_2 & \cdots & \sigma_{22^K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{2^K 1} & \sigma_{2^K 2} & \cdots & -\sigma_{2^K} \end{bmatrix} \\
\sigma_i &= \sum_{j=1, j\neq i}^{2^K} \sigma_{i,j} \\
\boldsymbol{\lambda} &= (\lambda_1, \lambda_2, \cdots, \lambda_{2^K})^{\mathrm{T}} \\
\Lambda &= diag(\boldsymbol{\lambda}^{\mathrm{T}}).
\end{aligned}
\tag{4.5}
$$

From the steady-state vector $\boldsymbol{\pi}$ and arrival rate vector $\boldsymbol{\lambda}$, it can derive the mean arrival rate λ_{total} of the composite MMPP as follows

$$\lambda_{total} = \boldsymbol{\pi}\boldsymbol{\lambda}. \tag{4.6}$$

4.2.2 Estimation of Cache and PIT miss rate

A router search for content in its own cache called the Content Store when an interest arrives at the router. In the event that the interest has not been resolved on the cache of the router, it will be passed to look up the PIT table. Only if the interest does not hit both the cache and PIT table will it be sent to the next router. For purposes of estimating the number of interests that have passed through each intermediate node, in this subsection, we model the impact of cache and PIT on content transfer. Here, we first discuss the model of the cache miss rate, and then we derive an analytical model of the PIT miss rate

Cache miss rate

Now, we model the effects of ICN caching on content delivery. To perform this analysis, we need to present the cache miss probabilities at the intermediate nodes along the

path of requests. Since the cache miss rate of content depends on its popularity, we assume content belongs to an arbitrary popularity class k. Let $p_k(i)$ donate the cache miss probability for content of class k at the intermediate node i. Under the assumption that all caches adopt the Least Recently Used (LRU) cache replacement policy, $p_k(i)$ can be derived through the following expression [58]

$$\begin{aligned} p_k(1) &= e^{-\frac{\lambda}{m} q_k g(1) c_1^{\alpha}} = e^{-(\frac{C_1}{m\delta\Gamma(1-\frac{1}{\alpha})})^{\alpha}} \\ \frac{g(i)}{g(1)} &= \frac{\mu(i-1)}{\lambda(i)}. \end{aligned} \tag{4.7}$$

where C_1 is the cache size of the i^{th} router, $\mu(i-1)$ is the frequency of request miss at the $(i-1)^{th}$ router, δ is the average content size in number of chunks and α is the constant parameter of Zipf distribution. In parallel, $1/g = \lambda d \delta^{\alpha} m^{\alpha-1} \Gamma(1-\frac{1}{\alpha})^{\alpha}$ and $g(i)/g = \sigma(i-1)/\lambda(i)$ [58]. This expression is based on the probability mass function of Poisson distribution. Here, c_i denotes the cache size in number of chunks and two consecutive misses for the same chunk cannot take place before gc_i^{α}. Therefore, cache miss rate can be translated to the probability that chunks of content in class k have not been requested during the gc_i^{α}.

For the stationary miss probabilities at node i, using the i^{th} and $(i+1)^{th}$ routers along the path from consumer to the provider as an example. The requests arriving at $(i+1)^{th}$ router should not be hit at previous i routers. Based on this, the miss probability at $(i+1)^{th}$ router can be derived by iteratively calculating the miss probability of $1^{st}, , i^{th}$ routers. Evaluating this relation gives

$$p_k(i) = p_k(1)^{(\frac{c_i}{c_1})^{\alpha} \prod_{j=1}^{i-1} p_k(j)}, \qquad \forall i > 1. \tag{4.8}$$

PIT miss rate

We then examine the impact of PIT on content transfer. The PIT table functions by tracking pending interests at intermediate nodes and preventing the dispatch of the same interest when previous interest for the same chunk has been emitted and the chunk has not yet been received. Taking the impact of PIT into account has the promising to increase the accuracy of content transfer model as it impacts both throughputs the cache behavior of the following node.

In ICN, only if the interest does not hit on the cache of the router will be passed to look up the PIT table. Thus, the request arrival rate for content of class k at PIT table of first router is $\frac{\lambda_k}{m}(1-p_k(1))$.

In steady state, $\mathrm{RVRTT}_k(i)$ denotes the residual virtual round trip time of class k at node i, that is the average elapsed time between the emit of interest and the reception of its corresponding chunk at node i.

Based on the probability mass function, the probability that PIT miss rate at the router is $\mathbb{P}(X=0)=\frac{\gamma^X e^{-\gamma}}{X!}=e^{-\gamma}$, where γ here denotes the average interest arrival frequency for the same chunk within the $\mathrm{RVRTT}_k(i)$ and $\mathbb{P}(X=0)$ indicates the probability that no new interest for the same chunk arrived during the $\mathrm{RVRTT}_k(i)$. Thus, the PIT miss rate for content of class k at the first router $p_{pit,k}$ can be written as

$$p_{pit,k}=e^{-\frac{\lambda_k}{m}(1-p_k(1))\mathrm{RVRTT}_k(1)}. \tag{4.9}$$

The fundamental feature of PIT is to avoid chunk requests flooding is interest aggregation. In the case of linear topology, no interest will be dispatched to the next router on the condition that interest for the same chunk has already been emitted and the chunk has not yet been received. With regard to other network topologies, the interests for the same

chunk are mainly aggregated at the first level of routers, and only interest from different paths for the same chunk has the chance to be forwarded to the second level of routers. The number of same interests that reach the second level of routers is orders of magnitude smaller than the first level of routers. Therefore, we only consider the PIT miss rate at the first router, while the interest aggregation and PIT rate at the second level of routers are negligible. The expression of calculate $\mathrm{RVRTT}_k(1)$ is shown in Eq. (4.23). Momentarily following the Eq. (4.23) illustrates how the iterative method can be used to calculate PIT miss rates.

Now, we take both cache and PIT into account. We use $q_k(i)$ to represent the popularity distribution for class k at i^{th} router. Having $q_k(1)$ follow the Zipf distribution at the first router, the popularity distribution at the i^{th} router can be written as

$$q_k(i) = \frac{q_k p_{pit,k} \prod_{j=1}^{i-1} p_k(j)}{\sum_{l=1}^{K} q_l p_{pit,l} \prod_{j=1}^{i-1} p_l(j)}. \tag{4.10}$$

As the input content request rate of router is equal to the intensity of miss request at the front node, and thus the content request rate at the i^{th} router $\lambda(i)$ is given by

$$\lambda(i) = \lambda_{total} \sum_{k=1}^{K} p_{pit,k} q_k \prod_{j=1}^{i-1} p_k(j), \qquad \forall i > 1. \tag{4.11}$$

Under the assumption that every router implements the LRU cache replacement policy. For linear topology, the miss probability for class k at the first router $\mathrm{P}_k(1)$ can be direct derived by $\mathrm{P}_k(1) = p_k(1) p_{pit,k}$, and the miss probability for class k at the i^{th} router $\mathrm{P}_k(i)$ can be computed by accounting for the modified popularity distribution and modi-

fied content request rate. It results:

$$\begin{aligned}
\mathrm{P}_k(i) &= e^{-\frac{\lambda(i)}{m} q_k(i) g(i) c_i^{\alpha}} \\
&= e^{-\frac{\lambda_{total} \Sigma_{k=1}^{K} p_{pit,k} q_k \Pi_{j=1}^{i-1} p_k(j)}{m} \frac{q_k p_{pit,k} \Pi_{j=1}^{i-1} p_k(j)}{\Sigma_{l=1}^{K} q_l p_{pit,l} \Pi_{j=1}^{i-1} p_l(j)} g(\frac{c_i}{c_1})^{\alpha} c_1^{\alpha}} \\
&= e^{-\frac{\lambda_{total} q_k p_{pit,k} \Pi_{j=1}^{i-1} p_k(j)}{m} g c_1^{\alpha} (\frac{c_i}{c_1})^{\alpha}} \\
&= p_k(1)^{(\frac{c_i}{c_1})^{\alpha} p_{pit,k} \Pi_{j=1}^{i-1} p_k(j)}, \qquad \forall i > 1.
\end{aligned} \tag{4.12}$$

In the case of the complex topology, we assume there are R routers at the first level connected to a single router at the second level. By using the Eq. (4.7), here we have $\frac{g(2)}{g(1)} = \frac{1}{R}$. Thus, the miss probability for the second router can be written as

$$\begin{aligned}
\mathrm{P}_k(2) &= e^{-\frac{R\lambda_{total} \Sigma_{k=1}^{K} p_{pit,k} q_k p_k(1)}{m} \frac{q_k p_{pit,k} p_k(1)}{\Sigma_{k=1}^{K} p_{pit,k} q_k p_k(1)} g(i) (\frac{c_2}{c_1})^{\alpha} c_1^{\alpha}} \\
&= e^{-\frac{R\lambda_{total} q_k p_{pit,k} p_k(1)}{m} g \frac{\mu(i-1)}{\lambda(i)} c_1^{\alpha} (\frac{c_2}{c_1})^{\alpha}} \\
&= p_k(1)^{(\frac{c_2}{c_1})^{\alpha} p_{pit,k} p_k(1)}.
\end{aligned} \tag{4.13}$$

Thus, no matter what the topology, the miss probability remains unchanged.

4.2.3 Superposition of multiple MMPPs

When MMPPs are superposed the state space of the resultant MMPP grows exponentially. Hence we approximate the *m* state MMPP (the resulting arrival process at the router) with a 2 state MMPP [97]

The approach taken by Kathleen is to match the first three noncentral moments of the instantaneous arrival rate of the MMPP, as well as a suitable time constant for the process, which is defined as the integral of the covariance function of the instantaneous arrival rate

of the MMPP. The r^{th} noncentral moment of the superposed MMPP (Q,Λ) is given by

$$\alpha_i = \pi\Lambda^i e. \tag{4.14}$$

where π is the steady-state vector of superposed MMPP.

Based on the covariance function, the appropriately defined time constant is calculated as

$$\begin{aligned}\rho &= v^{-1}\int_0^\infty \pi\Lambda(e^{Qt} - e\pi)\Lambda e \\ &= v^{-1}[\pi\Lambda(e\pi - Q)^{-1}\Lambda e - m^2].\end{aligned} \tag{4.15}$$

where v and m are the variance and mean of the MMPP arrival rate. Assuming the infinitesimal generator and coefficient matrix of approximated 2-state is given by

$$\tilde{Q} = \begin{bmatrix} -\tau_1 & \tau_1 \\ \tau_2 & -\tau_2 \end{bmatrix}, \quad \tilde{\Lambda} = \begin{bmatrix} \tilde{\lambda_1} & 0 \\ 0 & \tilde{\lambda_2} \end{bmatrix}. \tag{4.16}$$

After computing the first three non-central moments and the time constant for the 2-state MMPP, we have

$$\begin{aligned} &\alpha_1 = \tilde{\lambda_1}\tilde{\pi_1} + \tilde{\lambda_2}\tilde{\pi_2}, \qquad \alpha_2 = \tilde{\lambda_1^2}\tilde{\pi_1} + \tilde{\lambda_2^2}\tilde{\pi_2}, \\ &\alpha_2 = \tilde{\lambda_1^3}\tilde{\pi_1} + \tilde{\lambda_2^3}\tilde{\pi_2}, \qquad \rho = (\tau_1+\tau_2)^{-1}, \\ &\tilde{\pi_1} = \frac{\tau_1}{\tau_2+\tau_2}. \end{aligned} \tag{4.17}$$

Due to the fact that our approximated 2-state MMPP is an interrupted Poisson process, we have $\tilde{\lambda_2} = 0$. Now we can obtain the attributes of the 2-state MMPP with

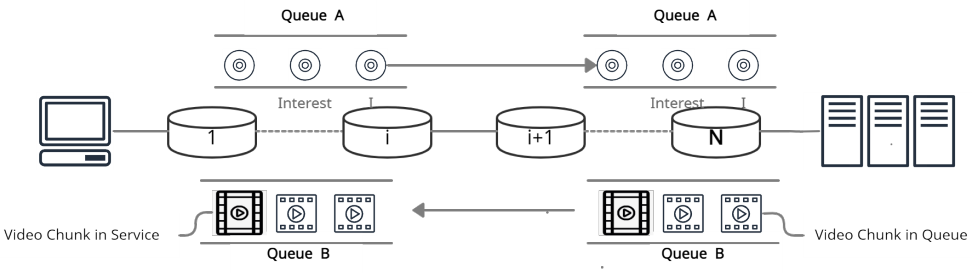


Figure 4.1: Traffic flow in a router

$$\tilde{\lambda_1} = \frac{\alpha_2}{\alpha_1}, \quad \tau_1 = \rho^{-1}\frac{\alpha_2 - \alpha_1^2}{\alpha_2}, \\ \tau_2 = \rho^{-1}\frac{\alpha_1^2}{\alpha_2}. \tag{4.18}$$

4.2.4 Content Delivery time of Wired Networt

One can directly calculate the propagation delay of the content transfer based on the result of the cache and PIT miss rate in the previous subsection. The majority of previous work on content transfer models in ICN lacks the consideration of transmission and queueing delays. In order to reveal the impact of transmission and queueing delay on content delivery time, in this subsection, we develop an analytical model for content delivery in ICN with both propagation delay and transmission delay taken into account.

Average system time

In ICN, the request of a chunk, referred to as interest, is yielded from the consumer node and transmitted hop-by-hop to the repository or cache that contains a temporary copy of the data chunk along the path. Once the request is satisfied, the chunk flows down to the consumer following the reverse path of the request. At peak periods, the network may encounter a high volume of traffic during the transmission of content. During this period, chunks are required to remain in the router awaiting transmission.

As depicted in Fig. 4.1. There are two queues for ICN traffic. Queue A is the queue for interest packets entering the router, while queue B is the queue for returned chunks

leaving the router. The service time of queue A can be performed in constant time as the size of interest is similar. The service time for queue B is also constant as each chunk has a fixed size. Therefore, queue A and queue B are both MMPP/D/1 queues. The size of interests is orders of magnitude smaller than chunk. Hence, the transmission time of the interest is negligible when compared to the propagation and transmission time of the chunk and is thus ignored. If necessary, the interest transmission time can be easily accounted for in the same manner we calculate the chunk transmission time.

We assume D is the size of the chunk, B_i indicates the bandwidth between node $i-1$ and i. Let $h = \frac{D}{B_i}$ be the service time and $h^{(2)}$ be the variance of the service time. The stochastic matrix G is given by

$$G = e^{Q-\Lambda+\Lambda G}. \tag{4.19}$$

The complexity of computing G is exponential growth with the increase of superposed MMPPs. So we approximate the multistate state MMPP with a 2 state MMPP in the last subsection. The steady-state probability vector b of the stochastic matrix G satisfies

$$bG = b, \qquad be = 1. \tag{4.20}$$

Thus, the mean waiting time W_N in the MMPP/D/1 queueing system can be computed as [95].

$$W_N = \frac{1}{2(1-\rho)}[2\rho + \lambda_{total}h^{(2)} - 2h((1-\rho)b + h\tilde{\pi}\tilde{\Lambda})(\tilde{Q} + e\tilde{\pi})^{-1}\tilde{\boldsymbol{\lambda}}]. \tag{4.21}$$

The average system time T_N with both transmission delay and queueing delay considered, can be derived as $T_N = W_N + h$. The average system time $T_N(i)$ for each intermediate router can be derived by assigning the appropriate parameter $\tilde{\Lambda}$.

Wired networt Content delivery time

After obtaining the miss rate of the router and average system time, we proceed to analyze the average delivery time of the content. We define VRTT_k to indicate the average duration between the emission of an interest in class k and the reception of the corresponding chunk. VRTT_k is formed of the weighted sum of propagation delay and average system time of each router, where the value of the weight is miss probabilities $\mathrm{P}_k(i)$ at each intermediate router. Therefore, the virtual round trip time can be obtained by

$$\mathrm{VRTT}_k = \sum_{i=1}^{N} (T_N(i) + 2\theta_i) \prod_{j=1}^{i} \mathrm{P}_k(j-1). \tag{4.22}$$

where θ_i indicates the link delay from node $i-1$ to node i. As all the interests will be delivered to the first router, the residual virtual round trip time of class k at node 1 is can be written as

$$\mathrm{RVRTT}_k(1) = \sum_{i=2}^{N} (T_N(i) + 2\theta_i) \prod_{j=1}^{i} \mathrm{P}_k(j-1). \tag{4.23}$$

As the input parameter for $\mathrm{RVRTT}_k(1)$, $T_N(i)$ and $P_k(i)$ can be determined by applying Eq. (4.12) and Eq. (4.21). Prior to using Eq. (4.12) and Eq. (4.21), it is necessary to address the PIT miss rate presented in Eq. (4.9) . However, as shown in Eq. (4.9), $\mathrm{RVRTT}_k(1)$ is an input parameter in the process of calculating PIT miss rate. Consequently, the equation for calculating PIT miss requires $\mathrm{RVRTT}_k(1)$, while the process of calculating $\mathrm{RVRTT}_k(1)$ requires PIT miss rate, which makes them a closed loop. In light of this, to estimate the PIT miss rate, we apply the iterative method which uses $\mathrm{RVRTT}_k(1)$ that do not consider the influence of PIT aggregation as an initial guess to generate successive approximations to PIT miss rate. In the following section, we will present a chart showing the rate of convergence for the PIT miss rate. Note that VRTT_k is

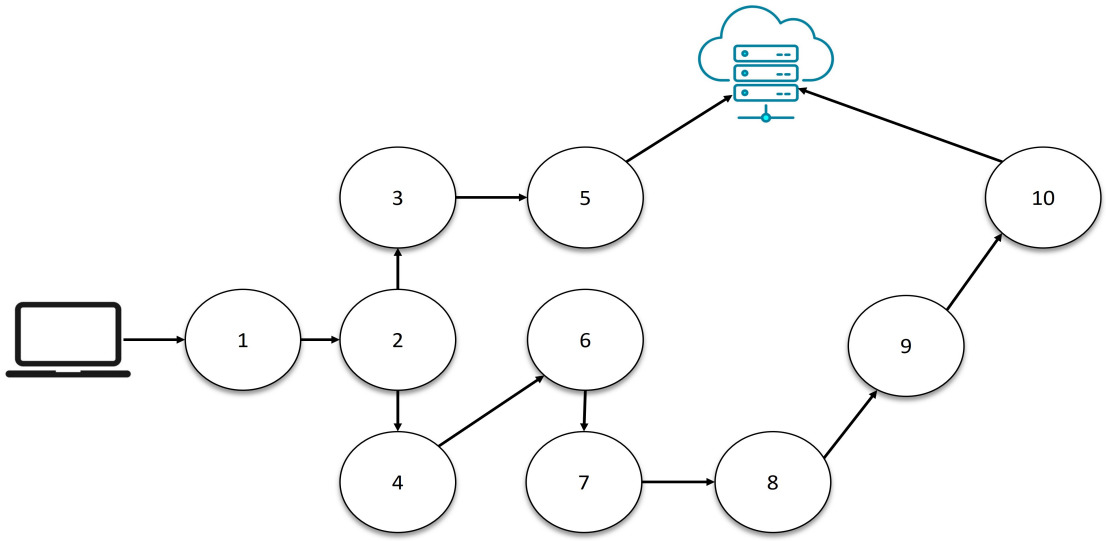


Figure 4.2: General topology

in term of chunk packet level. Delivery time for content level is mainly based on the forwarding strategy at the consumer. We modeled the content delivery time in ICN using two extreme scenarios, in which the chunk transmission window size are $W = 1$ and $W = \infty$, as examples. Readers can easily modify the attributes of W based on their own forwarding strategy. In the first case, the emit of interests of the content in a sequential manner, where only the previous interest has been served, the new interest will be delivered. In this case, the average delivery time for content items in popularity class k is

$$E_k(T_{one}) = \delta \sum_{i=1}^{N} (T_N(i) + 2\theta_i) \prod_{j=1}^{i} \mathrm{P}_k(j-1). \tag{4.24}$$

Another extreme scenario is that all the interests of the contents are sent out together. In this scenario, we have

$$E(T_{all}) = \sum_{i=1}^{N} (T_N(i) + 2\theta_i) \prod_{j=1}^{i} \mathrm{P}_k(j-1). \tag{4.25}$$

As the interest the interest rate control proposed for ICN [1] is initially a simple pipelining and in ICN it remains an open issue how to let W vary over time through a sliding window control. Therefore, we do not specify a numeric value for W. For future study, with the varying of the different forwarding protocols, one can easily infer the content delivery time as a function of the content size and chunk transmission window size, which is given by

$$E_k(W) = \frac{\delta}{W}\mathrm{VRTT}_k. \tag{4.26}$$

4.2.5 Content Delivery time of Wireless Network

ICN paradigms offer a potential for facilitating consumer mobility by separating identity and location [2], which is beneficial for seamless mobility. In this subsection, the model of delivery time for wireless networks will be presented from the perspectives of consumer mobility and provider mobility. Since the forwarding strategy of window size W is uncertain, for convenience, the discussion in this subsection is based on chunk-level.

To facilitate the evaluation of transfer performance at the chunk level, we assume the producer be the n^{th} node. Then average service time for chunk of content of class k, $T_{chunk}(n)$, for this case can be calculated as:

$$T_{k,chunk}(n) = \sum_{i=1}^{n}(T_N(i) + 2\theta_i)\prod_{j=1}^{i}\mathrm{P}_k(j-1). \tag{4.27}$$

Consumer Mobility

Supporting consumer mobility in ICN is generally simpler [98]. In contrast to traditional connection-oriented networks, ICN does not require reestablishing the connection with

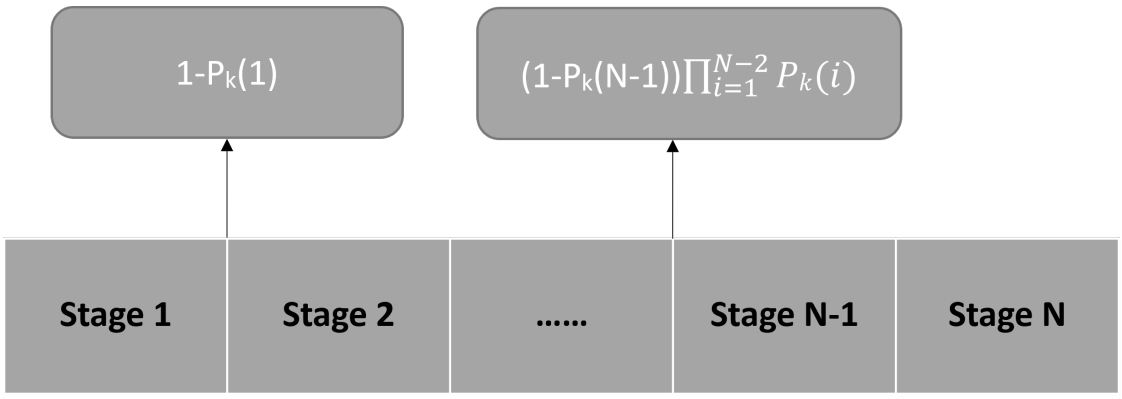


Figure 4.3: The stage of pending interest

the provider. As for consumer mobility in ICN, the moving consumer upon relocation and reattachment to the network can reissue a request for those pending interests. Pending interests refer to those interests that have not yet been served before the consumer moves. Fig. 4.3 Shows the distribution of pending interest at different stages. The probability of pending interest in stage i is $(1-P_k(i))\prod_1^{i-1}P_k(i)$. Therefore, the probability density function of pending interest in stage i is given by

$$g(x)=\begin{cases} 1-P_k(1) & x<T_N(1)+2\theta_1 \\ P_k(1)(1-P_k(2)) & T_N(1)+2\theta_1\le x<\sum_{i=1}^{2}T_N(i)+2\theta_i \\ \vdots & \vdots \\ (1-P_k(N-1))\prod_{i=1}^{N-2}P_k(i) & \sum_{i=1}^{N-2}T_N(i)+2\theta_i\le x<\sum_{i=1}^{N-1}T_N(i)+2\theta_i \\ (1-P_k(N))\prod_{i=1}^{N-1}P_k(i) & \sum_{i=1}^{N-1}T_N(i)+2\theta_i\le x<\sum_{i=1}^{N}T_N(i)+2\theta_i \end{cases} \tag{4.28}$$

Thus, the average waiting time of pending interest before consumer moving T_p can be

written as

$$
\begin{aligned}
T_{k,p} &= \frac{\int_0^{T_N(1)+2\theta_1} xg(x)dx}{T_N(1)+2\theta_1} + \frac{\int_{T_N(1)+2\theta_1}^{\sum_{i=1}^{2} T_N(i)+2\theta_i} xg(x)dx}{(\sum_{i=1}^{2} T_N(i)+2\theta_i)-(T_N(1)+2\theta_1)} + \cdots\cdots \\
&\quad + \frac{\int_{\sum_{i=1}^{N-2} T_N(i)+2\theta_i}^{\sum_{i=1}^{N-1} T_N(i)+2\theta_i} xg(x)dx}{(\sum_{i=1}^{N-1} T_N(i)+2\theta_i)-(\sum_{i=1}^{N-2} T_N(i)+2\theta_i)}) + \frac{\int_{\sum_{i=1}^{N-1} T_N(i)+2\theta_i}^{\sum_{i=1}^{N} T_N(i)+2\theta_i} xg(x)dx}{(\sum_{i=1}^{N} T_N(i)+2\theta_i)-(\sum_{i=1}^{N-1} T_N(i)+2\theta_i)} \\
&= \frac{(1-P_k(1))\int_0^{T_N(1)+2\theta_1} xdx}{T_N(1)+2\theta_1} + \frac{(1-P_k(N-1))\int_{T_N(1)+2\theta_1}^{\sum_{i=1}^{2} T_N(i)+2\theta_i} xdx}{(\sum_{i=1}^{2} T_N(i)+2\theta_i)-(T_N(1)+2\theta_1)} + \cdots\cdots \\
&\quad + \frac{(1-P_k(N))\prod_{i=1}^{N-1} P_k(i)\int_{\sum_{i=1}^{N-2} T_N(i)+2\theta_i}^{\sum_{i=1}^{N-1} T_N(i)+2\theta_i} xdx}{(\sum_{i=1}^{N-1} T_N(i)+2\theta_i)-(\sum_{i=1}^{N-2} T_N(i)+2\theta_i)}) + \frac{(1-P_k(N))\prod_{i=1}^{N-1} P_k(i)\int_{\sum_{i=1}^{N-1} T_N(i)+2\theta_i}^{\sum_{i=1}^{N} T_N(i)+2\theta_i} xdx}{(\sum_{i=1}^{N} T_N(i)+2\theta_i)-(\sum_{i=1}^{N-1} T_N(i)+2\theta_i)} \\
&= \frac{(1-P_k(1))(T_N(1)+2\theta_1)}{2} + \frac{(P_k(1)(1-P_k(2)))(T_N(1)+2\theta_1+\sum_{i=1}^{2} T_N(i)+2\theta_i)}{2} + \cdots\cdots \\
&\quad + \frac{(1-P_k(N-1))\prod_{i=1}^{N-2} P_k(i)(\sum_{i=1}^{N-2} T_N(i)+2\theta_i+\sum_{i=1}^{N-1} T_N(i)+2\theta_i)}{2} \\
&\quad + \frac{(1-P_k(N))\prod_{i=1}^{N-1} P_k(i)(\sum_{i=1}^{N-1} T_N(i)+2\theta_i+\sum_{i=1}^{N} T_N(i)+2\theta_i)}{2} \\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{i}(T_N(i)+2\theta_i)(1-P_k(j))\prod_{h=1}^{i-1} P_k(h)).
\end{aligned}
\tag{4.29}
$$

After T_p is derived, the duration of the consumer is connected to a cellular base station or a Wi-Fi point should be determined. A generalized Gamma distribution [99] can be used to describe the duration of time. We denote the expectation of this duration by η_1, and η_1 is evaluated as follows:

$$
\eta_1 = \frac{\pi R}{2v}. \tag{4.30}
$$

where v represents the average speed of the producer. Here, every cell is assumed to be a circle, and R is the radius of the cell.

In the case of moving of consumer (i.e. the consumer is not connected to the network).

Let us denote the distance to the producer (number of routers on the path from consumer to producer) before the moving by d and the new path after the moving by $\tilde{d}$. Let η_2 denote the average time required for moving consumer relocation and reattachment to the network. This implies that once the moving occurs, the consumer is not in the network for time η_2 . The average service time T_2 in this case is given by

$$T_2 = T_p + \eta_2 + T_{chunk}(\tilde{d}). \tag{4.31}$$

Let the probability that the consumer is reachable at any arbitrary time be denoted by r. Then, we can calculate r as follows:

$$r = \frac{\eta_1}{\eta_1 + \eta_2}. \tag{4.32}$$

Using Equations (4.27) and (4.31), when consider the provider mobility, we can estimate the mean service time for chunk in wireless network $W_{k,consumer}$ as follows:

$$W_{k,consumer} = rT_{k,chunk}(d) + (1-r)T_2. \tag{4.33}$$

Provider Mobility

To characterize provider mobility, the key problem needed to be addressed is that when a provider is moving and becomes unreachable, the probability that the pending interest of the consumer could be served by the in-network caching.

The working flow as shown in Fig. 4.4. Assume that the duration of a provider's connection to a cellular base station or Wi-Fi point follows a generalized Gamma distribution with $\eta_1 = \frac{\pi R}{2v}$, and η_3 is the average time taken for the new address of a provider to be updated at all the relevant nodes. This implies that once the provider moves, the producer is not reachable for interval time η_3, Note that during this period, interests can still be served if the requested data is present in any of the in-network caching along the path to provider. Let the probability that the consumer is reachable at any arbitrary time be denoted by ρ. Then, we can calculate ρ as follows

$$\rho = \frac{\eta_1}{\eta_1 + \eta_3}. \tag{4.34}$$

In the event that the provider moves, let ψ indicate the probability that the interest will not be served by in-network caching. We have $\psi = \prod_{i=1}^{N-1} P_k(i)$. The average service time after moving of provider T_3 is given by

$$T_3 = (1-\psi)(T_{k,chunk(d)} - \psi \sum_{i=1}^{d} (T_N(i) + 2\theta_i)) + \psi(\eta_3 + T_{k,chunk}(\tilde{d})). \tag{4.35}$$

Based on Equation (4.34) and (4.35) , the mean service time for chunk in wireless network $W_{k,provider}$ in term of provider mobility is given by

$$W_{k,provider} = \rho T_{k,chunk}(d) + (1-\rho)T_3. \tag{4.36}$$

Following our discussion so far, we derive the content delivery time based on two

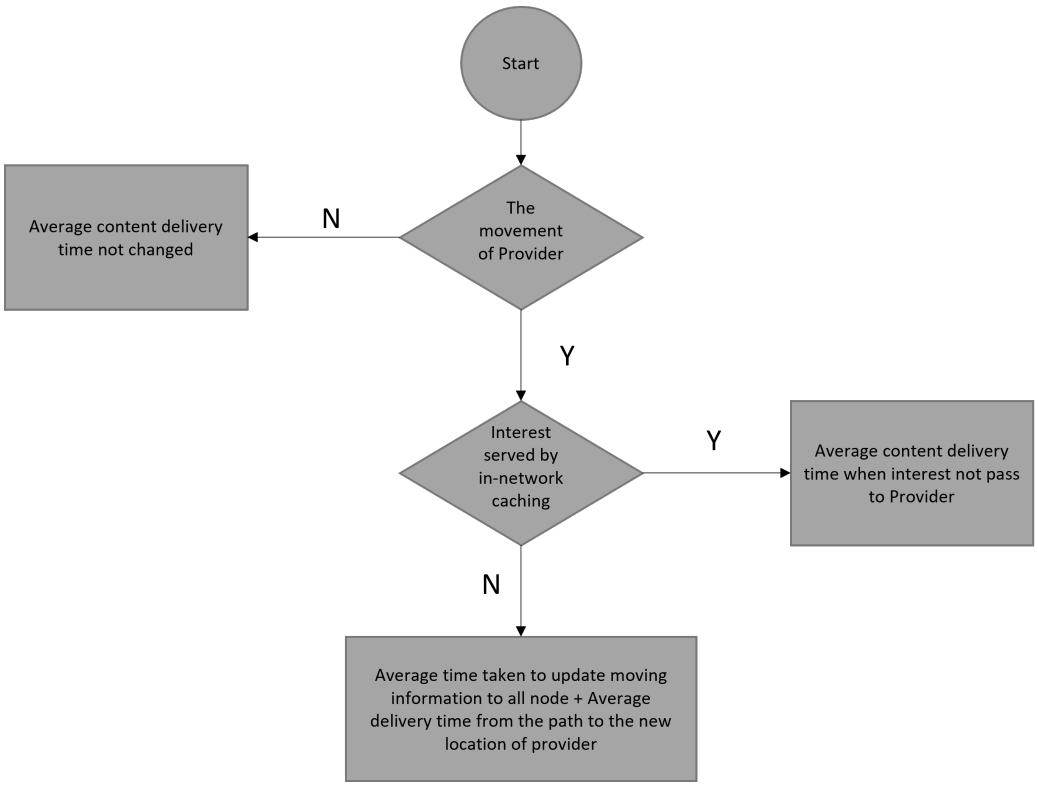


Figure 4.4: The work flow of Provider mobility

different scenarios of forwarding strategies for window size. Additionally, we evaluate the average delivery time for chunks in wireless networks. In the next section, we will validate our proposed model.

4.3 Validation and Performance Analysis

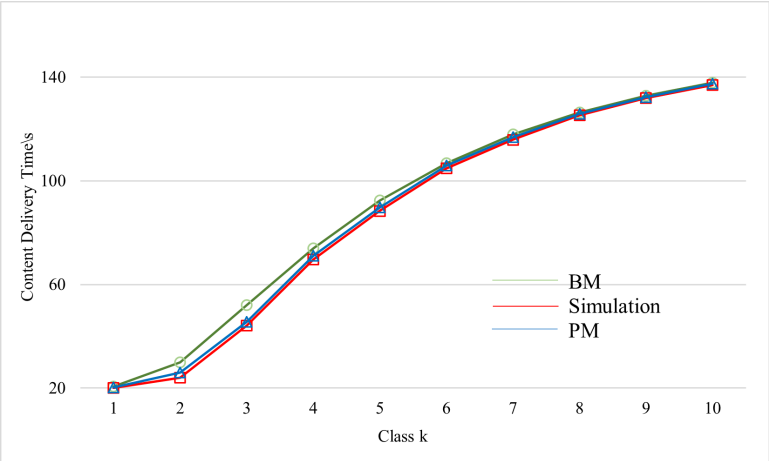
The effectiveness and accuracy of the developed analytical models are validated via a modular NDN simulator ndnSIM [90], developed under the ns-3 framework. This open-source simulator implements the CS, PIT and FIB data structures, and content retrieve operations of ICN. For all the simulation scenarios mentioned, we run the simulations 50 times and present our findings based on the mean value of the simulation results.

We consider a population of $M = 10^3$ different contents, all the contents are evenly allocated in $K = 10$ classes, each class has $m = 100$ contents which are split into chunks of 1024 bytes. and the size of content is geometrically distributed with average $\delta = 10^3$

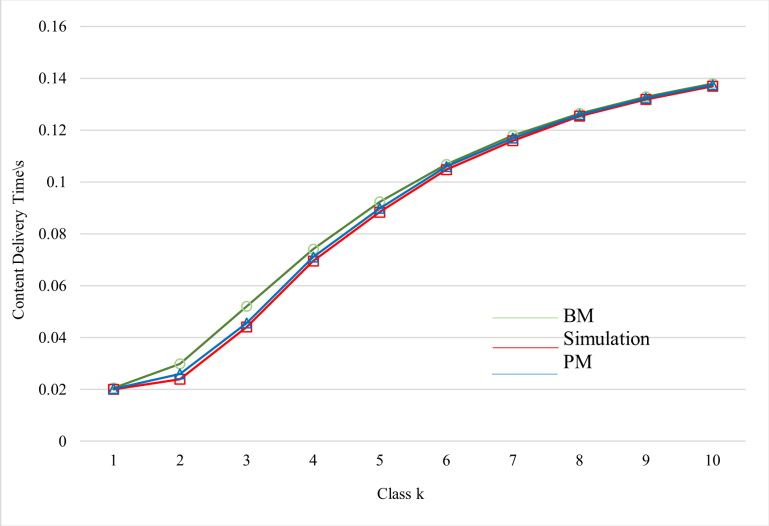
chunks. The content requests generated by mobile users are modeled by MMPP with total intensity $\lambda_{total} = 10\ content/sec$. Content popularity follows the Zipf distribution with the exponent parameter $\alpha = 2$. Jia et al. [84] indicates that the realistic Internet video content of YouTube follows Zipf distribution with exponent $\alpha = 2$. We assume each chunk with the same size of 1024 Bytes, which has good bandwidth savings result in ICN scenario [91]. In our experiment, we use the topology shown in Fig. 4.2. The propagation delay between two nodes has been set as 10 ms and the bandwitch capacity is allocated with 0.1 Gbps. Every router is equipped with a cache with size $c = 10^5$ chunks (0.1GBs) and implement an LRU replacement policy.

4.3.1 Validation

Based on ndnSIM, we run simulations under two extreme forwarding strategies. One of the forwarding strategies is that a new interest is only emitted after an existing interest has been satisfied, so we call it the single window strategy. Another forwarding strategy is to send all the interests of content at once, which we refer to as ”infinite window”. Under the same setting of parameters with simulations, we calculate the content delivery time using our PIT based model (PM) and the model in chapter 3 (BM), which does not consider the impact of PIT. Then we compare and analyze the numerical results of the two models and the simulation results. As depicted in Fig. 4.5 .Compared with the QM, our new model is closer to the simulated result, particularly for classes 2 and 3. The content of classes 2 and 3 are more likely to be aggregated in the PIT because they have a lower probability of being stored in the cache when compared with class 1 and a greater likelihood of being requested when compared with other classes. Comparing the content delivery time depicted in Fig. 4.5. There is a large difference in the delivery time of two different strategies. Content delivery time can be significantly decreased if the



(a) single window strategy

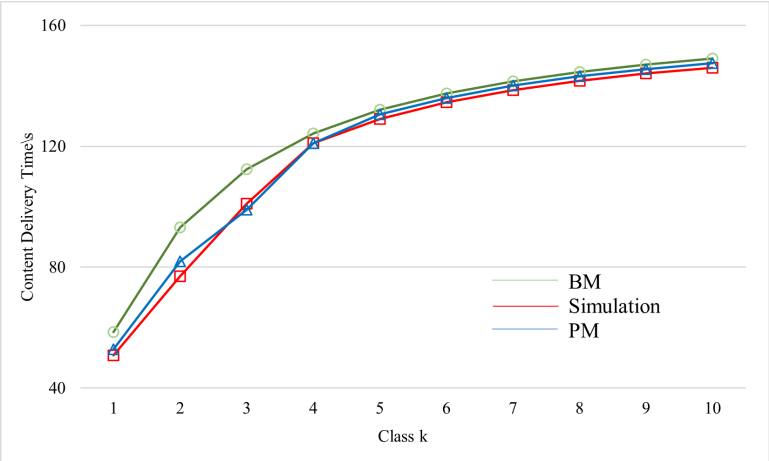


(b) infinite window strategy

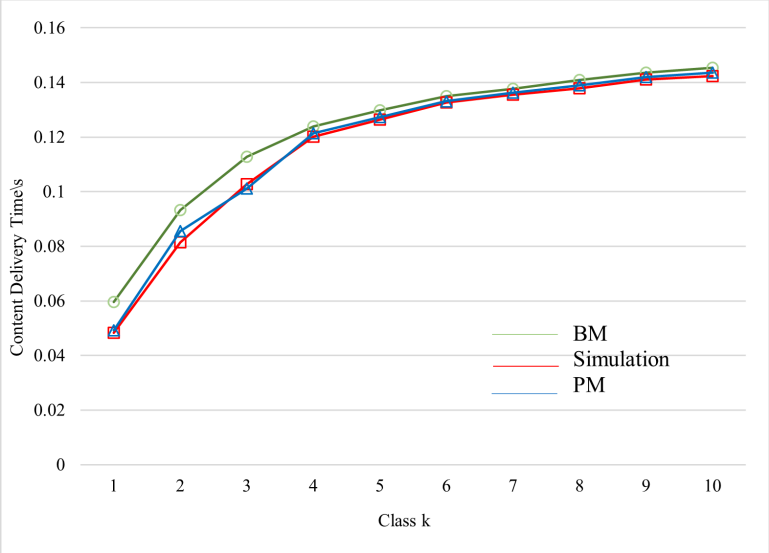
Figure 4.5: Average content delivery time with $\alpha = 2$

window size reaches infinity. However, it will cause network flooding if all the interests are emitted at once. Hence, a suitable forwarding strategy that is both within the tolerance of the network and provides efficient content delivery is important in ICN. In this regard, our model can serve as a foundation for further research in the forwarding strategy area.

The model can also be applied to different forms of content dissemination. A further type of short-form video content is user-generated content (UGC),whose popularity is calculated using the Zipf distribution, with parameter α typically between 1.2 and 2.0. [100, 101]. As shown in Fig. 4.6, since the α is decreasing, content requests can be more diverse, thereby resulting in more content being hit at PIT. Compared to the results of



(a) single window strategy



(b) infinite window strategy

Figure 4.6: Average content delivery time with $\alpha = 1.3$

BM, our model is more accurate with the decreasing of the popularity distribution α.

Taking both Fig. 4.5 and Fig. 4.6 into account, it is obvious that the window size has a big impact on content delivery performance. Ideally, the window size should be set equally to the number of chunks in the content. However, all the interests of the contents are sent out together may cause network congestion. In parallel, the kernel buffer of consumer could not be infinite. Therefore, it remains an open issue as to how to allow W vary over time through a sliding window control to achieve the most efficient content delivery in the future.

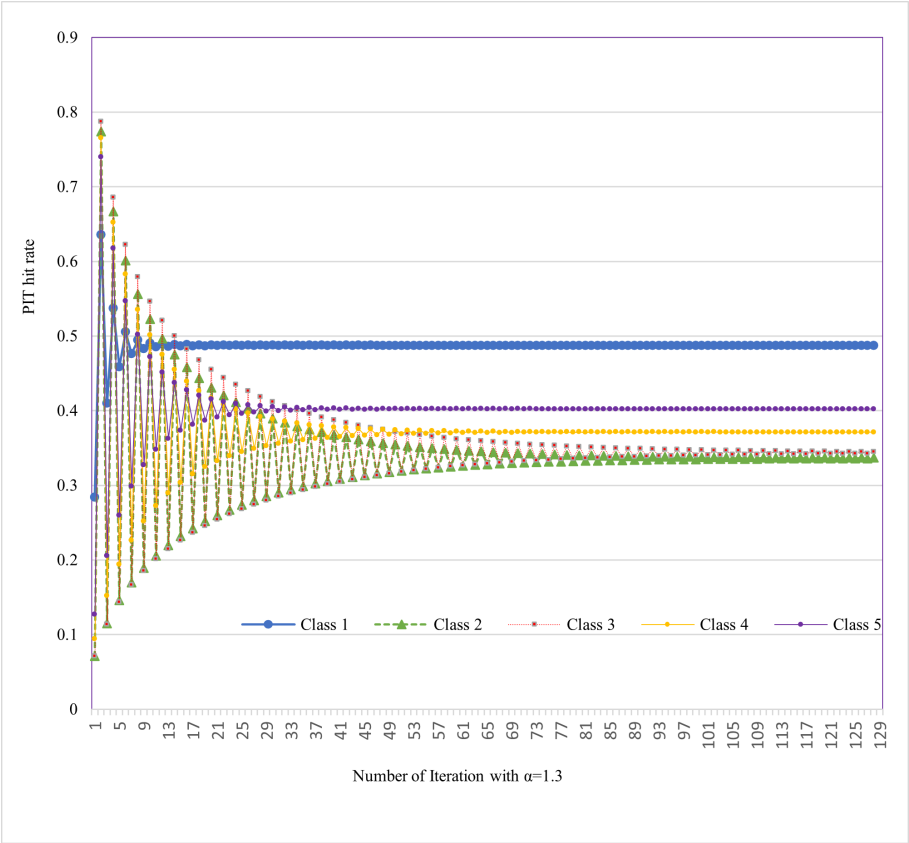


Figure 4.7: Number of iteration till convergence with $\alpha = 1.3$

4.3.2 Performance Analysis

In this subsection, we first use our proposed PIT miss rate calculation method to predict the impact of the Zipf exponent α on PIT miss rate, after that, the developed model is used as a cost-efficient tool to predict the impact of the number of different content M and cache size c on the performance of content transfer of ICN under bursty content requests.

Here, we demonstrate how long it takes PIT miss rate to reach convergence by our proposed method of PIT miss rate calculation. As shown in Fig. 4.7, for $\alpha = 1.3$, it takes more than 100 times of iteration to reach the convergence, while when $\alpha = 1.5$, which shown in Fig. 4.8, only need less than 30 times to reach convergence. $\alpha = 1.7$, as depicted in Fig. 4.9, just need 12 times to convergence. This illustrates

As far as the α of Zipf distribution influences the effectiveness for PIT miss rates to reach convergence. In Fig. 4.10, the impact of α on PIT miss rate is depitcted. It is

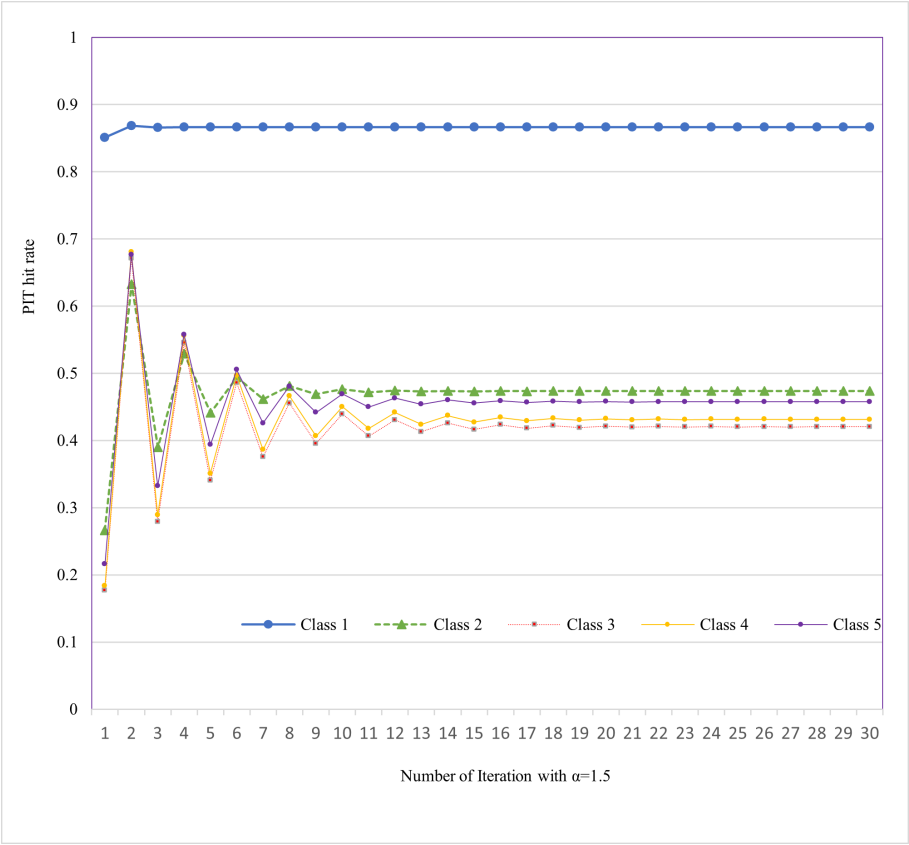


Figure 4.8: Number of iteration till convergence with $\alpha = 1.5$

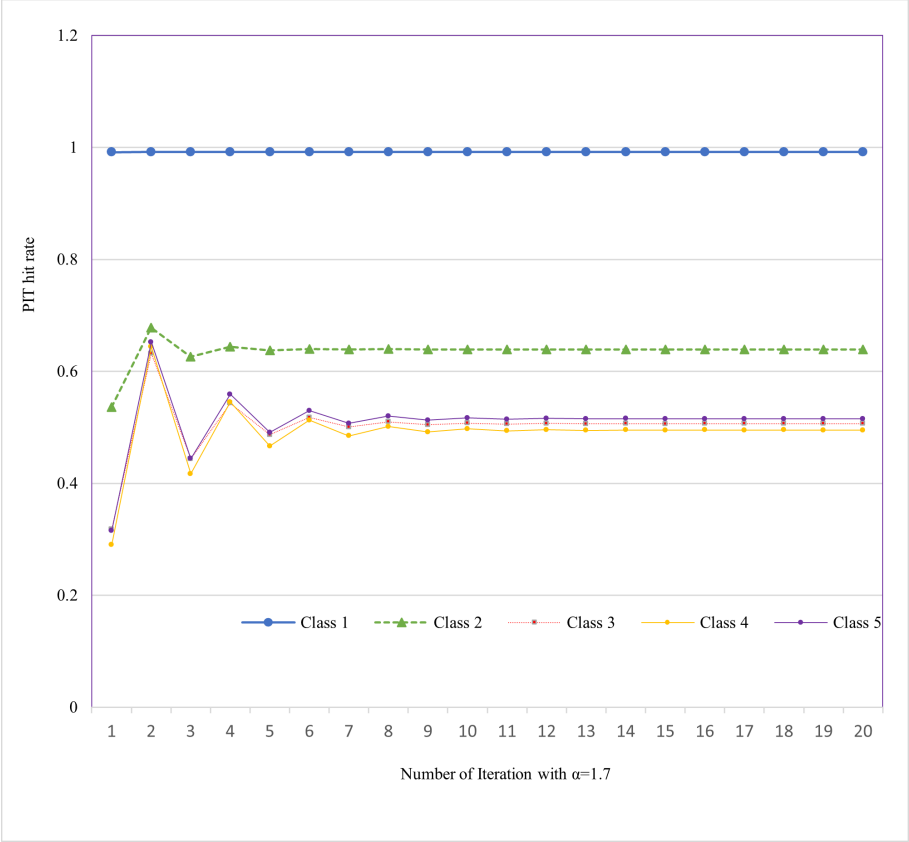


Figure 4.9: Number of iteration till convergence with $\alpha = 1.7$

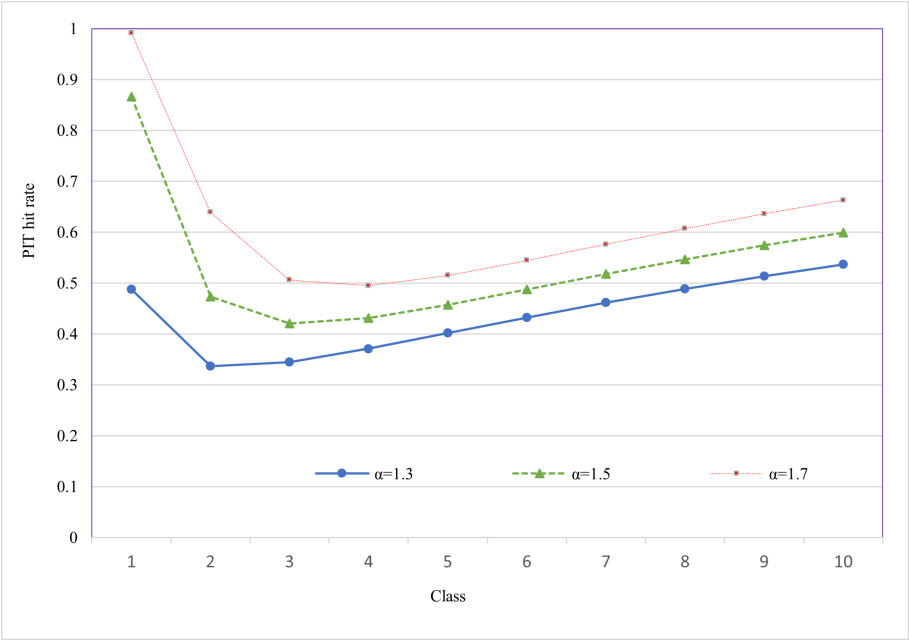


Figure 4.10: PIT miss rate for different Zipf distribution α

observed that PIT miss rates increase as α increases. High value α means the gap of distribution for different class is large, which leads to the interval of two consecutive request increases. Consequently, the probability that a new interest will not find a duplicate interest stored in the PIT increases.

In following work of performance analysis, α and window size have been allocated with $\alpha = 2.0$ and $W = 1$. The result cache size c influence on the performance of content delivery is shown in Fig. 4.11. When the cache size grows, the number of contents stored close to the consumer increases, improving the cache hit rate and decreasing the impact of request aggregation, and finally, reducing the delivery time of content.

The population of different contents M also plays an important role in predicting content delivery time. With $K = 10$ classes. The number of contents of each class m is adapted accordingly. The result is depicted in Fig. 4.12. When the number of contents of each class m grows, the number of contents stored close to the consumer decreases, reducing the cache hit rate, and finally, increasing the delivery time of content.

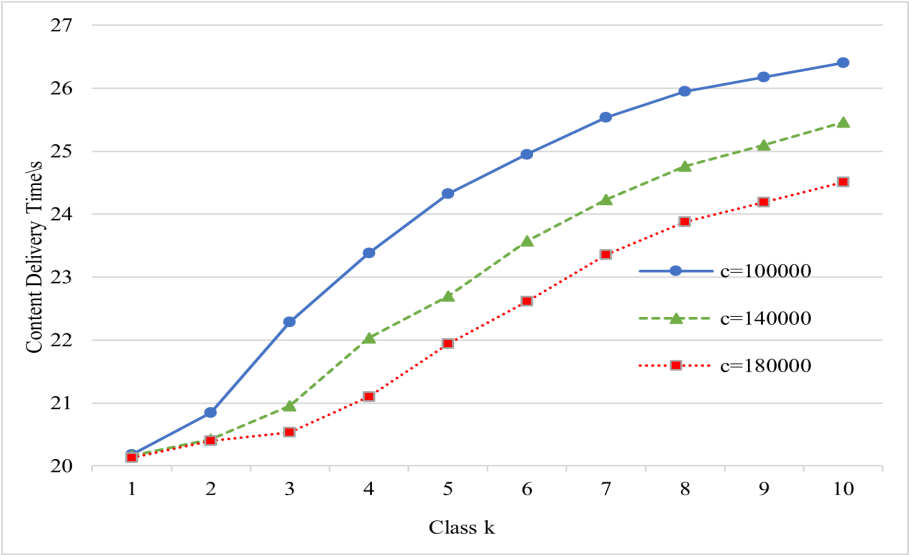


Figure 4.11: Content delivery time predicted by the model for different cache size c

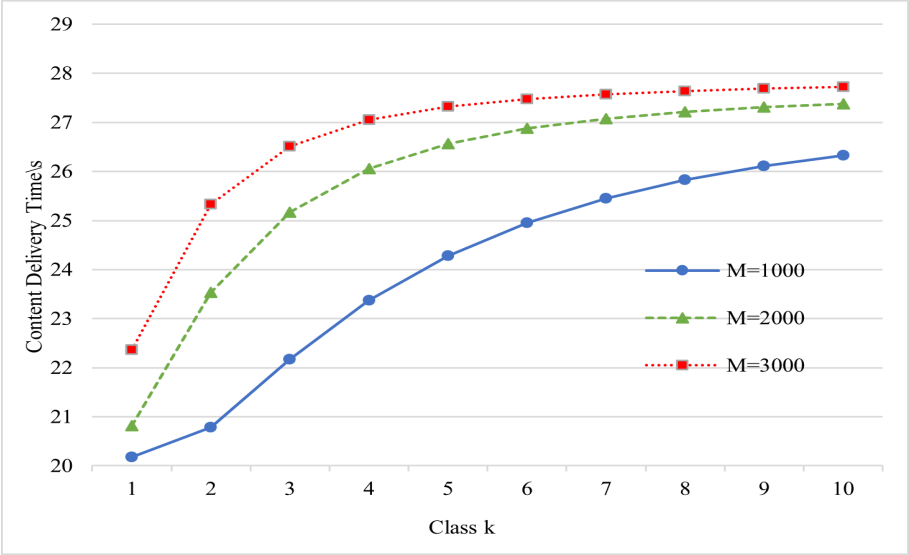


Figure 4.12: Content delivery time predicted by the model for different content population M

4.4 Summary

In this chapter, we have characterized the miss rate of PIT and investigated the impact of PIT on content delivery time under bursty requests in ICN. Since the complexity of calculating the content delivery time of the superposed MMPP grows exponentially, we approximate the multi-state MMPP with a 2-state MMPP to make computation easier. Our proposed model has been used to predict the influence of chunk transmission window size on content transfer performance. We also developed an analytical model to evaluate the mean service time based on consumer and provider mobility. In comparison with

the simulation results, the experimental results demonstrate the accuracy of our proposed model.

Chapter 5

Modelling Distribution of Content Delivery Time and Minimal Pending Interest table size in ICN

5.1 Introduction

The Pending Interest Table (PIT) is one of the essential components of content retrieval in ICN [76]. The function of PIT is to keep track of the interface of incoming interest (the request for chunk) to allow requested chunks to be returned back to the consumer. PIT can also aggregate requests for the same chunk to avoid chunk requests flooding. In order to ensure the proper operation of the PIT in ICN routers, the PIT is required to operate at wirespeed in the forwarding plane [75, 76, 76, 77, 79]. This makes PIT a costly resource, where the cost of an ICN router increases with its increasing size. Hence, for cost-saving purposes, allocating a very small size to the PIT seems to be a reasonable decision. However, ICN routers with inadequate PIT sizes may cause unacceptably high drop rates for interests arriving at the router. If the drop of interest occurs, the content delivery time will be adversely affected because the content needs to be retransmitted, thus increasing the

delay experienced by consumers. Thus, in order to ensure that interests drop rate less than the requirement, it is necessary to have an analytical model that determines the minimum PIT size. Such a model could be used to investigate the relationship between PIT size and content delivery time. As ICN holds the promise to become the next generation Internet architecture. The developed analytical model can be applied to industries as well. For example, a minimal PIT size could result in cost savings when deploying ICN routers by Internet service provider.

In ICN, a consumer retransmits the interest packet if the corresponding chunk packet is not received within a retransmission timeout (RTO) interval [65]. Like the calculation method of RTO in IP networks, it relies on the predicted Round Trip Time (RTT), which is the time interval between the transmission of the Interest packet and arrival of the corresponded content [65]. However, because of in-network caching, the transmission delay can vary significantly, which makes it harder to exactly estimate RTT [63]. To provide reliable delivery, a timer is encoded in each unsatisfied Interest packet, and requests are retransmitted when the timer expires [64, 65]. Our previous work as shown in chapter 3 and 4 has predicted the Virtual Round Trip Time (VRTT). VRTT, however, only represents the average time that elapses between dispatching the chunk request and receiving it, which is not appropriate for designing the interest timeout in ICN. It is more intuitive to design the interest timeout when you have the distribution of RTT. Therefore, to accurately predict the RTO in ICN, we propose an analytical model that exhibits the distribution of RTT, which we refer to as elapsed time distribution in this chapter.

The rest of this chapter is organised as follows: Section 5.2 derives the comprehensive analytical model to determine the minimum PIT size. An analytical model for the distribution of RTT in ICN is also developed in this section. The accuracy of the developed model is validated in Section 5.3through extensive simulation experiments. Finally,

Section 5.4 concludes this chapter.

5.2 Analytical Model of PIT Occupancy

In this section, we model the PIT occupancy in ICN router under bursty arrival to find the minimal PIT size and the distribution of interest elapsed time at consumer. In order to do this, the detailed description of our model is presented in the first subsection. After that, the input process at PIT will be introduced. Then, we derive an expression of mean service time. Finally, we derive the minimal PIT size and the elapsed time distribution of interest.

5.2.1 Model of PIT occupancy of router

In this subsection, a model of PIT occupancy of ICN router has been proposede. The detailed description of our model and the parameters used in this chapter can be found below.

In our model, we assume that the arrival of consumers generated interests follows Markov Modulated Poisson Process (MMPP), which is a doubly stochastic process. MMPP has been widely applied in modeling the bursty content requests of real communication systems since it could describe how arrival changes with time shifting while remaining computationally tractable. The details of interests arrival after filtering by the content store (CS) are discussed in section 5.2.2. We assume that the elapsed time between the dispatch of interest and the reception of corresponding data of interest follows an exponential distribution with mean μ. The detail of service time μ of the queuing model can be found in section 5.2.3. We define the size of our PIT as C interests, i.e., no more than C interests can be stored in the PIT table.

In existing research on PIT [80], additional buffers are not taken into account. On the basis of our analysis of previous work, the PIT system does not use an additional buffer

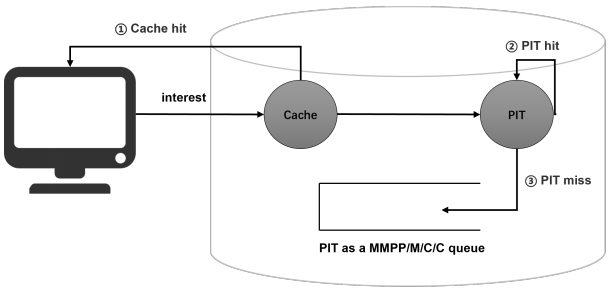


Figure 5.1: Traffic flow in ICN router

for the reason that router has limited resources and PIT requires to work at wirespeed. Thus, an extra buffer will significantly increase the cost of router. Furthermore, the experiment results indicate that even with an extra buffer at PIT, the content delivery time is approximately equivalent to the strategy of dropping and retransmitting the interest if PIT is full. Consequently, The maximum interests in our queuing model is equivalent to the PIT size C. With attributes have been defined, the PIT occupancy of ICN router can be modelled as a MMPP/M/C/C queue. To facilitate reading, the notations of system parameters are summarised in Table 5.1.

Table 5.1: SYSTEM PARAMETERS INVESTIGATED IN CHAPTER 5

Parameter	Meaning
C	PIT size (number of interest could be stored in PIT)
N	System size at PIT
K	Number of different popularity classes
δ	Average content size in number of chunks
α	Zipf exponent
ξ	The maximum desired interest drop rate
θ_i	The propagation delay from node $i-1$ to node i
π	Stationary probability Vector of the PIT
π_i	The steady state distribution of i interests in the system
λ_k	Content request rate for class k
$q_k(i)$	Popularity distribution for class k at node i
$p_k(i)$	Cache miss rate for class k at node i
$p_{pit,k})$	The PIT miss rate for content of class k at the first router
$P_k(i)$	Total Miss probability for class k at node i
$T_N(i)$	Queueing delay for one chunk experienced at node i
$\text{RVRTT}_k(i)$	The residual virtual round trip time of class k at node i
$E(T)$	Elapsed time distribution of interest at the consumer

5.2.2 Input process at PIT

In this subsection, we describe the arrival of interests after filtering by the content store and PIT aggregation. As shown in Fig. 5.1, when interest generated by consumers reaches the router, There are three different scenarios that can occur in ICN routers. The first scenario is when interest arrives at a router, the router first checks whether the chunk requested by the interest has been cached on the content store (cache) of the router. If a hit occurs, the router will return the corresponding chunk back. In the event of a miss, interest will meet the second scenario. As shown in Fig. 5.1. The interest will first be delivered to the PIT. Then the router looks up the PIT which stores the interests that are pending for service. If an interest is already stored in the PIT, then the new income of interest will update the information stored in the PIT to add the network interface of the incoming interest, and the new interest will not be forwarded to the next router. In the event that the new interest arrival finds no match in the PIT, it will jump to the third scenario of Fig.5.1. The router will add the information of the new interest into the PIT. A Forwarding Information Base (FIB) is then searched, which maps name prefixes to the next hop router, and finally, the interest is forwarded to the next router. A potential interest that could possibly arrive the PIT queueing system is one that has missed both the content store and PIT.

To model the realistic network, we consider M contents are equally divided into K classes and let the mean interest arrival rate be λ_k. Each class represents a specific type of service that contains $m = M/K$ different contents. Furthermore, contents belonging to class k are requested with the identical probability q_k following the Zipf distribution, where Zipf distribution is widely used for characterizing the popularity of Internet.

As MMPPs are superposed the state space of the resultant MMPP grows exponentially. The m state MMPP has been approxiamted by a 2-state MMPP based on the technique

of matching the first three noncentral moments of the instantaneous arrival rate of the MMPP, which is characterized by an infinitesimal generator Q_k of the underlying Markov process as well as the coefficient matrix Λ_k. Matrix Q_k and Λ_k are given by

$$Q_k = \begin{bmatrix} -\sigma_{12} & \sigma_{12} \\ \sigma_{21} & -\sigma_{21} \end{bmatrix}, \quad \Lambda_k = \begin{bmatrix} \lambda_{k1} & 0 \\ 0 & \lambda_{k2} \end{bmatrix}. \tag{5.1}$$

where σ_{12} indicates the transition rate from state 1 to state 2, and σ_{21} donates the transition rate from state 2 to 1. The request arrival rate is λ_{k1} when the Markov chain is in state 1 and request arrival rate is λ_{k2} when the Markov chain is in state 2. according to the definition of MMPP in cookbook [95], The steady-state vector π_k need to satisfy $\pi Q = 0$ and $\pi e = 1$, where $e = (1,1,...,1)^{\mathrm{T}}$ is the column vector. We also have the normalizing equation $\pi_1 + \pi_2 = 1$. On the basis, we have

$$\pi_1 = \frac{\sigma_{21}}{\sigma_{12} + \sigma_{21}}, \quad \pi_2 = \frac{\sigma_{12}}{\sigma_{12} + \sigma_{21}}. \tag{5.2}$$

then, the mean arrival rate of the 2-state MMPP arrival in steady state, λ_k is given by

$$\lambda_k = \lambda_{k1} \times \pi_1 + \lambda_{k2} \times \pi_2 = \frac{\sigma_{21}\lambda_{k1} + \sigma_{12}\lambda_{k2}}{\sigma_{12} + \sigma_{21}}. \tag{5.3}$$

The process of deriving the superposition of K individual 2-state MMPPs could be found in section 4.2.1. From the steady-state vector $\boldsymbol{\pi}$ and arrival rate vector $\boldsymbol{\lambda}$, it can derive the mean arrival rate λ_{total} of the composite MMPP as follows

$$\lambda_{total} = \boldsymbol{\pi}\boldsymbol{\lambda}. \tag{5.4}$$

Following the above analysis of superposed MMPP, we are able to model the interest

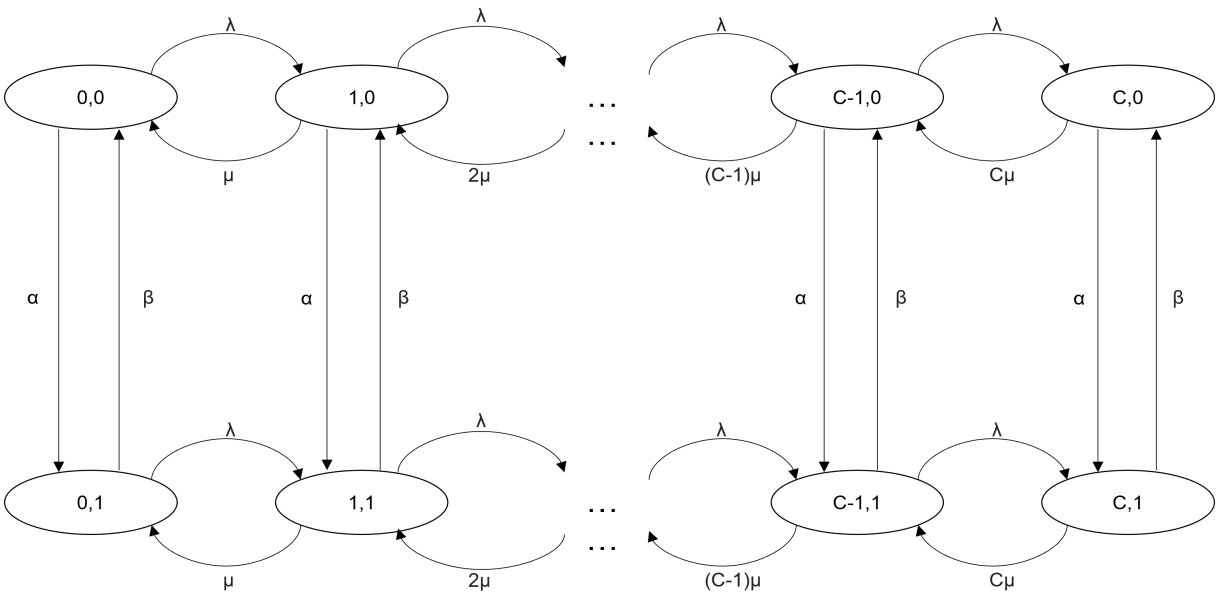


Figure 5.2: State transition diagram of MMPP(2)/M/C/C

arrival distribution at the content store. As described, a potential interest that could possibly arrive the PIT queueing system is one that has missed both the content store and PIT. Let p_k and $p_{pit,k}$ be the average cache miss rate and PIT miss rate for interests of class k. The detail of how to calculate average cache miss rate and PIT miss rate can be found in chapter 4. The input MMPP for class k at router has been be defined as $\{Q_k, \Lambda_k\}$. Then the arrival rate of PIT queueing system for class k, $\Lambda_{pit,k}$ can be derived as

$$\Lambda_{pit,k} = p_k p_{pit,k} \Lambda_k. \tag{5.5}$$

A superposition of MMPPs results in an exponential growth of the state space of the resultant MMPP. It is difficult to direct build up the transition rate matrix of the superposed MMPP. Hence, we implement the technique of Kathleen [97] to approximate the multistate state MMPP with a special case of the 2-state MMPP called Interrupted Poisson Process (IPP). The approach taken by Kathleen is to match the first three noncentral moments of the instantaneous arrival rate of the MMPP. As we already get the steady-state vector $\boldsymbol{\pi}$ infinitesimal generator Q and arrival rate vector $\boldsymbol{\lambda}$ of the superposed MMPPs

according to the above calculation. [99] give a brief introduction of how to approximate a superposed MMPPs into a IPP.

In this subsection, we derive the interest arrival distribution after filtering by the content store and PIT aggregation. Following the input process for the queueing model, we will discuss the average service time of the queueing model in the next subsection.

5.2.3 Average service time

The average service time for class k is the residual virtual round trip time of class k at the assessed router. The details of determining the residual virtual round trip time of class k are outlined in chapter 4. We briefly summary the methodology in this subsection.

Let $p_k(i)$ donate the cache miss probability for content of class k at the intermediate node i. Under the assumption that all caches adopt the Least Recently Used (LRU) cache replacement policy, $p_k(i)$ can be derived through the following expression

$$p_k(1) = e^{-\frac{\lambda}{m} q_k g(1) c_1^{\alpha}} = e^{-(\frac{C_1}{m\delta\Gamma(1-\frac{1}{\alpha})})^{\alpha}}. \tag{5.6}$$

where C_1 and δ represent the cache size of the first router and average content size in number of chunks. The PIT miss rate for content of class k at the first router $p_{pit,k}$ can be written as

$$p_{pit,k} = e^{-\frac{\lambda_k}{m}(1-p_k(1))\mathrm{RVRTT}_k(1)}. \tag{5.7}$$

We use $q_k(i)$ to represent the popularity distribution for class k at i^{th} router. Having $q_k(1)$ follow the Zipf distribution at the first router, the popularity distribution at the i^{th} router can be written as

$$q_k(i) = \frac{q_k p_{pit,k} \prod_{j=1}^{i-1} p_k(j)}{\sum_{l=1}^{K} q_l p_{pit,l} \prod_{j=1}^{i-1} p_l(j)}. \tag{5.8}$$

We can derive the probability of interest for class k miss on both the cache and PIT at router i as follows:

$$\mathrm{P}_k(i) = p_k(1)^{(\frac{c_i}{c_1})^{\alpha} p_{pit,k} \prod_{j=1}^{i-1} p_k(j)}, \qquad \forall i > 1. \tag{5.9}$$

The residual virtual round trip time RVRTT_k can be obtained by

$$\mathrm{RVRTT}_k(j) = \sum_{i=j}^{N} (T_N(i) + 2\theta_i) \prod_{l=1}^{i} \mathrm{P}_k(l-1). \tag{5.10}$$

where θ_i indicates the link delay from node $i-1$ to node i and $T_N(i)$ is the queueing delay for one chunk experienced at node i. As we have the popularity distribution at the i^{th} router $q_k(i)$, The average service time for interest at node i is given by

$$\frac{1}{\mu} = \sum_{i=1}^{K} \mathrm{RVRTT}_k(i) * q_k(i). \tag{5.11}$$

Based on the expression, we can discover that the average service time follows an exponential distribution

5.2.4 Minimal PIT size

In the subsection, based on the result of section 5.2.2, we use 2-state MMPP to describe the arrivial of interest after filtering by the content store and PIT aggregation. Two arrival states correspond to two Poisson process with arrival rate λ and 0 separately. Fig.5.2 shows the state transition of the MMPP(2)/M/C/C queueing system.

According to the diagram of state transition, the corresponding transition matrix Q can be given as

$$Q = \begin{bmatrix} q_0 & & & \\ & q_1 & & \\ & & \ddots & \\ & & & q_C \end{bmatrix}. \tag{5.12}$$

The sub-matrices q_0, q_1 and q_C are shown in Eq. (5.13).

$$\begin{aligned} q_0 &= \begin{bmatrix} -(\lambda+\alpha) & \alpha & \lambda & 0 \\ \beta & -\beta & 0 & 0 \end{bmatrix} \\ q_1 &= \begin{bmatrix} \mu & 0 & -(\lambda+\alpha+\mu) & \alpha & \lambda & 0 \\ 0 & \mu & \beta & -(\beta+\mu) & 0 & 0 \end{bmatrix} \\ q_c &= \begin{bmatrix} C\mu & 0 & -(\lambda+\alpha+C\mu) & \alpha & 0 & 0 \\ 0 & C\mu & \beta & -(\beta+C\mu) & 0 & 0 \end{bmatrix}. \end{aligned} \tag{5.13}$$

We assume π_i is the steady state distribution of i interests in the system. The steady state vector π_i need to satisfy

$$\begin{aligned} \pi Q &= 0 \\ \sum_{i=1}^{C} \pi_i \mathrm{e} &= 1. \end{aligned} \tag{5.14}$$

According to the detailed balance equation, the stationary distribution of the number of interests in the queueing system can be derived as

$$\pi_i \mathrm{e} = \frac{d}{i!}\left(\frac{\lambda\beta}{\alpha\mu+\beta\mu}\right)^i, \qquad i \le C. \tag{5.15}$$

where d is a constant satisfying $\sum_{i=1}^{C} \pi_i \mathrm{e} = 1$.

To derive the minimal size of PIT at router, we calculate the steady state vector π of

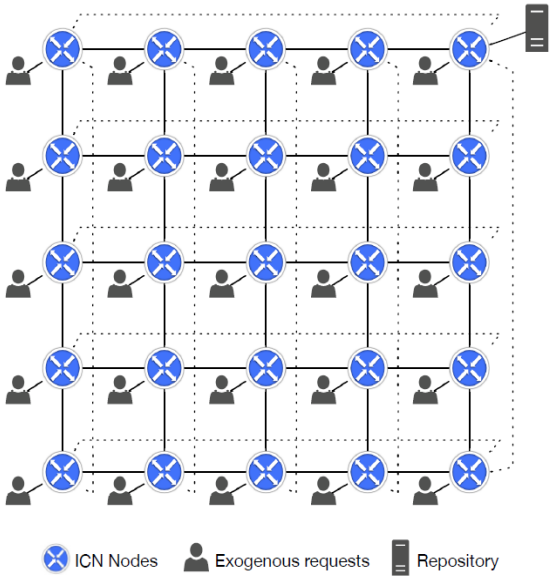


Figure 5.3: Network topology: 5×5 two-dimensional torus network with one repository connecting to a random node.

our MMPP/M/C/C queueing model. Based on the arrival theorem of queueing theory, π_C indicates that PIT is full and the incoming interests will be dropped. Accordingly, the drop rate at the PIT is π_C, Let the ξ represents the maximum desired interest drop rate. The minimal size of PIT is subject to

$$\pi_C \mathrm{e} < \xi, \qquad C \in \mathbb{N}_+. \tag{5.16}$$

In this subsection, we provide a methodology for finding the minimum PIT size at the router. Next, we will model the elapsed time distribution of interest at the consumer.

5.2.5 Elapsed time distribution

It is of interest to us to examine the elapsed time distribution of interest at the consumer because it can be used to represent the distribution of content delivery time. This model can be used to estimate the impact of network parameters on distribution of content delivery time. Researcher could utilize the elapsed time distribution of interest to assign a desirable interest timeout.

In the previous subsection, PIT occupancy is modeled as MMPP/M/C/C queueing at the router level. An extra buffer at the router level was not considered due to the fact that it would increase the cost of the ICN router without decreasing the average delivery time. On the consumer side, the cost of extra buffer at consumer side is not expensive. Moreover, when an interest finding PIT on the consumer side is full, allowing the interest to wait in the queue of the PIT is preferable to dropping it and retransmitting it. L and C indicate the interest buffer and PIT size at consumer respectively. Let $N = C + L$, we could model the PIT occupancy at consumer side as a MMPP/M/C/N queue. The transition matrix Q of the queueing model can be written as

$$Q = \begin{bmatrix} q_0 & & & & & & \\ & q_1 & & & & & \\ & & \ddots & & & & \\ & & & q_C & & & \\ & & & & q_{C+1} & & \\ & & & & & \ddots & \\ & & & & & & q_N \end{bmatrix}. \tag{5.17}$$

The difference between the transition matrix of MMPP/M/C/C and MMPP/M/C/N is

$$\begin{aligned} q_C &= \begin{bmatrix} C\mu & 0 & -(\lambda+\alpha+C\mu) & \alpha & \lambda & 0 \\ 0 & C\mu & \beta & -(\beta+C\mu) & 0 & 0 \end{bmatrix} \\ q_N &= \begin{bmatrix} C\mu & 0 & -(\lambda+\alpha+C\mu) & \alpha & 0 & 0 \\ 0 & C\mu & \beta & -(\beta+C\mu) & 0 & 0 \end{bmatrix} \\ q_i &= q_C, \qquad \forall i \in (C,N). \end{aligned} \tag{5.18}$$

With steady state distribution satisfy $\pi Q = 0$ and $\sum_{i=1}^{C} \pi_i \mathrm{e} = 1$, the steady state distri-

bution of the MMPP(2)/M//N can be written as

$$\pi_i e = \begin{cases} \frac{d}{i!}(\frac{\lambda\beta}{\alpha\mu+\beta\mu})^i, & i \leq C \\ \\ \frac{d}{C!C^{i-C}}(\frac{\lambda\beta}{\alpha\mu+\beta\mu})^i, & C < i \leq N. \end{cases} \tag{5.19}$$

After derived the steady state distribution of the queueing model, we start to analyze the elapsed time distribution of the interest. The arriving interest may encounter two mutually exclusive situations. In the first instance, when an interest arrives and the PIT is not full, the interest will be forwarded immediately and the information of the interest will be stored. Another scenario is when an interest arrives and the PIT is full, then the interest can be buffered as it arrives.

Let $X(t)$ and $Y(t)$ be the service and the waiting time of the CDFs when the state is k, respectively. As the queue is modeled as a $MMPP/M/C/N$ queue, $X(t)$ and $Y(t)$ follow the exponential and the Erlang distributions, respectively, and are given by

$$X(t) = 1 - ue^{ut} \tag{5.20}$$

$$Y(t) = 1 - \sum_{j=0}^{i-C} \frac{e^{-C\mu x}(C\mu x)^j}{j!} \tag{5.21}$$

Thereby, the elapsed time distribution of the interest $\mathrm{E}(t)$ can be written by

$$
\begin{aligned}
\mathrm{E}(t) &= \sum_{i=0}^{C-1} \pi_i \mathrm{eX}(t) + \sum_{i=C}^{N} \pi_i \mathrm{eP}(\mathrm{X}(t) \mid \mathrm{Y}(t)) \\
&= \sum_{i=0}^{C-1} \pi_i \mathrm{e}(1-e^{-\mu t}) + \\
&\quad \sum_{i=c}^{N} \pi_i \mathrm{e} \int_0^t (1 - \sum_{j=0}^{i-C} \frac{e^{-C\mu x}(C\mu x)^j}{j!}) \mu e^{-\mu(t-x)} dx \\
&= \sum_{i=0}^{N} \pi_i \mathrm{e}(1-e^{-\mu t}) - \sum_{i=C}^{N} \pi_i \mathrm{e} \sum_{j=0}^{i-C} \frac{(C\mu)^j \mu e^{-\mu t}}{(C\mu-\mu)^{j+1}} \cdot \\
&\quad \int_0^t (C\mu-\mu)^{k+1} \frac{x^j e^{-(C-1)\mu x}}{j!} dx \\
&= \sum_{i=0}^{N} \pi_i \mathrm{e}(1-e^{-\mu t}) - \sum_{i=c}^{N} \pi_i \mathrm{e}(\sum_{j=0}^{i-C} \frac{(C\mu)^j \mu e^{-\mu t}}{(C\mu-\mu)^{j+1}} \\
&\quad \cdot (1 - \sum_{a=0}^{j} \frac{(C\mu t-\mu t)^a}{a!} e^{-(C-1)\mu t}).
\end{aligned}
\tag{5.22}
$$

In this section, we model the PIT occupancy in ICN router as a MMPP/M/C/C queue, we also derive the elapsed time distribution of interest to help allocate an appropriate interest timeout. To demonstrate the accuracy of our model, we will discuss the validation of our model in the following section.

5.3 Validation and Performance Analysis

In this section, we evaluate the performance of our proposed PIT occupancy and elapsed time distribution of interest stay in the PIT of the consumer based on a modular NDN simulator ndnSIM [90], which developed under the ns-3 framework. This open-source simulator implements the CS, PIT and FIB data structures, and content retrieve operations of ICN. For all the simulation scenarios mentioned, we run the simulations 50 times and present our findings based on the mean value of the simulation results.

We consider a population of $M = 1000$ different contents. All the contents are evenly allocated in $K = 10$ classes, each data packet is 1024 bytes. Every router is equipped with a cache capable of storing 1 percent of all content available on the Internet. There are 5

consumers connected to the edge router. The content requests generated by consumers are modeled by MMPP with total intensity $\lambda_{total} = 7.5\ content/sec$. We adopt a two-dimensional 5×5 torus network as shown in Fig. 5.3. Torus is adopted as the topology since it can be formed into different topologies according to our needs. For example, when using deterministic routing, a torus topology can be modified into a cascade topology. In this chapter, we use the shortest path routing method in the torus topology. The link capacities and delays are set as follows. This link between the outermost node has a capacity of 0.1 Gbps and a 5 ms propagation delay. The link between the outermost and internal node has a capacity of 0.5 Gbps where propagation delay is 10 ms. The link between the internal node has a capacity of 1 Gbps and the propagation delay is 20 ms.

5.3.1 Validation

In this subsection, we first use our proposed model to estimate the minimal PIT size at the bottleneck router of the ICN network under different thresholds of drop rate ξ. In the simulation, after driving the minimal PIT size through our model, we set the derived value as the PIT size of the bottleneck router to record the interest drop rate in the simulation. 0.001, 0.01, 0.05, 0.1, 0.2 are the values chosen for ξ in the model. The results are depicted in Fig.5.4. The figure reveal that the analytical performance results closely match those obtained from the simulation experiments based on the thresholds of drop rate. The result also shows that the minimum PIT size increased as the requirement for ξ decreased.

We now evaluate the elapsed time distribution, The PIT size at consumer is 5, with extra buffer of 5 interests (L=5). As shown in Fig. 5.5, the delay distribution obtained from simulation is very close to the results of our model, validating the accuracy of the developed analytical model. Accordingly, the value of RTO can be allocated based on the acceptable drop rate of interest.

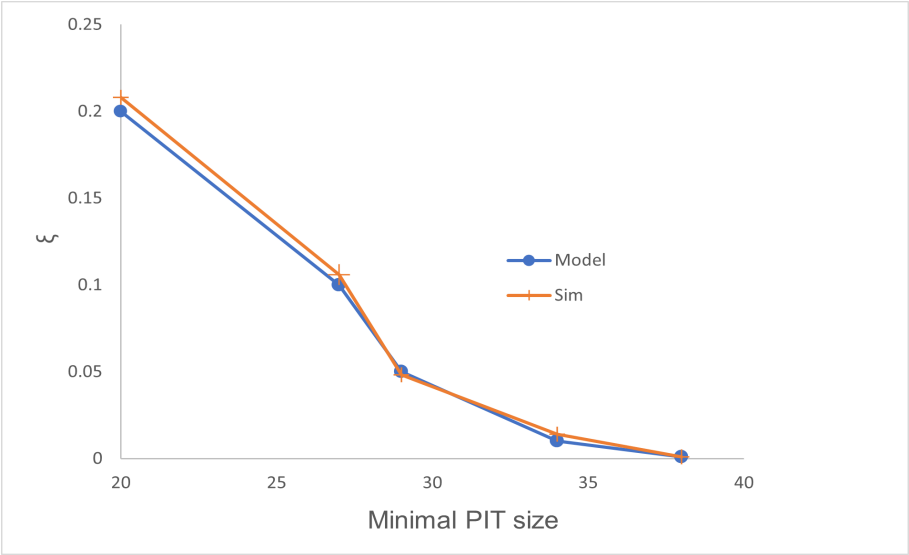


Figure 5.4: Effect of ξ on optimal PIT size

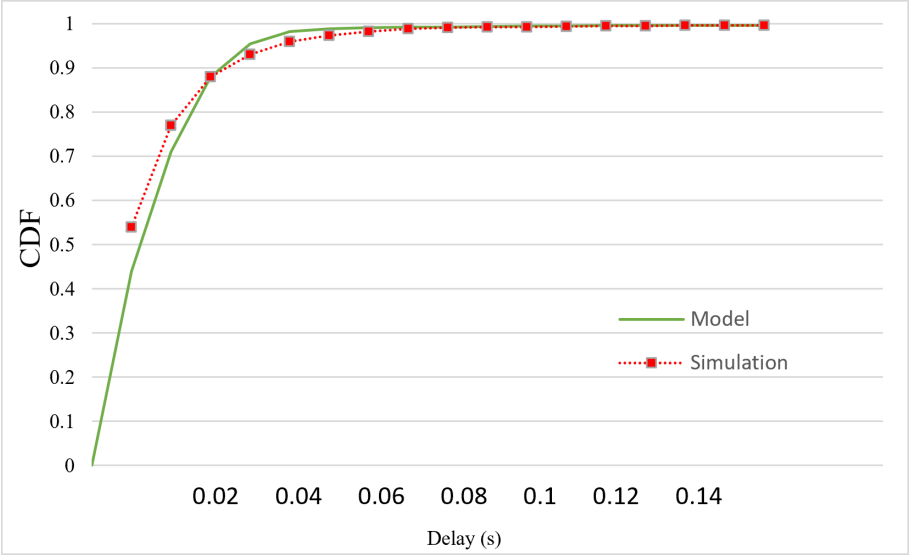


Figure 5.5: Elapsed time distribution

5.3.2 Performance Analysis

In this subsection, the developed model is used as a cost-efficient tool to estimate the impact of the network traffic λ_{total} and cache size l on the minimal size of PIT in the ICN network.

We first study the effect of varying request arrival rate on the minimal PIT size. The value of λ_{total} are 7.5, 10, 12.5, 15, 17.5 and 20. The value of ξ is set to 0.01 for all the cases. As depicted in Fig. 5.6, with the increase in the content arrival rate λ_{total}, the minimal PIT size has risen. This is due to the fact that high arrival of requests causes

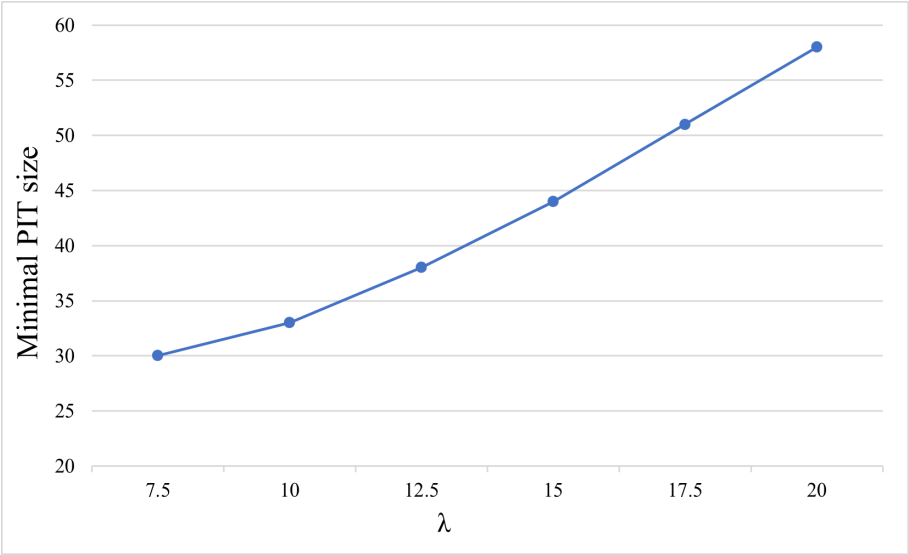


Figure 5.6: Effect of request arrival rate on the minimal PIT size

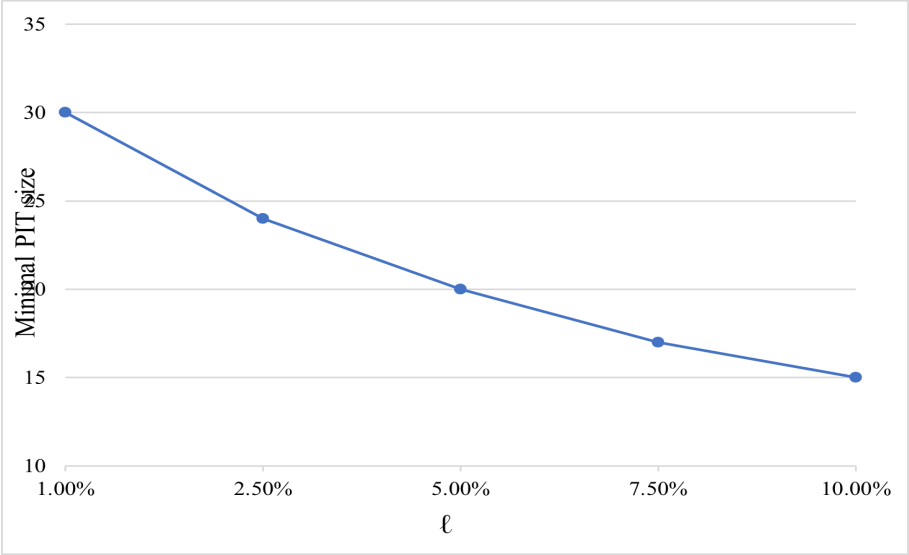


Figure 5.7: Effect of cache size on the minimal PIT size

more interest to need to be stored in PIT

We then examined the effect of cache size on the minimal PIT size. The value of cache are 1%, 2.5%, 5%, 7.5% and 10% of content universe. As shown in Fig. 5.7. The minimal PIT size of router decreased with the increase of cache size. It is due to the fact that only if the interest misses at the cache, it will be sent to the PIT. By increasing the cache size, the cache miss rate will decrease, thus resulting in fewer interests being sent to the PIT

5.4 Summary

In this chapter, we have modeled the PIT occupancy of ICN router under bursty arrival using a MMPP/M/C/C queue. The aggregation of PIT interests and filtering of content stores are taken into consideration to drive the arriving interest at PIT. The model for estimating the service time for interest has also been developed. Based on our proposed model, we present a methodology for determining the minimum PIT size at the router. Additionally, the elapsed time distribution of interest at the consumer was modeled to help determine the threshold of interest timeout. Moreover, the accuracy of our analytical model has been demonstrated by comparing it to simulated results.

Chapter 6

Conclusions and Future Work

In this thesis, the content transfer performance in ICN has been evaluated. In the academic research, having a unified and accurate analytical model of content delivery in ICN is imperative. Such a model could be exploited to evaluate the performance of different cache replacement policies and to monitor the Quality of Service (QoS) of services, which are important for the wide adoption of ICN. As ICN holds the promise to become the next generation Internet architecture. The developed analytical model can be applied to industries as well. For instance, short-form video platforms such as TikTok and Instagram Reels can use the model as tool to investigate the content delivery time based on different popularity. When the delivery time of different popularity of content is known, short-form video platforms can devise strategies for those videos whose delivery time exceeds the consumers' maximum waiting time. For example, when a short-form video platform requests extra time to deliver a video, it might invite customers to engage in a mini-game.

‘

6.1 Conclusions

This research work has presented new analytical models for performance evaluation of ICN content transfer in the presence of practical network environment for multimedia

Services. The accuracy of the proposed models has been validated through comparing the analytical results with those obtained from simulation experiments of an ICN simulator, ndnSIM. The developed analytical models have been used to explore the impact of the key networking metrics on the performance of ICN content transfer. Moverover, a new analytical model for PIT occupancy has been developed to determine the minimal PIT size. We also proposed an analytical model that exhibits the distribution of RTT to help design the timeout of the interest. The main contributions of this research are summarised as follows:

- To evaluate the performance of content transfer in ICN considering various content requests and service popularity, an analytical model has been developed by exploiting the queueing network theory. As one of the key components of content transfer delay, queueing delays have been taken into account in each intermediate router. The proposed model has been used to investigate the performance of content transfer under arbitrary network topology. Extensive ICN simulation experiments based on the ns-3 framework have been performed to validate the effectiveness and accuracy of the analytical model.

- A novel analytical model has been developed to evaluate the content transfer performance of ICN under bursty multimedia content requests. MMPP has been adopted by the model to capture the bursty characteristics of multimedia services. The PIT miss rate as the key performance index of ICN has been derived to accurately evaluate content delivery time. The proposed model has been used to examine the influence of chunk transmission window size on content transfer performance. The accuracy of the analytical model is validated through extensive simulation experiments.

- ICN is intrinsically compatible with wireless networks. To evaluate the mean service time of interest in wireless network of ICN, an analytical model has been developed. The model considers both the mobility of the consumer and the mobility of the provider in evaluating the chunk-level transfer performance.

- In order to determine the minimum PIT size at the router in accordance with the interest drop rate requirement, an analytical model of PIT occupancy has been developed. Additionally, the elapsed time distribution of interest at the consumer was modeled to help determine the threshold of interest timeout. Moreover, the accuracy of the analytical model has been demonstrated by comparing it to simulated results.

6.2 Future Work

The research of this thesis mainly concentrates on analysis of content transfer performance in wired and wireless ICN with heterogeneous multimedia traffic and arbitrary topology. As a panacea for content-oriented Internet service, ICN provides potential solutions to various problems in the current host-based communications. The main research outcomes achieved in the research can be extended to accommodate several interesting yet challenging research directions and can benefit other emerging networking technologies.

- Interest rate control proposed for CCN is initially a simple pipelining. If we keep the value sliding window control equal to one, where $W = 1$, the forwarding of a 1 MB file, however, still requires over 20 seconds as shown in chapter 4. Moreover, If the link delay increases, for example, content retrieval between two countries, the content delivery time will be influenced by the link delay more significantly. It may be opposed to ICN's quick retrieval feature. Currently, it remains an open issue how to let W vary over time through a sliding window control W.

- Experiment result shows that the capacity utilization of content store after the first three routers decreases significantly. Due to this, we aim to find a method for improving the utilization of the cache resources of routers except for the three front-facing routers. Considering that the content's name is transparent to the ICN, it provides the router with the opportunity to determine the popularity of the content based on the content's name. It is our goal to develop a statistical analysis system to analyze names with high popularity over a period of time and within a geographic region. The system could be implemented on edge computing resources or in a data center. Delivering the analytical results to each router periodically in order to inform them of the popularity of content, thereby enabling the router to make decisions regarding which content should be cached. A router may also employ a hybrid strategy for replacing caches.

- There is no clear consensus regarding the use of hierarchical or flat names. Hierarchical names can be human-readable and are easier to aggregate in principle, but it is unclear whether they can scale to Internet levels without turning into DNS names due to aggregation. Conversely, flat names are easier to manage, they do not impose processing requirements for longest prefix matching, they can be self-certifying, and they can be handled with highly scalable structures such as DHTs, but it is uncertain whether DHTs can provide satisfactory performance. It is also unclear how the router's FIB could contain such a large table since all names of content need be stored. The bottleneck of ICN structure needs to be addressed to rekindle the research interest of ICN among the research institutes.

- A notable trend in wireless networks is the deployment of cache-enabled small base stations (SBSs) to offload traffic from macro base stations (BSs). There has

been much attention paid to caching content at small BSs in order to enhance the quality of service for users and to alleviate congestion in the backhaul connection [102–104]. Nevertheless, this heterogeneous wireless infrastructure may have one shortcoming: small BSs are connected to the core network through low-capacity, unreliable backhaul links because of high density and cost constraints. Therefore, these small cell solutions alone will not be sufficient to meet the QoS requirements associated with peak traffic volumes.

There is a promising solution that can enhance the accessibility of service contents by storing the most popular contents in the local caches of SBSs. However, selecting the content that should be cached in the limited storage space available at the SBSs is an NP-hard problem. In [105], a mobility-aware heuristic solution has been proposed to address the NP-hard optimisation problem of maximising the caching hit ratio of mobile users. Additionally, in the current work, the popularity distribution is assumed to be well known. In practice, such assumption is not justified, thus learning-based approaches have been proposed to estimate the popularity distribution in [102, 106]. Since the estimation process requires real-time response and is computationally intensive, it poses new challenges. Thus, intelligent caching placement strategies based on learning-based approaches are noteworthy for improving the quality of service.

- An outstanding advantage of ICN is the ability to deploy ubiquitous caching across unreliable networks in order to increase resilience. However, ICN will pose significant challenges in terms of content discovery when it is used in mobile networks, particularly MANETs and DTNs [62]. Compared with host-centric routing, ICN-based wireless environments are characterized by frequent routing churn due to

caching and replacement behavior of network contents. In order to address the unpredictable content relocation problem in a mobile network, we must design and implement an efficient name resolution and routing mechanism. In the process, the caching of unpopular contents should be properly tackled. Because the potential gains may be unsatisfactory by using more sophisticated structured routing policies (e.g., GHT) to deal with unpopular contents [107].

References

[1] Van Jacobson, Diana K Smetters, James D Thornton, Michael F Plass, Nicholas H Briggs, and Rebecca L Braynard. Networking named content. In *Proceedings of the 5th international conference on Emerging networking experiments and technologies*, pages 1–12, 2009.

[2] George Xylomenos, Christopher N Ververidis, Vasilios A Siris, Nikos Fotiou, Christos Tsilopoulos, Xenofon Vasilakos, Konstantinos V Katsaros, and George C Polyzos. A survey of information-centric networking research. *IEEE communications surveys & tutorials*, 16(2):1024–1049, 2013.

[3] Salahadin Seid Musa, Marco Zennaro, Mulugeta Libsie, and Ermanno Pietrosemoli. Convergence of information-centric networks and edge intelligence for iov: Challenges and future directions. *Future Internet*, 14(7):192, 2022.

[4] Naveen Kumar and Naveen Kumar Gupta. Comparing the performance of tcp/ip with named data networking using ns-3. In *Proceedings of the 2022 Fourteenth International Conference on Contemporary Computing*, pages 560–563, 2022.

[5] Sumit Kumar and Rajeev Tiwari. An efficient content placement scheme based on normalized node degree in content centric networking. *Cluster Computing*, 24(2):1277–1291, 2021.

[6] Cisco. Cisco visual networking index: global mobile data traffic forecast update, 2017–2022. 2019.

[7] Xiaolong Jin and Geyong Min. Performance analysis of priority scheduling mechanisms under heterogeneous network traffic. *Journal of Computer and System Sciences*, 73(8):1207–1220, 2007.

[8] Seyyed Naser Seyyed Hashemi and Ali Bohlooli. Analytical modeling of multi-source content delivery in information-centric networks. *Computer Networks*, 140:152–162, 2018.

[9] Oussama Serhane, Khadidja Yahyaoui, Boubakr Nour, and Hassine Moungla. A survey of icn content naming and in-network caching in 5g and beyond networks. *IEEE Internet of Things Journal*, 8(6):4081–4104, 2020.

[10] Andriana Ioannou and Stefan Weber. A survey of caching policies and forwarding mechanisms in information-centric networking. *IEEE Communications Surveys & Tutorials*, 18(4):2847–2886, 2016.

[11] Ghassan Jaber, Natallia Pastei, Fatima Rahal, and Ahmad Abboud. Naming and routing scheme for data content objects in information-centric network. In *2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, pages 1–5. IEEE, 2020.

[12] Mihaela Ion, Jianqing Zhang, and Eve M Schooler. Toward content-centric privacy in icn: Attribute-based encryption and routing. In *Proceedings of the 3rd ACM SIGCOMM workshop on Information-centric networking*, pages 39–40, 2013.

[13] Nikolai Leshov, Muhammad Azfar Yaqub, Muhammad Toaha Raza Khan, Sungwon Lee, and Dongkyun Kim. Content name privacy in tactical named data net-

working. In *Proceedings of 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 570–572. IEEE, 2019.

[14] The size of the world wide web (the internet). [EB/OL]. `http://www.worldwidewebsize.com/line-spacing-in-latex-documents/` Accessd July 4, 2022.

[15] Ali Ghodsi, Scott Shenker, Teemu Koponen, Ankit Singla, Barath Raghavan, and James Wilcox. Information-centric networking: seeing the forest for the trees. In *Proceedings of the 10th ACM Workshop on Hot Topics in Networks*, pages 1–6, 2011.

[16] Matteo D'Ambrosio, Christian Dannewitz, Holger Karl, and Vinicio Vercellone. Mdht: A hierarchical name resolution service for information-centric networks. In *Proceedings of the ACM SIGCOMM workshop on Information-centric networking*, pages 7–12, 2011.

[17] Jia Zhou, Yanhua Li, Vijay Kumar Adhikari, and Zhi-Li Zhang. Counting youtube videos via random prefix sampling. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 371–380, 2011.

[18] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 1–14, 2007.

[19] Geetanjali Rathee, Ashutosh Sharma, Rajiv Kumar, Farhan Ahmad, and Razi Iqbal. A trust management scheme to secure mobile information centric networks. *Computer Communications*, 151:66–75, 2020.

[20] Mauro Tortonesi, Marco Govoni, Alessandro Morelli, Giulio Riberto, Cesare Stefanelli, and Niranjan Suri. Taming the iot data deluge: An innovative information-centric service model for fog computing applications. *Future Generation Computer Systems*, 93:888–902, 2019.

[21] Nitul Dutta, Shobhit K Patel, Osama S Faragallah, Mohammed Baz, and Ahmed Nabih Zaki Rashed. Caching scheme for information-centric networks with balanced content distribution. *International Journal of Communication Systems*, 35(7):e5104, 2022.

[22] Boubakr Nour, Hakima Khelifi, Hassine Moungla, Rasheed Hussain, and Nadra Guizani. A distributed cache placement scheme for large-scale information-centric networking. *IEEE Network*, 34(6):126–132, 2020.

[23] Quang Ngoc Nguyen, Jiang Liu, Zhenni Pan, Ilias Benkacem, Toshitaka Tsuda, Tarik Taleb, Shigeru Shimamoto, and Takuro Sato. Ppcs: A progressive popularity-aware caching scheme for edge-based cache redundancy avoidance in information-centric networks. *Sensors*, 19(3):694, 2019.

[24] Thavavel Vaiyapuri, Velmurugan Subbiah Parvathy, V Manikandan, N Krishnaraj, Deepak Gupta, and K Shankar. A novel hybrid optimization for cluster-based routing protocol in information-centric wireless sensor networks for iot based mobile edge computing. *Wireless Personal Communications*, pages 1–24, 2021.

[25] Marzieh Sadat Zahedinia, Mohammad Reza Khayyambashi, and Ali Bohlooli. Fog-based caching mechanism for iot data in information centric network using prioritization. *Computer Networks*, 213:109082, 2022.

[26] Yayuan Tang, Kehua Guo, Jianhua Ma, Yutong Shen, and Tao Chi. A smart caching

mechanism for mobile multimedia in information centric networking with edge computing. *Future generation computer systems*, 91:590–600, 2019.

[27] Li Zeng, Hong Ni, and Rui Han. An incrementally deployable ip-compatible-information-centric networking hierarchical cache system. *Applied Sciences*, 10(18):6228, 2020.

[28] Faria Khandaker, Sharief Oteafy, Hossam S Hassanein, and Hesham Farahat. A functional taxonomy of caching schemes: Towards guided designs in information-centric networks. *Computer Networks*, 165:106937, 2019.

[29] Dapeng Man, Qi Lu, Hanbo Wang, Jiafei Guo, Wu Yang, and Jiguang Lv. On-path caching based on content relevance in information-centric networking. *Computer Communications*, 176:272–281, 2021.

[30] Sen Wang, Jun Bi, Jianping Wu, Zhaogeng Li, Wei Zhang, and Xu Yang. Could in-network caching benefit information-centric networking? In *Proceedings of the 7th Asian Internet Engineering Conference*, pages 112–115, 2011.

[31] Yating Yang and Tian Song. Energy-efficient cooperative caching for information-centric wireless sensor networking. *IEEE Internet of Things Journal*, 9(2):846–857, 2021.

[32] Giovanna Carofiglio, Massimo Gallo, and Luca Muscariello. Joint hop-by-hop and receiver-driven interest control protocol for content-centric networks. *ACM SIGCOMM Computer Communication Review*, 42(4):491–496, 2012.

[33] S Salsano, A Detti, M Cancellieri, M Pomposini, and N Blefari-Melazzi. Receiver-driven interest control protocol for content-centric networks. In *Proceeding of ACM SIGCOMM Workshop on Information Centric Networking (ICN)*, 2012.

[34] Yongmao Ren, Jun Li, Shanshan Shi, Lingling Li, and Guodong Wang. An explicit congestion control algorithm for named data networking. In *2016 IEEE conference on computer communications workshops (INFOCOM WKSHPS)*, pages 294–299. IEEE, 2016.

[35] Lorenzo Saino, Cosmin Cocora, and George Pavlou. Cctcp: A scalable receiver-driven congestion control protocol for content centric networking. In *2013 IEEE international conference on communications (ICC)*, pages 3775–3780. IEEE, 2013.

[36] Yaogong Wang, Natalya Rozhnova, Ashok Narayanan, David Oran, and Injong Rhee. An improved hop-by-hop interest shaper for congestion control in named data networking. *ACM SIGCOMM Computer Communication Review*, 43(4):55–60, 2013.

[37] Massimo Gallo, Bruno Kauffmann, Luca Muscariello, Alain Simonian, and Christian Tanguy. Performance evaluation of the random replacement policy for networks of caches. *Performance Evaluation*, 72:16–36, 2014.

[38] Mostafa Dehghan, Bo Jiang, Ali Dabirmoghaddam, and Don Towsley. On the analysis of caches with pending interest tables. In *Proceedings of the 2nd ACM Conference on Information-Centric Networking*, pages 69–78, 2015.

[39] Jiang Liu, Guoqing Wang, Tao Huang, Jianya Chen, and Yunjie Liu. Modeling the sojourn time of items for in-network cache based on lru policy. *China Communications*, 11(10):88–95, 2014.

[40] Riheng Jia, Zhe Liu, Xiong Wang, Xiaoying Gan, Xinbing Wang, and Jun Jim Xu. Modeling dynamic adaptive streaming over information-centric networking. *IEEE Access*, 4:8362–8374, 2016.

[41] Yusung Kim and Ikjun Yeom. Performance analysis of in-network caching for content-centric networking. *Computer networks*, 57(13):2465–2482, 2013.

[42] Michele Mangili, Fabio Martignon, and Antonio Capone. Performance analysis of content-centric and content-delivery networks with evolving object popularity. *Computer Networks*, 94:80–98, 2016.

[43] Guoqiang Zhang, Yang Li, and Tao Lin. Caching in information centric networking: A survey. *Computer networks*, 57(16):3128–3141, 2013.

[44] Haozhe Wang, Geyong Min, Jia Hu, Hao Yin, and Wang Miao. Caching of content-centric networking under bursty content requests. In *Proceedings of 2014 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 2522–2527. IEEE, 2014.

[45] Yongmao Ren, Jun Li, Lingling Li, Shanshan Shi, Jiang Zhi, and Haibo Wu. Modeling content transfer performance in information-centric networking. *Future Generation Computer Systems*, 74:12–19, 2017.

[46] Asanga Udugama, Sameera Palipana, and Carmelita Goerg. Analytical characterisation of multi-path content delivery in content centric networks. In *Proceedings of 2013 Conference on Future Internet Communications (CFIC)*, pages 1–7. IEEE, 2013.

[47] Predrag R Jelenković and Xiaozhu Kang. Characterizing the miss sequence of the lru cache. *ACM SIGMETRICS Performance Evaluation Review*, 36(2):119–121, 2008.

[48] Giovanna Carofiglio, Massimo Gallo, and Luca Muscariello. On the performance

of bandwidth and storage sharing in information-centric networks. *Computer Networks*, 57(17):3743–3758, 2013.

[49] Ikram Ud Din, Suhaidi Hassan, Muhammad Khurram Khan, Mohsen Guizani, Osman Ghazali, and Adib Habbal. Caching in information-centric networking: Strategies, challenges, and future research directions. *IEEE Communications Surveys & Tutorials*, 20(2):1443–1474, 2017.

[50] Rui Hou, Shuo Zhou, Mengtian Cui, Lingyun Zhou, Deze Zeng, Jiangtao Luo, and Maode Ma. Data forwarding scheme for vehicle tracking in named data networking. *IEEE Transactions on Vehicular Technology*, 70(7):6684–6695, 2021.

[51] Boubakr Nour, Spyridon Mastorakis, Rehmat Ullah, and Nicholas Stergiou. Information-centric networking in wireless environments: Security risks and challenges. *IEEE wireless communications*, 28(2):121–127, 2021.

[52] Wenjie Li, Sharief MA Oteafy, and Hossam S Hassanein. Rate-selective caching for adaptive streaming over information-centric networks. *IEEE Transactions on Computers*, 66(9):1613–1628, 2017.

[53] Benjamin Rainer, Daniel Posch, and Hermann Hellwagner. Investigating the performance of pull-based dynamic adaptive streaming in NDN. *IEEE Journal on Selected Areas in Communications*, 34(8):2130–2140, 2016.

[54] Boubakr Nour, Kashif Sharif, Fan Li, Song Yang, Hassine Moungla, and Yu Wang. ICN publisher-subscriber models: Challenges and group-based communication. *IEEE Network*, 33(6):156–163, 2019.

[55] Jacques Samain, Giovanna Carofiglio, Luca Muscariello, Michele Papalini, Mauro Sardara, Michele Tortelli, and Dario Rossi. Dynamic adaptive video streaming:

Towards a systematic comparison of ICN and TCP/IP. *IEEE transactions on Multimedia*, 19(10):2166–2181, 2017.

[56] Sunghwan Ihm and Vivek S Pai. Towards understanding modern web traffic. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 295–312, 2011.

[57] Somaya Arianfar, Pekka Nikander, and Jörg Ott. Packet-level caching for information-centric networking. In *ACM SIGCOMM, ReArch Workshop*, volume 4, 2010.

[58] Giovanna Carofiglio, Massimo Gallo, Luca Muscariello, and Diego Perino. Modeling data transfer in content-centric networking. In *2011 23rd International Teletraffic Congress (ITC)*, pages 111–118. IEEE, 2011.

[59] George Xylomenos, Xenofon Vasilakos, Christos Tsilopoulos, Vasilios A Siris, and George C Polyzos. Caching and mobility support in a publish-subscribe internet architecture. *IEEE Communications Magazine*, 50(7):52–58, 2012.

[60] Yongqiang Huang and Hector Garcia-Molina. Publish/subscribe in a mobile environment. *Wireless Networks*, 10(6):643–652, 2004.

[61] Gareth Tyson, Nishanth Sastry, Ruben Cuevas, Ivica Rimac, and Andreas Mauthe. A survey of mobility in information-centric networks. *Communications of the ACM*, 56(12):90–98, 2013.

[62] Gareth Tyson, Nishanth Sastry, Ivica Rimac, Ruben Cuevas, and Andreas Mauthe. A survey of mobility in information-centric networks: Challenges and research directions. In *Proceedings of the 1st ACM workshop on Emerging Name-Oriented*

Mobile Networking Design-Architecture, Algorithms, and Applications, pages 1–6, 2012.

[63] Longzhe Han, Seung-Seok Kang, Hyogon Kim, and Hoh Peter In. Adaptive retransmission scheme for video streaming over content-centric wireless networks. *IEEE Communications Letters*, 17(6):1292–1295, 2013.

[64] Marica Amadeo, Claudia Campolo, and Antonella Molinaro. Design and analysis of a transport-level solution for content-centric vanets. In *Proceedings of 2013 IEEE International Conference on Communications Workshops (ICC)*, pages 532–537. IEEE, 2013.

[65] Marica Amadeo, Claudia Campolo, and Antonella Molinaro. Enhancing content-centric networking for vehicular environments. *Computer Networks*, 57(16):3222–3234, 2013.

[66] Vasilios A Siris, Xenofon Vasilakos, and George C Polyzos. Efficient proactive caching for supporting seamless mobility. In *Proceeding of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks 2014*, pages 1–6. IEEE, 2014.

[67] Vasilios A Siris, Xenofon Vasilakos, and George C Polyzos. A selective neighbor caching approach for supporting mobility in publish/subscribe networks. In *Proceedings of ERCIM Workshop on eMobility. Held in conjunction with WWIC*, page 63. Citeseer, 2011.

[68] Pengcheng Jiang, Yuehui Jin, Tan Yang, Joost Geurts, Yaning Liu, and Jean Charles Point. Handoff prediction for data caching in mobile content centric network. In

Proceedings of 2013 15th IEEE International Conference on Communication Technology, pages 691–696. IEEE, 2013.

[69] Do-hyung Kim, Jong-hwan Kim, Yu-sung Kim, Hyun-soo Yoon, and Ikjun Yeom. End-to-end mobility support in content centric networks. *International Journal of Communication Systems*, 28(6):1151–1167, 2015.

[70] Jihoon Lee, Sungrae Cho, and Daeyoub Kim. Device mobility management in content-centric networking. *IEEE Communications Magazine*, 50(12):28–34, 2012.

[71] Do-hyung Kim, Jong-hwan Kim, Yu-sung Kim, Hyun-soo Yoon, and Ikjun Yeom. Mobility support in content centric networks. In *Proceedings of the second edition of the ICN workshop on Information-centric networking*, pages 13–18, 2012.

[72] Dookyoon Han, Munyoung Lee, Kideok Cho, Ted Taekyoung, Yanghee Choi, et al. Publisher mobility support in content centric networks. In *Proceedings of the International Conference on Information Networking 2014 (ICOIN2014)*, pages 214–219. IEEE, 2014.

[73] Guido Appenzeller, Isaac Keslassy, and Nick McKeown. Sizing router buffers. *ACM SIGCOMM Computer Communication Review*, 34(4):281–292, 2004.

[74] Jingcao Hu, Umit Y Ogras, and Radu Marculescu. System-level buffer allocation for application-specific networks-on-chip router design. *IEEE Transactions on Computer-Aided Design of integrated circuits and systems*, 25(12):2919–2933, 2006.

[75] Xiong Wang, Wei Wang, Chunhui Zeng, Rui Dai, Sheng Wang, and Shizhong Xu. Reducing the size of pending interest table for content-centric networks with

hybrid forwarding. In *Proceedings of 2016 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2016.

[76] Matteo Varvello, Diego Perino, and Leonardo Linguaglossa. On the design and implementation of a wire-speed pending interest table. In *Proceedings of 2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 369–374. IEEE, 2013.

[77] Haowei Yuan and Patrick Crowley. Scalable pending interest table design: From principles to practice. In *Proceedings of IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pages 2049–2057. IEEE, 2014.

[78] Zhaogeng Li, Jun Bi, Sen Wang, and Xiaoke Jiang. Compression of pending interest table with bloom filter in content centric network. In *Proceedings of the 7th International Conference on Future Internet Technologies*, pages 46–46, 2012.

[79] Huichen Dai, Bin Liu, Yan Chen, and Yi Wang. On pending interest table in named data networking. In *2012 ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS)*, pages 211–222. IEEE, 2012.

[80] Diego Perino and Matteo Varvello. A reality check for content centric networking. In *Proceedings of the ACM SIGCOMM workshop on Information-centric networking*, pages 44–49, 2011.

[81] Amuda James Abu, Brahim Bensaou, and Ahmed M Abdelmoniem. A markov model of ccn pending interest table occupancy with interest timeout and retries. In *Proceedings of 2016 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2016.

[82] Giuseppe Rossini and Dario Rossi. A dive into the caching performance of content centric networking. In *2012 IEEE 17th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pages 105–109. IEEE, 2012.

[83] Wei Koong Chai, Diliang He, Ioannis Psaras, and George Pavlou. Cache "less for more" in information-centric networks. In *Proceedings of International conference on research in networking*, pages 27–40. Springer, 2012.

[84] Adele Lu Jia, Siqi Shen, Dongsheng Li, and Shengling Chen. Predicting the implicit and the explicit video popularity in a user generated content site with enhanced social features. *Computer Networks*, 140:112–125, 2018.

[85] Dapeng Man, Yao Wang, Hanbo Wang, Jiafei Guo, Jiguang Lv, Shichang Xuan, and Wu Yang. Information-centric networking cache placement method based on cache node status and location. *Wireless Communications and Mobile Computing*, 2021:1–13, 2021.

[86] Kelvin HT Chiu, Jason Min Wang, Ahmed M Abdelmoniem, and Brahim Bensaou. A two-tiered caching scheme for information-centric networks. In *2021 IEEE 22nd International Conference on High Performance Switching and Routing (HPSR)*, pages 1–6. IEEE, 2021.

[87] Seyyed Naser Seyyed Hashemi and Ali Bohlooli. Analytical characterization of cache replacement policy impact on content delivery time in information-centric networks. *International Journal of Communication Systems*, 32(18):e4154, 2019.

[88] Thomas E Stern and Anwar I Elwalid. Analysis of separable markov-modulated

rate models for information-handling systems. *Advances in Applied Probability*, pages 105–139, 1991.

[89] Olivier Brun and Jean-Marie Garcia. Analytical solution of finite capacity m/d/1 queues. *Journal of Applied Probability*, pages 1092–1098, 2000.

[90] Spyridon Mastorakis, Alexander Afanasyev, and Lixia Zhang. On the evolution of ndnsim: An open-source simulator for ndn experimentation. *ACM SIGCOMM Computer Communication Review*, 47(3):19–33, 2017.

[91] Diego Perino, Matteo Varvello, and Krishna PN Puttaswamy. Icn-re: redundancy elimination for information-centric networking. In *Proceedings of the Second Edition of the ICN workshop on Information-Centric Networking*, pages 91–96, 2012.

[92] I Cisco. Cisco visual networking index: Forecast and methodology, 2013–2018. *CISCO White paper*, 2018, 2013.

[93] Ioannis Psaras, Richard G Clegg, Raul Landa, Wei Koong Chai, and George Pavlou. Modelling and evaluation of ccn-caching trees. In *Proceedings of International Conference on Research in Networking*, pages 78–91. Springer, 2011.

[94] Stefan Podlipnig and Laszlo Böszörmenyi. A survey of web cache replacement strategies. *ACM Computing Surveys (CSUR)*, 35(4):374–398, 2003.

[95] Wolfgang Fischer and Kathleen Meier-Hellstern. The markov-modulated poisson process (mmpp) cookbook. *Performance Evaluation*, 18(2):149–171, 1993.

[96] Alexander Graham. *Kronecker products and matrix calculus with applications*. Courier Dover Publications, 2018.

[97] Kathleen S Meier-Hellstern. The analysis of a queue arising in overflow models. *IEEE Transactions on Communications*, 37(4):367–372, 1989.

[98] Yiannis Thomas, Nikos Fotiou, Stavros Toumpis, and George C Polyzos. Improving mobile ad hoc networks using hybrid ip-information centric networking. *Computer Communications*, 156:25–34, 2020.

[99] Vignesh Sivaraman, Dibyajyoti Guha, and Biplab Sikdar. Optimal Pending Interest Table Size for ICN With Mobile Producers. *IEEE/ACM Transactions on Networking*, 28(4):1615–1628, 2020.

[100] Christine Fricker, Philippe Robert, James Roberts, and Nada Sbihi. Impact of traffic mix on caching performance in a content-centric network. In *2012 Proceedings IEEE INFOCOM Workshops*, pages 310–315. IEEE, 2012.

[101] Muhammad Aminul Islam. *Popularity Characterization and Modelling for User-generated Videos*. PhD thesis, University of Saskatchewan, 2013.

[102] Pol Blasco and Deniz Gündüz. Learning-based optimization of cache content in a small cell base station. In *Proceedings of 2014 IEEE international conference on communications (ICC)*, pages 1897–1903. IEEE, 2014.

[103] BN Bharath, Kyatsandra G Nagananda, and H Vincent Poor. A learning-based approach to caching in heterogenous small cell networks. *IEEE Transactions on Communications*, 64(4):1674–1686, 2016.

[104] Sabrina Müller, Onur Atan, Mihaela van der Schaar, and Anja Klein. Context-aware proactive content caching with service differentiation in wireless networks. *IEEE Transactions on Wireless Communications*, 16(2):1024–1036, 2016.

[105] Yang Guan, Yao Xiao, Hao Feng, Chien-Chung Shen, and Leonard J Cimini. Mobicacher: Mobility-aware content caching in small-cell networks. In *Proceedings of 2014 IEEE Global Communications Conference*, pages 4537–4542. IEEE, 2014.

[106] Pol Blasco and Deniz Gündüz. Multi-armed bandit optimization of cache content in wireless infostation networks. In *2014 IEEE International Symposium on Information Theory*, pages 51–55. IEEE, 2014.

[107] Matteo Varvello, Ivica Rimac, Uichin Lee, Lloyd Greenwald, and Volker Hilt. On the design of content-centric manets. In *Proceedings of 2011 Eighth International Conference on Wireless On-Demand Network Systems and Services*, pages 1–8. IEEE, 2011.