The Hong Kong Polytechnic University
Department of Computing

Design of Scalable and Efficient Information Retrieval Systems

by

Xiaocui Sun

A thesis submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

June, 2011

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

______________________ (Signature)

______________________ (Name of Student)

Abstract

As the fast growing of the Internet, scalability and efficiency become the critical issues for future Internet design. Caching is an effective way to improve the scalability and efficiency for information retrieval systems. As the media applications such as YouTube and Facebook explosively increase, video caching has gained significant attention recently. By caching video chunks at the proxy servers close to end users, the server bandwidth requirement can be significantly reduced, and hence it can greatly improve the system performance.

Ethernet is a widely used layer 2 technology for information retrieval system. Recently, Ethernet has gained popularity to be deployed in Metropolitan Area Networks (MANs). Metro Ethernet is highly attractive mainly due to its cost effectiveness, minimal management and maintenance, easy interoperability and convenience. Clearly, deploying the same technology, such as Ethernet, in both MAN and Local Area Networks (LAN) segments can potentially reduce the complexity and cost in network design and management, and hence improve packet forwarding performance. Even though Ethernet has many advantages, to be deployed in MANs, it has to solve the scalability issues. Ethernet has poor scalability due to using a flat addressing scheme (i.e., non-hierarchical Media Access Control (MAC) addresses) and broadcasting based address resolution scheme. In a MAN composing of a large number of LAN segments, a node in the provider network needs to keep a potential large number of MAC-to-port-mapping entries for frame forwarding. This may either cause MAC learning table explosion or excessive frame flooding, thus significantly degrading the system performance. Moreover, a broadcast-based address resolution scheme in Metro Ethernet introduces a large amount of broadcast messages into the provider network. Many protocols such as Address Resolution Protocol (ARP) and Dynamic Host Configuration Protocol (DHCP) use the broadcast service as a service discovery mechanism. The broadcast based address resolution schemes make Ethernet extremely convenient and easily accomplished. However, for MANs with millions of end users, high frequency broadcast messages waste a lot of bandwidth for address resolution. Frequently broadcast frames also accelerate the replacement of the table entries in Ethernet switches, which may lead to forwarding table explosion and hence trigger excessive frame flooding. Moreover, every end user has to take resource to handle every broadcast message.

This thesis aims at developing efficient and scalable information retrieval systems: to achieve this goal, an optimized cache replacement/group algorithm for video system has been designed to improve the scalability of the video system; To make Metro Ethernet scalable and efficient, an enabled Cache effect on Forwarding Table (CFT) scheme, an End user enabled Mac-in-Mac (EMiM) encapsulation scheme and two distributed Registration based address resolution Protocol (RP) are proposed in this thesis.

Firstly, an Optimized Cache Replacement (OCR) scheme for video streaming is designed. OCR groups the users into different cache groups based on their request patterns. It calculates the user density among all possible intervals, and then selects the maximized user density group to cache. Based on the optimized scheme in a single cache, we also extend OCR for cooperative cache. The simulation results show that OCR can increase the hit ratio and reduce the server load.

Secondly, we describe the CFT scheme in Metro Ethernet. CFT learns the IP and MAC mapping pair in a frame and eliminates the subsequent broadcast frames asking for this mapping. To further reduce the forwarding table size of Provider Edge (PE) node, CFT can be cooperated with EMiM. The forwarding table entries in Customer Edge (CE) node and PE node in this scheme are modified to learn both the IP and MAC addresses. By receiving a frame, the CE and PE nodes cache the IP-MAC address mapping carried in the frame. Once the mapping is recorded, it can be served for its subsequent requests by searching it in the cache. Hence, the CE and PE nodes not only response for forwarding but also answer the ARP request. The proposed architecture is easy to be accomplished and fully backward compatible. The simulation results show that the proposed scheme can decrease both the communication messages for address resolution and forwarding table size in PE nodes.

Thirdly, an EMiM encapsulation scheme is proposed. In the proposed scheme, a user's MAC address as well as its PE node's MAC address are associated with its ARP entry. The modified ARP entry allows an end user to do Mac-in-Mac (MiM) encapsulation directly by adding the destination user's MAC address and PE node's MAC address in the frame when it initiates a session. Hence a PE node does not need to maintain the entries of mapping end user's MAC address to PE node's MAC address, thus significantly reducing the forwarding table size. The simulation results show that the proposed scheme could provide high scalability by reducing up to 65% maximum forwarding table size in PE nodes.

Finally, we propose two RPs. In a RP, multiple ARP registers are allocated to support address resolution. Each IP address has a home register which stores its ARP entry. When an end user moves to another location but keeps its IP address, its current PE or CE node is considered to be its foreign register. A foreign register temporally caches the ARP entry for an immigrated user and is in charge of the ARP entry updating in the home register. The IP address is used as an index to locate the corresponding home register through unicast, thus eliminating the broadcast to solve an unknown address. The proposed schemes can save more than 60% messages for address resolution and reduces up to 80% forwarding table size in PE nodes.

Publication

Journal Papers

1. **X. Sun**, and Z. Wang, ”An Efficient and Scalable Metro-Ethernet Architecture”, *International Journal of Future Generation Communication and Networks*, v3(4), pp. 25-42, 2010.

2. **X. Sun**, and Z. Wang, ”An Efficient Caching on Forwarding Table Scheme for Metro Ethernet”, *International Journal of Advanced Science and Technology*, v23, pp 10-26, 2010.

3. **X. Sun**, and Z. Wang, ”An Optimized Cache Replacement (OCR) scheme for video cache”, *Submitted to Computer Networks*.

Conference Papers

1. **X. Sun**, and Z. Wang, ”Enable Cache Effect on Forwarding Table in Metro-Ethernet”, *Advanced Communication and Networking*, 2010.

2. **X. Sun**, and Z. Wang, ”An End User Enabled MAC-in-MAC Encapsulation Scheme for Metro-Ethernet”, *IEEE International Symposium on Parallel and Distributed Processing with Applications*, 2008.

Acknowledgement

First, I would like to thank my supervisor, Dr. Zhijun Wang, for his rigorous supervision of my research. I thank him for his support, patience and encouragement during my Ph.D. study. He unremittingly trained me to be a good researcher. He taught me how to find research issues and how to solve problems. Time after time, he showed me how to express my ideas and write academic papers. His vision, passion, and attitude towards the research deeply affected me. Without his help and support, this body of work would not have been possible. What I have learned and experienced during the time I spent in his research group will benefit me much in the future.

Next, I thank my co-supervisor, Prof. Jiannong Cao, for his great support and generous help during my Ph.D. study. He set a good example for me as accuracy and strict method. Also, I thank Dr. Hao Che for his great help and wise advices during my Ph.D. study. I wish to acknowledge my appreciation to Dr. Lei Yie, Dr. Yi Xie, Yi Lai, Xike Xie, Yang Liu, Yi Wang, Binbin Zhou, Kunfeng Lai and Yi Yuan, who shared with me the pleasure of the Ph.D. study at the Hong Kong Polytechnic University. Furthermore, I would like to thank all my teachers from whom I learned so much in my long journey of formal education. They are Dr. Zhili Shao, Dr. Bin Xiao at the Hong Kong Polytechnic University, and many others.

Finally, but most significantly, I thank my families for their continuous love, support, trust, and encouragement through the whole trip of my life. Without them, none of this would have happened.

Contents

List of Figures

List of Tables

List of Abbreviations

AAA	Authentication, Authorization, Accounting
ADMs	Add/Drop Multiplexers
ARP	Address Resolution Protocol
ATM	Asynchronous Transfer Mode
BRAS	Broadband Remote Access Server
CFT	an enabled Cache effect on Forwarding Table scheme
CE	Customer Edge
CNs	Core Nodes
CCR	Cooperative Cache Replacement
CLRU	Cooperative LRU
CRP	Customer edge based distributed Registration address resolution Protocol
DHCP	Dynamic Host Configuration Protocol
DSL	Digital Subscriber Line
DACS	Digital Access Cross-connect Systems
DM cache	Direct-Mapped Cache
DEMAC	Destination End user’s MAC address
DPMAC	Destination PE node’s MAC
EMiM	End user enabled Mac-in-Mac
E-LINE	Ethernet Line
E-TREE	Ethernet Tree
E-LAN	Ethernet Lan
EPL	Ethernet Private Line
EVPL	Ethernet Virtual Private Line
EPON	Ethernet Passive Optical Network
EVC	Ethernet Virtual Circuit
EMAC	End user’s MAC address
EU	EMiM User
IEEE	Institute of Electrical and Electronical Engineers

LAN	Local Area Networks
LRU	Least Recently Used
LSRs	Label Switch Routers
LSPs	Label Switched Paths
LER	Label Edge Router
LFU	Least Frequent Used
LAA	Locally Administered Address
MANs	Metropolitan Area Networks
MiM	Mac-in-Mac
MEF	Metro Ethernet Forum
MAC	Media Access Control
MTU	Maximum Transmission Unit
MPLS	Multi-Protocol Label Switching
MiM	MAC-in-MAC
MSTP	Multiple STP
MAT	MAC Address Translation
MLRU	Modified LRU
OSI	Open System Interconnection
OCR	Optimized Cache Replacement scheme
OUI	Organizational Unique Identifier
OCR	Optimized Cache Replacement
PE	Provider Edge
PVC	Permanent Virtual Circuit
PRP	Provider edge based distributed Registration address resolution Protocol
PMAC	PE node’s MAC
Qos	Quality of Service
RPRs	Resilient Packet Rings
RP	Registration based address resolution Protocol
RSTP	Rapid STP
STP	Spanning Tree Protocol

SONET	Synchronous Optical Networking
SDH	Synchronous Digital Hierarchy
SCMAC	Source Customer Edge MAC
SEMAC	Source EMAC
SPMAC	Source PMAC
TDM	Time Division Multiplexed
TU	Traditional end User
UNIs	User Network Interfaces
VLAN	Virtual Local Area Network
VoD	Video-on-Demand
VC	Virtual Circuits
WANs	Wide Area Networks
1-GigE	1-gigabit Ethernet
10-GigE	10-gigabit Ethernet

Chapter 1

Introduction

As the explosive growth of the Internet, scalability and efficient become important for an information retrieval system. A successful system should be highly scalable and efficiency.

Caching is an effective method to improve scalability and efficiency for information retrieval systems. Specially, video caching has been popular in recent years as the increase of various video applications such as online courses, news distribution and entertainment distribution. In order to save the server work load, caching chunks of popular videos can significantly reduce the server load. Due to limited cache space, how to replace chunks is critical for cache performance. This thesis proposes an Optimized Cache Replacement (OCR) scheme for video cache to maximize the cache performance.

Ethernet is a dominant influential technology for information retrieval systems. As the dominant layer 2 access technology, Ethernet is now considered as a candidate technology for metropolitan area. Four novel Metro Ethernet technologies, including an enabled Cache effect on Forwarding Table (CFT), an End user enabled Mac-in-Mac (EMiM) encapsulation scheme and two distributed Registration based address resolution Protocols (RP) are proposed. The proposed protocols take the advantage of Ethernet and achieve high performance in traffic load.

In this chapter, we first describe the video caching. Then, we introduce the general Ethernet, including the development, the frame format, the forwarding basic, the advantages and services. We also discuss the advantages of Metro Ethernet and describe the most challenging issues faced to Metro Ethernet. Finally, we summarize the contributions and outline the organization of the thesis.

1.1 Video Caching

Now video streaming applications have gained increasingly popularity as the come of the popular web sites such as Youtube. With the exponential increment of the users, how to present real-time video streaming services becomes a critical issue. Caching is considered as one of the important means to improve the performance of the video system. Copies of the video chunks can be recorded at proxy servers or servers close to end users. The subsequent requests from the users could be handled by searching a nearby cache instead of retrieving the same chunk from the original server. This not only reduces network bandwidth consumption and the retrieve latency but also the server load. The crucial issue for a caching mechanism is to employ an effective cache replacement algorithm which could achieve high cache hit ratio, as a cache miss leads to long response latency and high overhead. Video-on-Demand (VoD) streaming caching is different from the normal web caching as the future requests are more predictable, and the data items need to be cached are huge. Hence, it is necessary to develop new cache schemes for video cache.

Least Recently Used (LRU) is the most common cache replacement algorithm. The entry accessed longest time is replaced by the new comer, so as to keep the most recently used entries in the cache. Though LRU has many advantages such as easy to implement and robust, it has several limitations. It has low efficiency. Many researches are conducted to improve the performance of LRU. However, the cache replacement algorithm is focused on highlighting the utilization of individual data items in a single cache, which may introduce complex computation for every data item's caching.

1.2 General Ethernet

Ethernet was not the only technology considered in Local Area Network (LAN) operating system. A battle between Ethernet and IBM's Token Ring occurred to compete for LAN superior between late 1980s and mid 1990s. Benefiting from higher data transmission rate, lower cost and less complexity management, Ethernet survived while the Token-Ring technology left the stage. Today, over 90 percent of all LANs worldwide are dominated by Ethernet topology [1] [2] [3] [4] [5]. Dominating in the LAN for a long time, Ethernet has emerged into the Metropolitan Area Network (MAN), where it could be a substitution or a complementation to a host of alternative technologies. A series of important transformations have changed Ethernet into a much faster, robust, function-rich technology that is competitive for dominance in the metro market.

1.2.1 Ethernet forwarding basic

The data packet format transmitted in Ethernet is called frame. The so called preamble or the trailing frame check sequence which shows start bits is essential for all physical hardware and is contained in all frame types. The maximum size of the Ethernet frame is defined as the Maximum Transmission Unit (MTU). The standard size of MTU is 1526 Bytes, which could carry 1500 Bytes data payload maximally. Data payload larger than this is separated into smaller pieces and encapsulated in different Ethernet frames. Ethernet Version 2 frame (Figure 1.1) is the most common one nowadays, as it is usually directly used by the Internet Protocol.

This part takes printing as an example to illustrate the forwarding in a LAN. If a

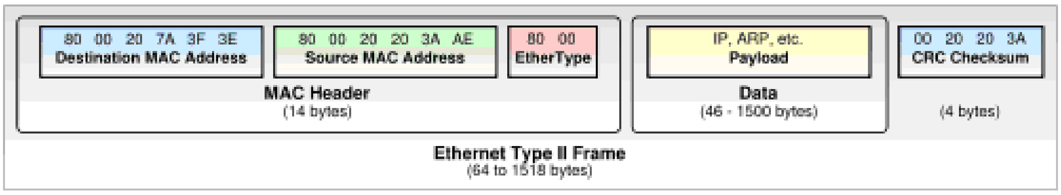


Figure 1.1: The most common Ethernet Frame format, type II (source: Wikipedia)

computer intends to use a printer in the same LAN, an Ethernet frame based on this computer's print application request is generated. In this Ethernet frame, the source address is filled by the Media Access Control (MAC) address of the computer and the destination address is filled by the printer's MAC address. This frame is transmitted either by half-duplex or full-duplex mode. When the frame is received by the printer, after checking the destination MAC address of the frame and confirming itself is the destination, the request is accepted and the corresponding print job is done.

Multiple LANs are usually interconnected by Ethernet bridges or switches which are considered as data link layer (or layer 2 as defined in Open System Interconnection (OSI) modal) devices. They have the following characters:

MAC learning ability : An Ethernet bridge or switch maintains a MAC learning table (the forwarding table) which associates ports with MAC addresses. This table is built in a way that the bridge or switch stores the unrecorded source MAC address in a frame and the corresponding incoming port. So, an Ethernet bridge or switch could identify where to send a frame based on the pre-learned MAC address table. Besides this table, a Virtual Local Area Network (VLAN) table which associates the port with VLAN is also kept. In VLAN, multiple LANs that are physically connected to a shared network are considered and operated as if they are in a single LAN. VLAN technology is a useful scheme to make Ethernet scalable [6] [7]. It uses a VLAN tag to partition a single large Ethernet into multiple VLANs [8][9]. The broadcasting messages for address resolution are limited to a single VLAN instead of the whole network, and hence the redundancy

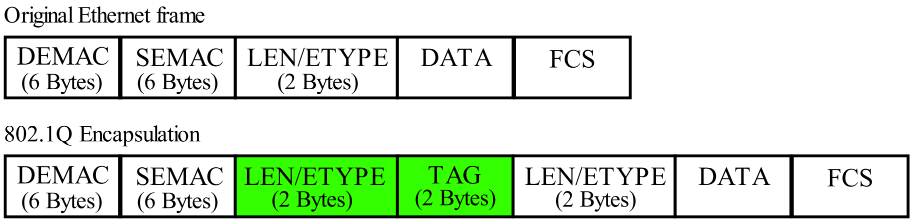


Figure 1.2: Add VLAN tag to a frame

traffic is reduced. The VLAN tag has only 12 bits and hence can only support up to 4096 active VLANs at any time in a network. To support more active VLANs, Q-in-Q or VLAN stacking encapsulation [10] [11] is proposed. In this scheme, a Provider Edge (PE) node inserts an additional Q-tag in the Ethernet frame to support more active VLANs.

Originally, a frame does not have a VLAN tag when it is generated. The switch receives the untagged frame and then assigns a VLAN tag to this frame based on the VLAN of the incoming port. Figure 1.2 shows the frame format before and after a VLAN tag is inserted. The 802.1Q tag is inserted in the middle of the untagged frame, which is between the source MAC address and the Type field. The Type field is set to 0x8100 which is defined for 802.1Q tagged frame. The tag control information field is divided into 2 parts. The left 3 bits are used to show the 802.1Priority which could identify eight different priority levels. Different level of services could be provided based on the priority level. The right five bits are used for the VLAN tag ID, which could offer 4096 $(2^{12})VLANnumbers.$

Flooding : Flooding happens when the destination MAC address in a received frame is not learned in the MAC learning table. Switch delivers the frame to all the ports that has the same VLAN of the received port expect the port where the frame came in. This is the convenient way to fetch the destination before this MAC address is recorded in the switch table. However, huge device and link resources are consumed by the flooding.

VLAN helps to constraint the size of the flooding.

Broadcast and Multicast supportive : If a frame needs to be broadcast to the network, the destination MAC address of the frame is set to all 1's that is FFFF FFFF FFFF. All the devices in the same domain receive this frame, and the interested receiver handles it or responds. Multicast is used to deliver frames to a smaller part of the whole network simultaneously. It could be considered as a subset of broadcast. Some users with the same interest could form a group, and a frame sent by a group member could multicast deliver to the other users in the same group.

1.2.2 Ethernet advantages

Generally speaking, Ethernet has the following advantages:

- high reliability
- low cost
- ease of management
- deployment convenient (almost 30,000,000 Ethernet subscribers are added every year)
- multiple choice of transmission speed (10 to 1000 Mbps to 10 Gig and 100 Gig)
- compatible with all upper-layer applications and operating system.

1.2.3 Ethernet services

Metro Ethernet Forum (MEF) [12] defined three categories of Ethernet services based on the flexibility of Ethernet. They are Ethernet Line (E-LINE), Ethernet Tree (E-TREE) and Ethernet Lan (E-LAN)[13][14].

E-LINE is the traditional point to point service which connecting two User Network Interfaces (UNIs), such as Ethernet Private Line (EPL) and Ethernet Virtual Private Line (EVPL).

E-TREE is a tree structure which supplies point to multipoint services similar to Ethernet Passive Optical Network (EPON) as described in [11].

E-LAN supplies multipoint to multipoint services between different sites. This service is convenient in a way that the established Ethernet Virtual Circuit (EVC) needs not to be reconfigured when a site is added or removed. In the provider network, the Spanning Tree Protocol (STP) could be used to prevent loops[15][16].

The corresponding models for the three services types are illustrated in Figure 1.3.

1.3 Metro Ethernet

1.3.1 Main MAN core technology

The MAN is generally a network that spans a metropolitan area covering multiple sites. Many technologies can be considered in MAN. Five main technologies will be investigated in the following. They are Synchronous Optical Networking (SONET)/ Synchronous Digital Hierarchy (SDH), Asynchronous Transfer Mode (ATM) [17], Resilient Packet Rings (RPRs), Multi-Protocol Label Switching (MPLS) [18] and pure Ethernet [19].

SONET[20][21]/SDH[22][23] which was originally the underlying transport network of the telephone system could also be used in all modern packet switching networks. The typical legacy SONET/SDH MAN architecture is composed of metro core and metro access rings interconnected by a combination of SONET/SDH Add/Drop Multiplexers (ADMs) and Digital Access Cross-connect Systems (DACS). SONET/SDH ring has the

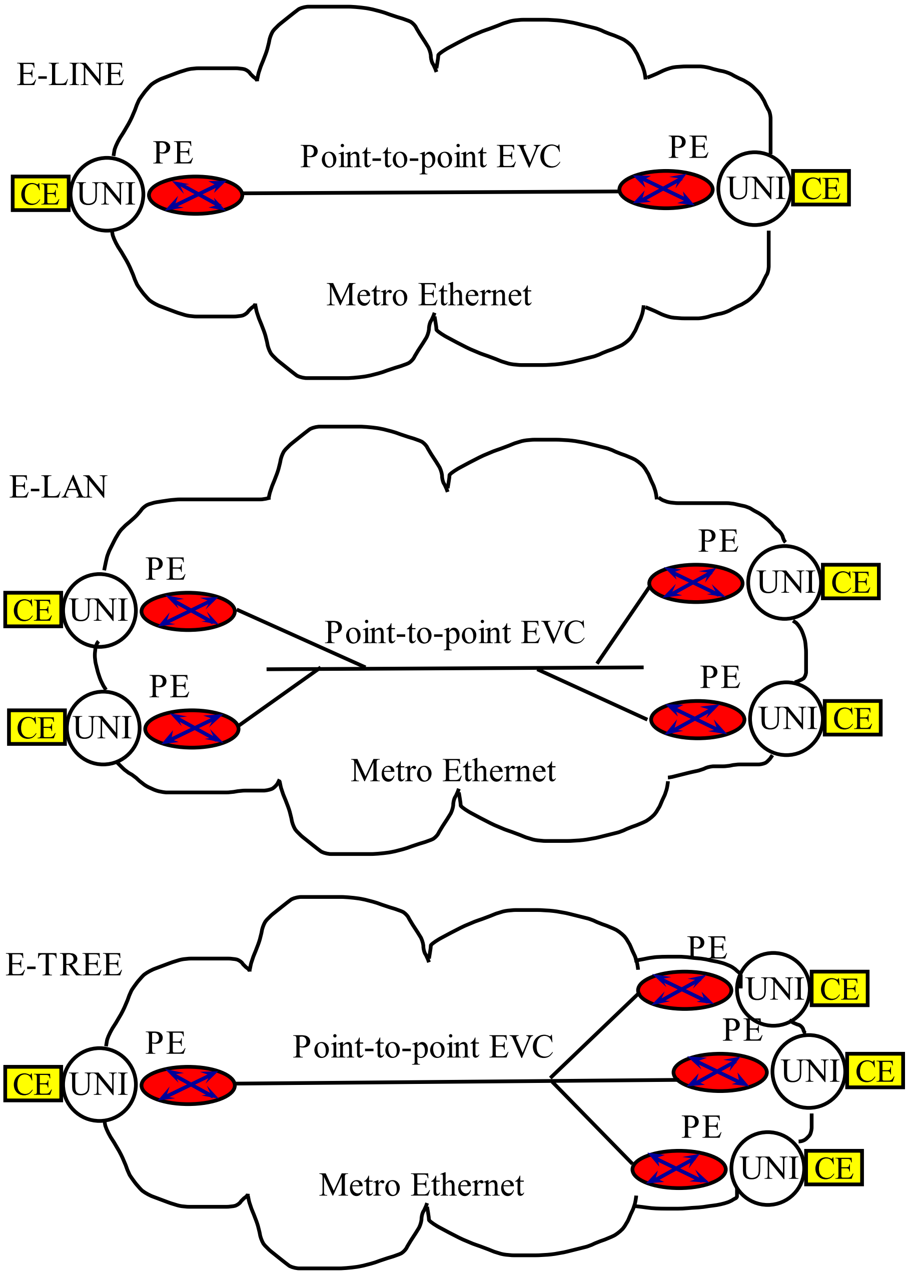


Figure 1.3: E-LINE, E-LAN and E-TREE services

self-healing character, that is, when a link failure happens, the SONET/SDH ring can be recovered automatically. Because of the voice-optimized nature, the data traffic in this network needs additional switches or routers to map data into Time Division Multiplexed

(TDM) channels for transiting the SONET/SDH network [24]. This affects the efficiency of the metro environments and restricts the hierarchical development. However, there are three main limitations of SONET/SDH MAN. They are: High maintenance expense, high devices expense and bandwidth inefficient.

ATM [25] is an end to end connection based technology, which is a powerful technology capable of providing not only regular voice but also IP based services. A Permanent Virtual Circuit (PVC) is usually constructed for every end user or every service. This circuit is terminated on the PE. In the Digital Subscriber Line (DSL)/ATM architecture, this PE node is called Broadband Remote Access Server (BRAS)[26]. Upon the DSL Forum [27] [28], the TR-59 DSL architecture model with ATM is a BRAS-centric infrastructure. Therefore, BRAS is the crucial point to deal with the traffic traveling both from the customer and the core network. The BRAS is responsible for multiple functions related to traffic, such as Authentication, Authorization, Accounting (AAA), Quality of Service (Qos), traffic aggregation and service differentiation. This leads to the drawback that all the IP traffic has to be directed to BRAS regardless of its physical location. Even when the communication is between the two end users resident in the same region, the traffic has to go through the BRAS which is a waste of bandwidth and causes an unnecessary communication delay. Another drawback is ATM can not handle multicast. As mentioned earlier, ATM is end-to-end technology, while multicast is an end-to-multiple transmission. Besides this, because of the expense of BRAS equipment the scalability of this infrastructure is limited.

RPRs was developed and promoted by the Institute of Electrical and Electronical Engineers (IEEE) [29] 802.17 RPRs Working Group and the RPRs Alliance. RPR metro networks have crucial benefits such as fast failure reaction, bandwidth fairness, multicast support, and physical layer diversity. However, it has the following limitations:

- As RPR can only operate over ring topologies, it can not be a full metro solution in itself.

- RPR is only a layer 2 technology. RPR is able to manage relation ships across nodes on a ring. Considering the complexity of the entire network which is definitely not a single ring topology [30], it is not suitable.
- RPR is too costly.

MPLS [31][32][33][34] has gained much attention for carrying Ethernet frames across metro networks. MPLS creates a virtual connection across a heterogeneous network of switches and routers. As illustrated in Figure 1.4, MPLS metro network is composed of access nodes (PE routers), Customer Edge (CE) and Label Switch Routers (LSRs). The Label Switched Paths (LSPs) are established between the two PE routers before the frame forwarding is started. As the layer 2 frames enter an MPLS service provider domain, MPLS uses a fix-format label layer 2 encapsulation to facilitates the transportation. According to destination MAC address/port/802.1Q information, two MPLS labels are inserted into the customer Ethernet frames by a Label Edge Router (LER) at the ingress of a network. The core LSRS only uses the tunnel label which is stacked at the top of the frame to carry the frame across the last outgoing LER. The LER uses the second inserted label, the Virtual Circuits (VC) label, to decide how to process the frame and transmit it on the appropriate egress interface. Because of the MPLS tunneling hierarchy, the VC label is hidden until the tunnel label is removed by the egress LER. MPLS offer benefits in both scalability and traffic control. However, it is far too complicated to be used for resiliency and is expensive.

Pure Ethernet MAN extends the native Ethernet protocol into the metro network and has attracted much attention recently[35]. Originally, Ethernet designed for LAN wasn't well fitted for MAN. However, due to the fast development of new techniques such as 1 and 10-gigabit Ethernet (1-GigE and 10-GigE) [36], Ethernet in MAN has become fea-

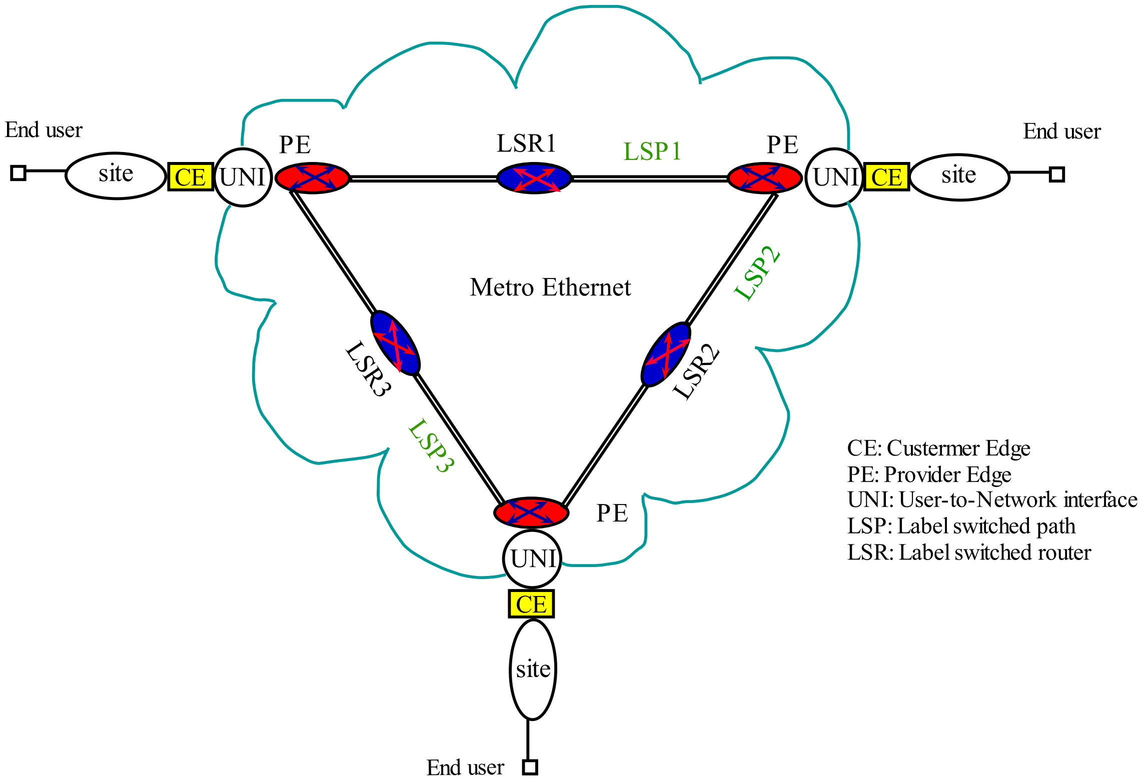


Figure 1.4: an MPLS metro network

sible. The speed of Gigabit Ethernet (100 BASE-X) is 10 times faster than Fast Ethernet (100 BASE-T) and 100 times faster than Ethernet (10 BASE-T). They are scalable and backward compatible. Figure 1.5 shows a general Metro Ethernet topology where STP is used to establish a tree like path to link all the customer sites in the same VLAN and Q-in-Q and MAC-in-MAC (MiM) encapsulation are employed[37][38]. Q-in-Q is used because the VLAN tag (Q-tag in IEEE 802.1Q) is limited to 12 bits. In order to support more VLANs, PE node inserts an additional Q-tag to the customer Ethernet frames before sending it to the metro domain. MiM encapsulation is used to solve the forwarding table explosion problem in the core network. PE node inserts its two additional MAC addresses, the source PE node MAC addresses and the destination PE node MAC addresses, to the customer Ethernet frames before sending it to the metro area. These two

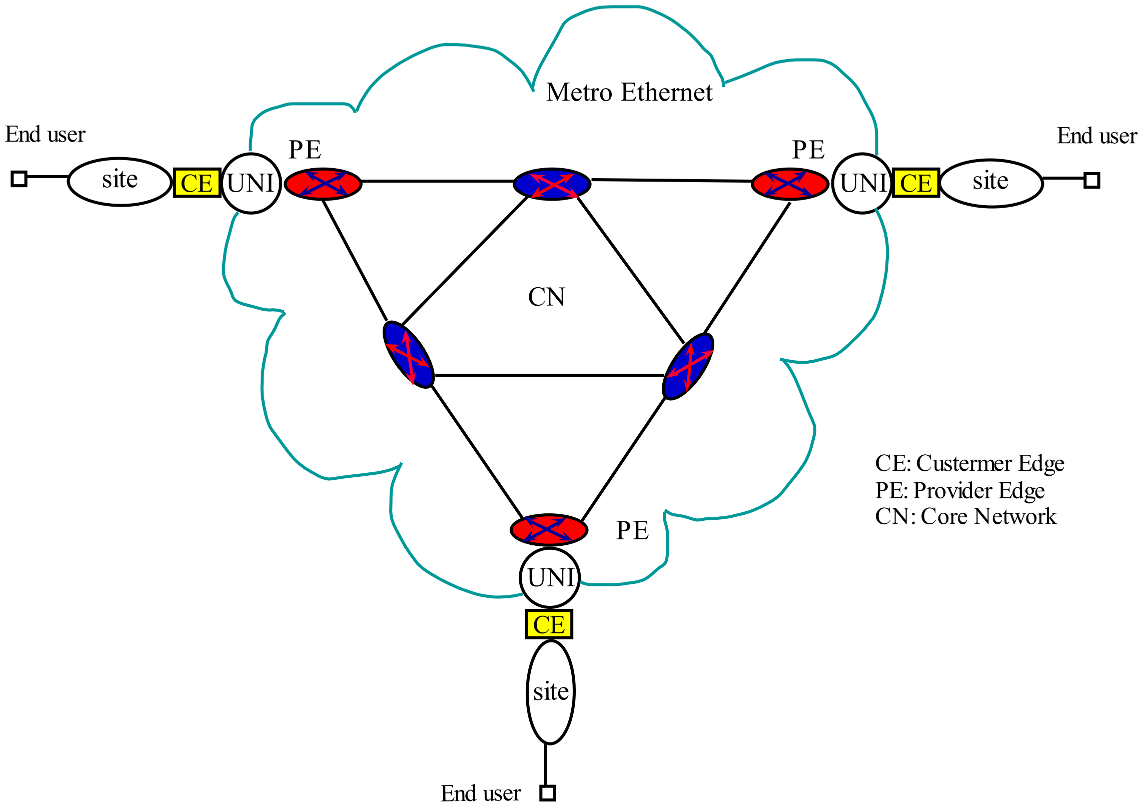


Figure 1.5: a pure Ethernet metro network

MAC addresses are removed by the destination PE node. Figure 1.6 describes the frame formats in Ethernet, MPLS metro network and pure Ethernet metro network respectively.

1.3.2 Motivation of Metro Ethernet

The traditional MANs are generally built based on layer 3 IP technology, usually using ATM as its layer 2 transport, whereas the LANs are dominated by layer 2 technologies, mostly based on Ethernet technologies [39]. As a result, when a data frame is sent across a MAN, it may have to travel through multiple heterogeneous networks, and hence it may need to be re-encapsulated in different frame formats a number of times and even be split into multiple smaller frames/cells or merged with other frames/cells into a larger

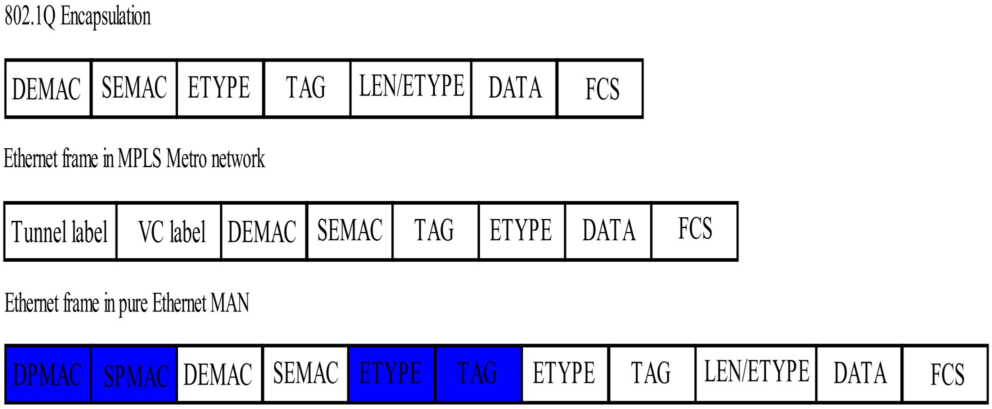


Figure 1.6: Frame formats in Ethernet, MPLS metro network and pure Ethernet metro network

frame when it passes through the MAN [40]. This not only makes the packet forwarding complex and inefficient, but also adds complexity and cost to the router/switch design, as well as network operation and management, in support of heterogeneous networks and interfaces. Clearly, deploying the same technology, such as Ethernet, in both MAN and LAN segments can potentially reduce the complexity and cost in network design and management, and improve packet forwarding performance[41][42]. To date, Ethernet has been considered as the technology to replace layer 3 technologies in MANs and is now encroaching into the Wide Area Networks (WANs).

Deploying Ethernet into the MAN core can establish a more homogeneous transport infrastructure, which will result in lower overheads, improved cost effectiveness, greater flexibility and aggregation. The following features pertaining to this technology are:

- **Reduced Overhead**: Without the frame size limitation and connection restriction, Ethernet proved to be IP adaptive in LAN environments, which reduces the huge overheads caused by re-encapsulating or splitting frames when they transit the MAN. Enabling Ethernet frames to transmit across a MAN, eliminates the abundant step to encapsulate Ethernet frames into ATM cells or other transport formats and then decapsulates them to the original frame format. This is impressive in saving traffic, as in ATM backbones the total overhead could utilize up to 25% of the traffic. For example, the throughput on a 155 Mbps circuit could drop to 116 Mbps [43]. This simplicity favorably results in a smaller delay time which guarantee less distortion;

- **Cost Effect**: Ethernet requires minimal management and maintenance cost and is ubiquitously deployed. Having been the dominant LAN technology for many years, mass Ethernet products are all over the technology market. Network administrators are already familiar and comfortable with Ethernet. The equipment of Ethernet is cheap and no pre-planning for optimal deployment is needed. The provided lowest per port cost becomes more and more significant as the port capacity increases;

- **Flexibility**: Dynamic configuration and plug-play nature result in better bandwidth achievement and higher aggregation efficiency. Having already spread worldwide, almost all the end systems support Ethernet interfaces;

- **Geographical support**: Today more and more companies run geographically distant campuses, offices, and data servers. To be able to create virtual LAN environments that link distant campuses, offices and servers together brings significant value to those companies for effective information exchanges and data sharing;

- **Bandwidth available**: The increased importance of the Internet for business applications, such as triple-play services, has led to a greater demand for higher MAN interface rates. Besides this, real-time applications such as voice and video are so sensitive to time delay that the shorter the access speed the greater the advantage gained by the subscribers. In addition, Ethernet provides the highest bandwidth-cost ratio. While 10 Gigabit Ethernet is widely available, 100 Gigabit Ethernet is under development and will be available in the near future[44][45]. So, it is possible to achieve a gigabit speed network by deploying all Ethernet networks.

1.4 Challenging issues

Ethernet was originally designed as a LAN technology that usually handles a small number of users. To be deployed in MANs, it has to solve the scalability issues [46] [9] [47] [48]. The main challenging issues when deploying Ethernet in MAN are:

Scalability: Because the universal MAC addressing scheme lacks the hierarchical structure, the forwarding of Ethernet is based on the fact that bridges learn MAC addresses automatically. Bridges inspect every incoming frame and learn the MAC and port mapping which can be used for the subsequent forwarding. This makes Ethernet simple. However, when it comes to the metro area where millions of end users reside, the MAC address table explosion problem comes[49]. Even worse, Ethernet uses broadcast to resolve unknown MAC addresses. This means, every node in the network has to process the broadcast frame, which also harms the node resource.

Resilience: Resilience is another factor that attributes to Metro Ethernet, which requires the ability to detect topology changes and network recovery automatically [50]. To guarantee the end-to-end services, the ideal network should perform smoothly and transparently when failures occur. After the fault detection, the network should perform network restoration automatically. That means an alternative path should be established through a network reconfiguration or a backup path quickly. This is related to the improvement of forwarding mechanisms of Ethernet, which is mainly based on STP. Though STP is simple and popular, the reconfiguration time which ranges from 1 to 2 seconds, is drastically small. Though some developments have been conducted such as Rapid STP (RSTP) and Multiple STP (MSTP), to fulfill the expectation of MAN, more work still has to go on.

1.5 Contribution of this thesis

In this thesis, some efficient and scalable information retrieval systems are proposed. The details are:

- Design an optimized cache replacement algorithm for video cache;
- Design novel Metro Ethernet technologies to satisfy the specific requirements of the MAN networks;
- Sustain the plug-and-play nature of Ethernet;
- Achieve high performance in terms of scalability and efficiency.

The proposed architectures include an Optimized Cache Replacement (OCR) scheme, an enabled Cache effect on Forwarding Table (CFT), an End user enabled MiM (EMiM) encapsulation scheme and two distributed Registration based address resolution Protocols (RP).

In OCR, the users are grouped into different cache groups based on their arrival patterns. It calculates the user density among all possible intervals, and then selects the maximized number of users which can be answered from the cache. Based on the optimized scheme in a single cache, we also extend OCR to cooperative caches. Simulations are conducted to verify the proposed schemes. The results show that the OCR can increase the hit ratio and reduce the server load.

In CFT, we modified the forwarding table entries in CE and PE nodes to learn both the IP and MAC addresses. By receiving a frame, the CE and PE nodes cache the IP-MAC address mapping carried in the frame. Once the mapping is remembered, the information can be served for the subsequent requests asking for the mapping. For protocols relying on broadcast to look for the MAC address of another end user, the designated CE and PE nodes can stop the broadcast and reply back directly with the help of the information stored in the new forwarding table. So, the CE and PE nodes not only can determine the next hop but also handle the Address Resolution Protocol (ARP) request coming from the segments behind it. CFT is easy to be accomplished and fully backward compatible.

In EMiM, the PE node's MAC address is associated with an end user's ARP entry. This modification allows an end user to encapsulate both the destination user's MAC address and its PE node's MAC address in the frame. All the PE nodes only need to swap addresses and hence do not need to maintain the entries mapping end user's MAC address to its PE node MAC address, thus significantly reducing their forwarding table sizes. We also discuss how the proposed schemes coexist with end users using traditional ARP technology.

In RP, multiple ARP registers are allocated to support address resolution. Each IP address has a home register which stores its ARP entry. When an end user moves to another location but keeps its IP address, its current PE or CE node is considered to be its foreign register. A foreign register temporally caches the ARP entry for an immigrated user and is in charge of the ARP entry updating in the home register. The IP address is used as an index to locate the corresponding home register through unicast, thus eliminating the broadcast to solve an unknown address. The proposed architectures are evaluated by simulations and the results show the proposed schemes can save more than 60% of messages for address resolution and reduce by up to 80% forwarding table size in PE nodes.

1.6 Outline of this thesis

The rest of the thesis is organized as follows: Chapter II gives the related work; Chapter III demonstrates the detail of the OCR scheme; The description of CFT scheme is shown in chapter IV; Chapter V presents the details of the proposed EMiM architecture. Chapter VI describes the RP schemes. The research work that will be conducted in the future is proposed in Chapter VII. Finally, Chapter VIII concludes the thesis.

Chapter 2

Literature Review

In this chapter, we provide an overview of the standards and schemes for information retrieval systems. The organization of this chapter is as follows. Section 2.1 summarizes the related work of caching; Section 2.2 reviews the VLAN technology; Section 2.3 depicts the MiM encapsulation; Section 2.4 discusses the novel Ethernet forwarding approaches.

2.1 Related Work with Caching

There are various caching algorithms designed for Web cache, such as LRU, Least Frequent Used (LFU)[123], Log(Size) [51], Day [52], Page Load Delay [53], Greedy Dual-Size [54], [55] and Logistic Regression [56]. Among them, LRU is the simplest cache algorithm and is widely used in the real world web cache. Many researchers developed new caching schemes based on LRU. A counter-based L2 cache replacement is introduced in [57]. Based on the certain principles, the event counter for each entry is changed. When this counter passed the defined threshold, this entry is discarded. As the data within the higher priority part of the LRU stack is more likely to be used, line distillation was proposed in [58]. In this scheme, the more frequently accessed data are stored in a cache line and the rare used one is eliminated. SF-LRU [59] combines LRU and LFU schemes so it could evict items that are not only used long time ago but also have low frequency to be hit again. Filtering cache (LBF cache) proposed in [60] is based on LRU-like algorithm which caches more recently used blocks. Different from the previous cache, its cache contains two parts, a Direct-Mapped Cache (DM cache) and a fully associative buffer. In the DM caches, one bit is used for each entry to indicate the most recently used block. The replacement of items is based on this bit. It can increase the hit ratio.

Video streaming caching is different from the Web caching[125]. The future request of the first one is more predictable and the data items which need to be cached are huge. Based on this character, many specific caching policies are proposed to minimize the accesses latency and balance the network traffic load with limited caching storage[61][62][63][64] [65][66][67]. [64] employed to use variable sized chunks to overcome the problem of low byte hit ratio during unsteady period. MiddleMan architecture is composed of a serial of cooperative proxy servers which are managed by a coordinator. The coordinator plays a key role in data caching, such as redirects requests to the corresponding proxy, decides the content of the caches and eliminates files when caches are full. Recently, researchers on hierarchical peer-to-peer architecture for VoD streaming are extremely popular. [68] utilizes a utility function to predict the future demand of each chunk. A control server is the crucial part on the operating video caches, the job keeps track of chunks hosted by each peers, directs requests to the corresponding peer and decides where this chunk should be accommodated. Another cooperative hybrid P2P architecture is described in [52]. In this scheme, a priority index which is calculated based on the sum cost of fetching this chunk from other peers and its hit frequency. Similarly, [53] also considered a P2P architecture with peers with limited cache size. Normally, Chunks are requested in subsequent order, but randomly skip could be supported as well. The chunks with higher supply deficit have more opportunities to be downloaded.

Several papers proposed segmentation caching. Many of them divided video files into predetermined fixed length segments including prefix and suffix [69][70][71][72]. The segments are replaced based on their popularity and a larger size segment is given to the video with higher popularity. The suffix which does not saved in cache could be fetched by the prefix. DECA[73] employed variable chunk sizes which could be changed dynamically according to the updated segment popularity. The prefix segment based partial caching algorithm improved the caching efficiency. However, the fetching between prefix and suffix may lead to initial time waiting.

FGS video-based caching system [124] adopted fine-grained scalable coding which has been used for MPEG-4 standard in proxy caching. An efficient caching management framework is developed considering both high cache chunk utilization and flexible bandwidth adaptation.

2.2 VLAN

Ethernet uses a broadcast-based address resolution scheme which introduces a large number of broadcast messages into MAN. Many protocols such as ARP[74] and DHCP[75] use the broadcast service as a service discovery mechanism. For example, to solve the MAC address for an IP address, an ARP request message is broadcast through the network so that the corresponding end user can receive and respond to an ARP replay message. The broadcast based address resolution schemes make Ethernet extremely convenient and easy to employ. However, for MANs with millions of end users [45], high frequency broadcast messages waste a lot of bandwidth for address resolution. Moreover, every end user needs to use resources to handle every broadcast message [76].

One way to reduce the broadcast message is to use the VLAN technology [77] to limit the broadcast messages. A large network is portioned into smaller separate segments by

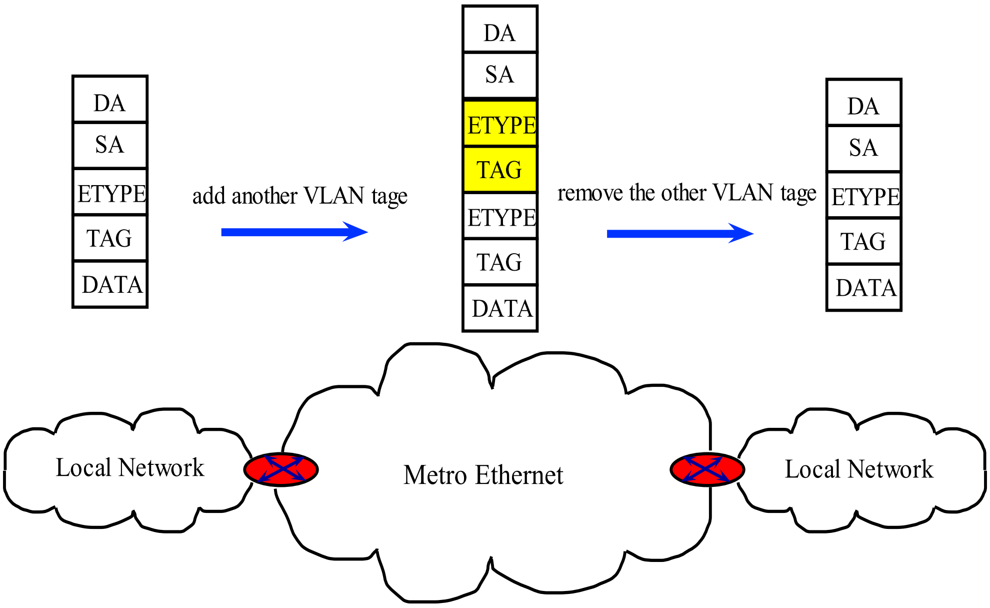


Figure 2.1: Q-in-Q encapsulation approach

assigning different VLAN tags to end users. The end users in one VLAN are logically located in one segment, while they may be located in different segments physically. The broadcasts initiated by end users are restricted to transmission only inside the VLAN segments the user belongs to and end users in other VLAN are not involved. This helps to alleviate the broadcast problems. Another advantage of VLAN is the flexibility of roaming. As the VLAN configuration is logical, the physical movement of an end user could be ignored. If a user physically moved from segment S to another segment T, as long as this user is logically configured to belong to S, it could be considered as still connected to S regardless of the new location.

One limitation of VLAN is that the communication between any pair of users should be in the same VLAN, and hence a large number of VLANs need to be created. However, as VLAN tags only have 12 bits, the number of VLANs that can be supported is limited. In order to serve more VLANs, Q-in-Q is defined to enhance intelligence and scalability. Another VLAN tag is added to the frame before it is sent out to the metro area by the provider edge. The Q-in-Q approach is shown in Figure 2.1. It does not affect the MAC address learning method. The provider network devices have to learn all the end users' MAC addresses. Besides the VLAN shortage problem, VLAN also has other problems:

Configuration overhead : VLANs, subnet configuration and address assignment are needed to be done manually. Careful planning and operations are needed to decide on VLAN tags for every port of every device.

Limited scalability : Although broadcasting is limited in every VLAN area, in metro area, the number of users located in one VLAN is still large. Moreover, VLAN could be overlapping. Users could belong to multiple VLANs. The transmission devices should be visible to every user who considerers them as the metro ingress device in every VLAN. As broadcasts in multiple VLANs traverse by these devices, the forwarding table is replaced frequently by the end users belonging to these VLANs. The devices often provide more VLANs than require and flexible user roaming, which aggravates this problem.

Sufficient efficiency : A single spanning tree is used to forward frames inside VLANs, which is not efficient in a large network.

2.3 MiM encapsulation

Ethernet has poor scalability due to the use of a flat addressing scheme (i.e., non-hierarchical MAC addresses). Upon receiving a frame, Ethernet switches or bridges learn where to direct a frame by associating the source address of an incoming frame with the incoming port. This information is recorded in the forwarding table and prepared for delivering the subsequent frames to the corresponding destination. The entry format is <*MAC*, *port*, *recordtime*, *age*>. When a frame with a new source MAC address $e1$ arrives at port $p1$, by inspecting the frame, the switches or bridges store the information in the form of <$e1$, $p1$, *recordtime*, *age*>. If a frame destined to $e1$ is received by this device, the frame will be sent out on port $p1$. The MAC forwarding table at a node in a provider network needs to keep potentially a large number of MAC-to-port-mapping entries for frame forwarding. In a MAN environment composed of a large number of LAN segments, this may either cause forwarding table explosion or necessitate excessive frame flooding, depending on the actual timeout values for the table entries. Entries in the forwarding table can either be replaced by a new coming entry when the table is full or be deleted after the timer expired. Flooding is required when a destination could not be found in the forwarding table. Actions such as end host power up/down and roaming from one place to another make the destination absence problem even worse. In order to guarantee that the destination end user receives the requirement frame, the flooding frames traverse all the ports of all the devices in the network except the port at which the original frame arrives. Increasing number of recent entries that could be used by the following frames are replaced by the flooding addresses, which lead to more frequent broadcasts. In the MAN area, where the number of entries in the forwarding table are much smaller than the number of MAC addresses, flat addressing not only intensifies the table size problems but also wastes link bandwidth and processing resources because the control overhead disseminates end users frames by flooding.

To solve the problem of forwarding table overflow in Metro Ethernet, the MiM encapsulation scheme [9][78][79] has been developed. Originally switches or bridges had to learn the MAC addresses of all the end users, which could be hundreds of thousands in a metro area. Using this method, core network switches only had to learn the MAC addresses of provider edges. As shown in Figure 2.2, when a frame is transmitted from the local network to the metro core network, an additional pair of MAC addresses is inserted by the ingress PE node. Before sending the frame to the core network, PE node encapsulates the frame with the MAC address of the PE node where the destination user resides and this PE's MAC address. The frame format of the MiM encapsulation is shown

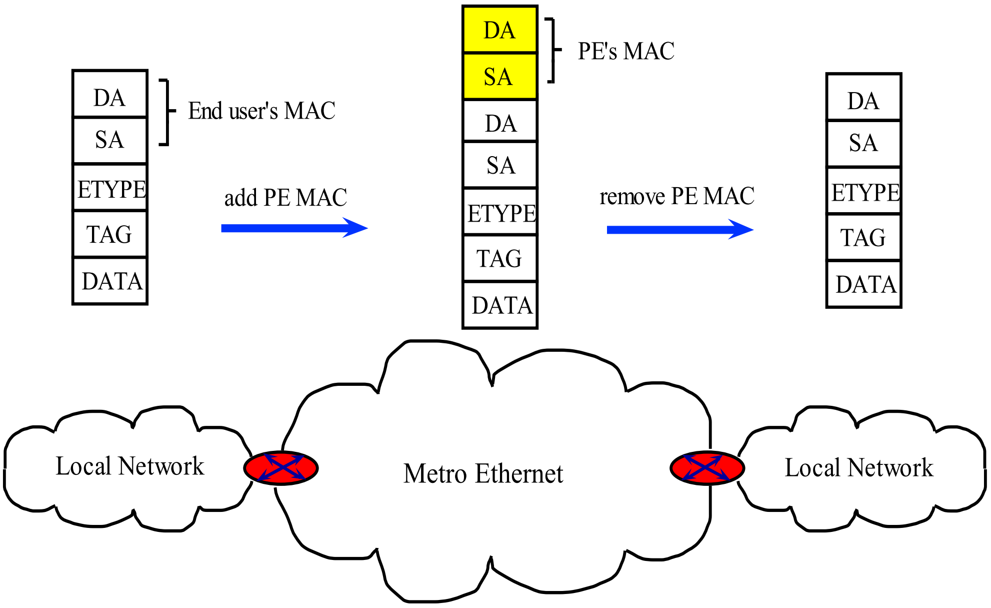


Figure 2.2: MiM encapsulation approach

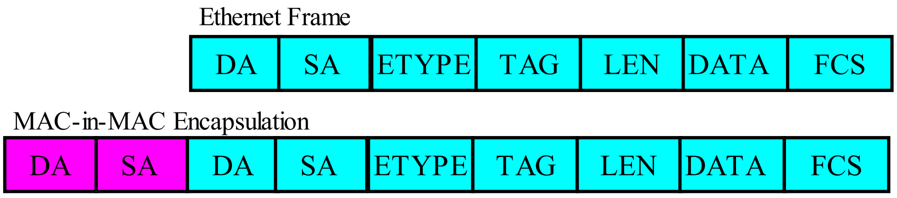


Figure 2.3: The MiM encapsulation

in Figure 2.3. The MAC addresses of the end users are hidden from view by the core network devices. So, only the MAC addresses of the PE node are learned. The added MAC addresses are stripped from a frame before it reaches the destination local network by the egress PE node. Leaving the forwarding in local network unaffected. MiM encapsulation extends Ethernet into Metro area without much complexity compared with other technologies. However, this scheme still cannot avoid learning table explosion.

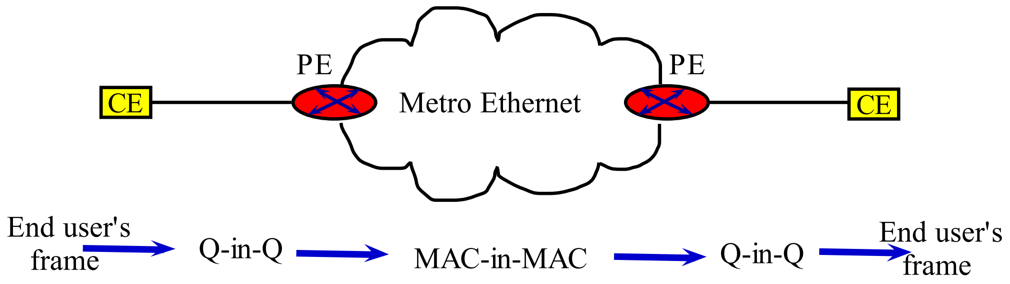


Figure 2.4: Hybrid approach of MiM and Q-in-Q

MiM is interpretable with VLAN and Q-in-Q encapsulation (Figure 2.4). According to Nortel Network [80], Q-in-Q is encapsulate in CE node and MiM is operated in PE node.

However, MiM can only reduce the forwarding table size in the Core Nodes (CNs), but not in the PE nodes which still have to maintain the entries of mapping user's MAC address to its PE node's MAC address for MiM encapsulation[81]. Huge end user's MAC addresses still have to be recorded at the provider edge nodes.To deploy Ethernet technology to MAN, it has to solve the issue.

2.4 The Novel Ethernet approaches enhancing the scalability

Traditionally, Ethernet relies on STP standardized in 1998 as IEEE 802.1 D to switch frames in a network [82]. STP is primarily used to provide loop-free communication among all the nodes. So at any time, only one Spanning Tree exists in the network. RSTP improves the reconvergence time of the STP, which reduces from minutes to less than one second. The main problem of RSTP is Count-to-infinity [83]. MSTP [84] [85]defined in IEEE 802.1s is proposed based on RSTP and VLAN. MSTP improve the performance of the network in several ways: A failure in one region not affects the traffc flow in other region and load balancing can be managed manually by assign different VLANs. However, MSTP inherits the other drawbacks of RSTP which is the only underlying protocol of MSTP. Besides, MSTP configuration is complex as VLANs must be assigned accurately for all bridges. Spanning tree protocols, as native routing protocol for Ethernet, draw mounting attentions when Ethernet is being considered in metro areas[86][87][88][89][90][91] [92][93][94][95][96].

In order to avoid the drawbacks of STP approaches, the shortest-path forwarding schemes are attracted much attention within the IEEE 802.1 working group [97] [98]. In [99] and [100], the link state routing protocols were proposed to replace the spanning tree based routing schemes [101][102][103]. Similar to routing, the link state information is exchanged between bridges to create the tree instances. RBridges [99] [104] could avoid broadcast when distributing of locally learnt MAC addresses by using a link state approach. The IP datagrams are normally performed by the routing protocol (e.g., IS-IS [105] [28]) learnt information. Without the restriction of STP, all the links are available to use, which results in better bandwidth efficiency and faster convergence. In [106], a hybrid scheme which uses the spanning tree protocol in the core network and the link state protocol in Metro access network respectively was proposed. The hybrid scheme can achieve better performance than using either spanning tree or link state scheme in the MANs. The main shortcomings for these schemes are the inconsistent intervals during the convergence time which is not optimized. In addition, the link state protocols provide no announcement after the reconfiguration which is necessary. All these schemes still have to use the broadcast based ARP. And the complexity and high cost should be considered.

Recently, a suggested solution to enhance the scalability for the traditional Ethernet is to use directory based ARP [47] to eliminate broadcast service for address resolution. In the scheme, a user needs to register its ARP entry in the ARP directory before its communication, and an unknown MAC address is resolved through the ARP directory lookup. In this scheme, the control plane is divided into two planes, a decision plane and a dissemination plane. The decision plane is in charge of calculating forwarding tables for each switch in the network. The dissemination plane is in charge of delivering network status information to the decision plane and distributing configuration information to the switches. Network topology, link status, and host status are gathered by the dissemination plane. The information is then used by the decision plane to calculate forwarding tables and answer the ARP requests. This scheme requires all the end nodes to register to the bridge attached to them when they join the network for the first time, and periodically re-register during the connection. So the bridges could perform link state routing as they have the MAC addresses of the end nodes. The information is exchanged with the neighbor bridges when a new end node is registered. Unicast forwarding is possible by the end user MAC addresses and attached bridges mapping information and the link state topology information. But for Metro Ethernet, a single ARP directory based solution is not scalable and efficient. Moreover, it creates single point of failure.

Some hash based address resolution schemes [107] [108] [109] [110] are proposed to eliminate the reliance broadcasting frame learning. Instead, when a destination MAC address is missed in the forwarding table, the frame is routed to a designated user based on the MAC hash value. In [107], two types of bridges are involved: advanced bridge and legacy bridge. A legacy bridge is the conventional bridge, while an advanced bridge is the new bridge that performs the proposed distributed address resolution. Only one advanced bridge is permitted to transmit packets to or from other segments, which is the designated advance bridge of this segment. Frame forwarding in the same segment follows the traditional Ethernet forwarding process. Frame forwarding between different segments will be delivered to the designated advance bridge of this segment, which determines the routing to the destination without resolve broadcast. SEATTLE [110][111] provides an alternative network architecture which not only sustained the plug-and-play feature but also provide shortest-path forwarding and end user address resolution via hash. A one-hop network-layer distributed hash table is used to store the mapping of the MAC address and the host location. After the location of the end user is found, the frame could be delivered along the shortest path based on the link-state information obtained by the switches. The cache update is based on unicast, which is efficient and prompt compared with the traditional broadcast based one. When network layer changes occur, the link state advertisement is exchanged between switches directly. This notified the switches to evict those invalid entries, which ensures all the frames are forwarded based on up-to-data state. Besides, switches in SEATTLE could cache responses to queries, such as ARP replies. These traffic-driven location resolutions and caching avoid excessive load when building the DHT. But routing frame with unknown destination addresses to designated user may take the frame travel more unnecessary hops to the destination, and make the traffic control more difficult.

SmartBridge [112] allows finding the shortest forwarding path by exchanging topology information among bridges. The topology knowledge is obtained by diffusing computation [113]. The computation process is described as follows: A bridge starts a topology acquisition process by asking its neighbors. The neighbors then direct the request to their neighbors. And the process is going on until all the bridges in the network got this request. A message is replied back after each request. So, the initiator could know when the whole distributed computation is over. When a topology change occurs, the bridge that notices the change triggers a diffusing computation that spreads to every bridge. The new topology information containing the links between different bridges and segments is exchanged among bridges. Packet dropping may happen during the topology request process. The traffic forwarding is performed based on the host MAC address and segment mapping table in Smartbridge. The forwarding between different segments is performed on shortest-paths. Smartbridge achieves shortest path forwarding between different segments and low re-convergence time. The shortages of this scheme are that a full knowledge of the network topology should be obtained and the storage consumption is extremely high for each Smartbirdge.

In [114], a MAC Address Translation (MAT) scheme is proposed for frame forwarding. The flat MAC address is translated to a hierarchical structured address for frame routing so that the number of forwarding table entries is reduced. MAT utilizes Locally Administered Address (LAA) which is traditionally used for network management to perform MAC address translation at PE routers. The PE routers replace the Organizational Unique Identifier (OUI) prefixes in legacy MAC address by the PE prefix before delivering the frame to the core network. So frames are forwarded by the PE prefix between different PEs. This resolves the MAC address table explosion problem as PEs are released from recording every MAC address in customer's network. However, it is possible that two MAC addresses are translated to the same structured address.

Both [115] and [116] provided the concept of using cache to suppress broadcast traffic. By caching the most recently used dynamic directory entries at every PE node and using some specified PE node maintaining the ARP entry, broadcasting messages can be reduced [115]. The Ethernet frames sending by the end user are encapsulated with an outer header when passing through PEs. The service discovery is processed by using a distributed directory which is only implemented on the PE switch. Two types of entries are included in the directory, static and dynamic entries. Static entries are servers such as DHCP, RADIUS, DNS etc. All the PEs maintain these static entries for fast discovering of these machines. Dynamic entries are used to record the information of the end user. At most two PEs keep the dynamic entry for one end user. One is the PE switch this end user directly connected to, and the other is the specified PE node which is selected based on the hashing function over the combination of service type and IP address of the end user. The most recently used dynamic directory entry could be cached at PEs until they time out. This could reduce the latency to fetch the entry from the specified PE node every time. But completely replace Ethernet devices to new ones is not feasible for real word application. Similarly, Etherproxy in [116] caches the ARP entries it learned and suppresses broadcast messages it received by looking up the entries it cached. This new kind of device could be set into an existing Ethernet directly. When broadcast arrived, EtherProxy's check the entries it cached to fetch the answer before broadcasts it to the rest of the network. If the corresponding entry is found, EtherProxy could reply to the requester and suppress the broadcast. Otherwise, the request is broadcast over the network. When the reply transit EtherProxy, it caches the response information which could be served for the subsequent request. It retains the plug and play nature of Ethernet and is backward compatible. However, broadcast still happenes under the Etherproxy, which scales with the size of the network. And these cache schemes need extra memory to store the cache entries.

Some data center network architectures, such as VL2 [117], PortLand [118] and MOOSE [119], achieve scalability by assigning user's flat names based on their hierarchical location. In these mechanisms, centralized directory systems are used to maintain the information of the end users. This is suitable for cloud service data centers where end user is virtual and totally controlled by provider, while a decentralized approach is much better when deploying in large autonomous and heterogeneous networks.

[120] designed a Metro Ethernet network architecture, which is combined of many optimization algorithms. The methodology is implemented in a prototype design tool and is used by ATT network planners. As the case studies in metro areas show, this methodology fulfills planner's expectation.

2.5 Summary

In this chapter, the relevant works with caching are summarized firstly. Then, the existing technologies in Ethernet is reviewed . The VLAN technology is included and shows that it is not suitable for the MAN area. After that the MiM encapsulation scheme is demonstrated. Then some approaches based on registration or cache to improve the scalability of the Metro Ethernet are described. However, none of these schemes could totally make Ethernet suitable when deploying in the MAN.

Chapter 3

An Optimized Cache Replacement algorithm for video on demand

This chapter presents an Optimized Cache Replacement (OCR) scheme and a Cooperative Cache Replacement (CCR) scheme. Section 3.1 gives the overview of this chapter. The details of OCR and CCR are shown in Section 3.2 and 3.3. The simulation results are given in Section 3.4. Some conclusions are drawn in Section 3.5.

3.1 Overview

With the explosive growth of the video streaming applications such as Youtube and Facebook [121][126], how to provide high quality video streaming to users becomes a key issue for successful service providers. Caching and multicast are two common schemes to improve the system performance. In the real world, the users are usually heterogenous and asynchronous, using multicast to group users can reduce the server load, but it increases the initial latency for most users.

Caching the popular video chunks for users' reuse/replay is another effective way to reduce the server load. However, the video streaming usually has huge space, it is hard to cache the whole video. Hence how to manage the cache for video streaming data plays a critical role for cache performance. Unlike the traditional Web Cache which caches individual Web documents, the request of the cached chunks for users has certain patterns, and hence the traditional cache replace schemes [122][52][51][53][56] may not be useful for video caching. For example, three users, say $u1$, $u2$ and $u3$, request the same video at time $t1$, $t2$ and $t3$. To make the cache useful, the video chunks requested by $u1$ should be cached till $u3$ requests the chunks, and then they can be removed from the cache. These three users form a user group which can use the cached chunks. Hence, the traditional cache replacement scheme such as LRU, LFU, and utility based cache replacement schemes are unsuitable for video streaming cache.

This chapter aims to design an OCR scheme for video cache. OCR groups the users into different cache groups based on their arrival pattern. It calculates the user density among all possible intervals, and then selects the groups with the maximized number of users to be answered from the cache. Based on the optimized scheme in a single cache, we also extend OCR to cooperative caches (called CCR). Simulations are conducted to verify the proposed schemes. The results show that OCR and CCR can greatly increase the original hit ratio and reduce the server load.

3.2 An Optimized Cache scheme for video streaming

Figure 3.1 shows a general video cache architecture. Due to the limited cache space, how to effectively use cache space to maximize the system performance becomes important. In this section, we propose the OCR scheme for video streaming.

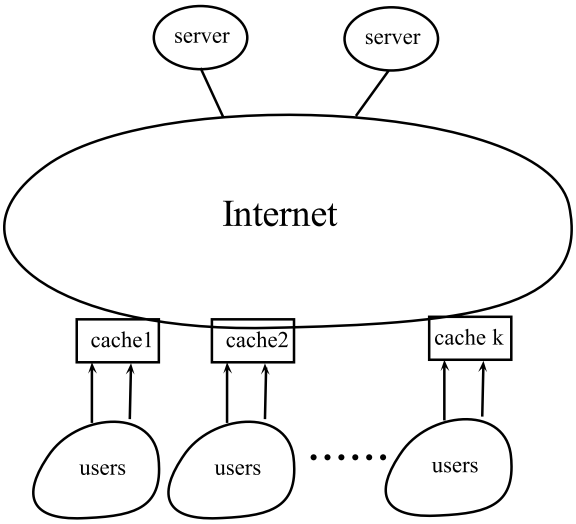


Figure 3.1: A video cache structure

3.2.1 An Optimized Cache Replacement scheme for video on demand

For video streaming data, if two users request the same video, they should ask for the same sequence of chunks in the following time. If we know the users' request patterns, we can organize the users in groups such that the cache can serve the maximum number of users in an optimization way. The OCR scheme optimally groups users to maximize the cache hit ratio. When a new user comes or an existing user leaves/finishes, the groups may be reorganized to maintain the maximum hit ratio. The details of the scheme is given below.

Consider a T_v seconds video, the data rate is R, then the total bytes for the video is RT_v. The video is split into chunks, and each chunk takes t_c seconds (i.e., Rt_c bytes). The number of chunks is T_v/t_c. If the cache size S_c is greater than RT_v, the whole video can be cached, and hence no replacement is needed. We consider the case that the cache space is less than the whole video data. Assume there are N_u users access the video during the T_v time (see Figure 3.2). User i (denoted as U_i) accesses the video at time t_i^u. Then the optimal cache replacement scheme is to achieve the maximum hit ratio to reduce the server load.

Definition 1: Assume users U_i and U_j request the video at time t_i^u and t_j^u, respectively, and the number of users (including both U_i and U_j) requested the video between t_i^u and t_j^u is n_{ij}, the *user density* D_{ij} is defined as

$$D_{ij} = \frac{n_{ij}}{t_j^u - t_i^u} \tag{3.2.1}$$

Definition 2: If all the chunks requested by U_i are cached and kept in the cache till u_j request it. Then all the users requested the video between t_i^u and t_j^u form a *caching group* G_{ij}.

Let us look an example shown in Figure 3.2(b), U_1, U_2 and U_3 are in a caching group; and U_4-U_7 form another caching group.

Definition 3: For a caching group G_{ij}, the *hit density* is defined as

$$H_{ij} = \frac{n_{ij} - 1}{t_j^u - t_i^u} \tag{3.2.2}$$

Definition 4: *Group Interval* of G_{ij} :IN_{ij} of caching group is defined as $t_j^u - t_i^u$.

OCR aims to maximize the cache hit density, i.e., to maximize the cache hit ratio. Suppose N_u users request the video, OCR first calculates the hit density from all possible caching groups if its group interval is less than the cache interval. OCR always picks a group with highest hit density, and then picks a caching group with the second highest hit density and so on. When a new user joins, the hit density of the groups starting from U_k is recalculated. If a new group has higher hit density, it should be cached. If no space, a group with smallest hit density should be removed from the cache. All caching groups may be merged with others and/or be split if some new group appears. If some new group

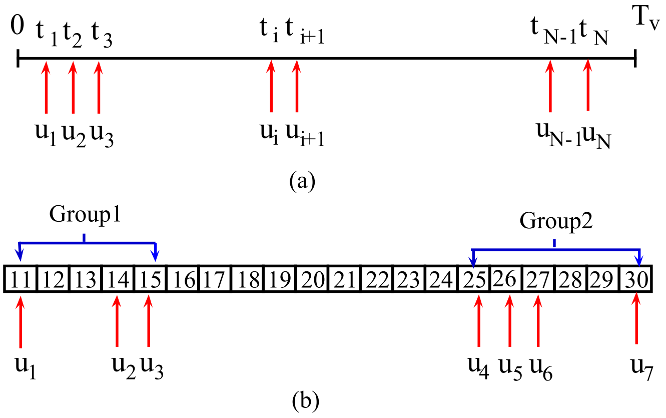


Figure 3.2: Group illustration in a cache

```
K: the number of current users
user  $U_i$  requests a chunk  $C_i$
IF ( $U_i$ is a cache user )
    return  $C_i$  to $U_i$ using cached item
ELSE IF  ( $U_i$ is a misuser )
    send the request to the server
ELSE
    For  ( j=1 to k )
        IF ( $t_i^u - t_j^u <= t_c$ )
            caculate $H_{i j}$
            IF ($H_{i 1}$ is selected as a cache group)
            mark $U_i$  as a cacheuser
            mark all users joined in the cache group as cacheuser
            mark all users left the cache groups as misuser
            ELSE
            mark $U_i$  as a misuser
```

Figure 3.3: OCR cache replacement algorithm

has higher hit density, it should replace the caching group with low hit density. We called this process as *regroup*.

During the regroup procedure, all the users (i.e. grouped user, etc.) in the process are considered.

If a user first requests a chunk, the cache checks if the chunk has been cached. If so, this chunk is returned, and the user joins to the cache group (marked as cacheuser). Otherwise, a cache miss is counted and the type of this user is examined (check for regroup), if the user can join in a group, the user is set to be a cachuser. Otherwise, it is marked as misuser. For a user tagged as a misuser, the following chunks are retrieved directly from the server, and the cache does not need to cache these chunks. For a cacheuser, the following chunks are given by the cache (except the first user in the group, it retrieves the chunk from the server, and the chunk is cached). Figure 3.3 shows the pseudo code of the OCR scheme.

3.2.2 A Cooperative optimized Cache Replacement scheme for video stream

The OCR scheme is designed for an individual cache. However, for some very popular videos, they should be cached in many caches, and the cost by retrieving a chunk from a neighbor cache usually is much lower than that of the server. In this section, we derive a CCR scheme.

One way for cooperation is just search the chunks from the neighbor caches. The replacement/regroup scheme of individual cache is the same as those in OCR. The other way is to consider the cost difference of the request from the neighbor caches. In CCR, we assume that the neighbor caches can exchange the cache information, and the cost ratio of retrieving a chunk from a neighbor cache to the server is r_u. Then the hit density of a group is defined as

$$H_{ij}^{c} = \frac{(n_{ij} - 1) + (1 - r_u)n_{ij}^{c}}{t_j^u - t_i^u} \tag{3.2.3}$$

Here n_{ij}^{c} is the number of users come from the neighbor caches. When a new user coming from a neighbor, the hit density can be recalculated and the group may be reorganized if necessary. The pseudo code of CCR is shown in Figure 3.4.

```
K: the number of current users
user $U_i$ requests a chunk $C_i$
IF ( $U_i$ is under the cache )
    execute OCR
    IF ( $U_i$ is a misuser )
    send the request to a neighbor cache
ELSE
    IF ( $U_i$ is a cacheuser )
        return $C_i$ back to $U_i$
    ELSE IF ( $U_i$ is a misuser )
        send the request to the server
    ELSE
        For ( j=1 to k )
            IF ( $t_i^u - t_j^u <= t_c$ )
                caculate $H_{ij}^c$
                IF ($H_{i1}^c$ is selected as a cache group)
                    mark $U_i$ as a cacheuser
                    mark all users joined in the cache group as cacheuser
                    mark all users left the cache groups as misuser
                ELSE
                    mark $U_i$ as a misuser
```

Figure 3.4: CCR cache replacement algorithm

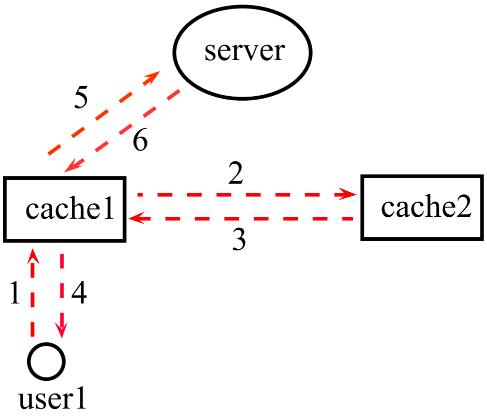


Figure 3.5: Example illustrating the CCR caching algorithm

Figure 3.5 illustrates an example of CCR. Suppose that user 1 located under cache 1 requests a chunk. Firstly, the request is sent to cache 1 (step 1). As the chunk is not available in cache 1, the request is directed to cache 2 which is the nearest neighbor of cache 1(step 2). After receiving the request, cache 2 searches the data recorded. Either the fetched chunk or an unfounded message is replied back to cache 1 (step 3). If the chunk is returned, cache 1 redirect this chunk to user 1 without caching it, and this request is finished (step 4). The following requests of the user is direct sent to cache 2. Otherwise, cache 1 redirects this request to the server (step 5), and the server delivers the chunk back (step 6). Before sending the chunk to user 1, cache 1 performs the OCR caching algorithm to decide if it needs to do regroup.

For hit density calculation, we only need to calculate the density between the existing users to the new user. Hence the calculation cost is linear.

3.3 Performance evaluation

In this section, we evaluate the performance of the proposed OCR and CCR schemes. The results are compared with the traditional LRU algorithm.

We assume the video length T_v is 7200s and the chunk size is 1M. The link transmission rate R is 200k/s. The initial time of a new user request is assumed to follow the Poisson distribution with a mean arrival rate of N/T_v. Here N is the number of users watching the video. The cooperative HR_{co} hit ratio defined as the hit ratio in the other caches is calculated in the following way:

$$HR_{co} = h_{co}/q_{tol} \tag{3.3.1}$$

where h_{co} is the total number of hits in the foreign caches, while q_{tol} is the total number of user queries including both served by the server and the caches. The total hit ratio is calculated as following:

$$HR_{tol} = (h_{loc} + h_{co})/q_{tol} \tag{3.3.2}$$

where h_{loc} is the total number of hits in the local cache.

Firstly, we evaluate the single cache to illustrate the benefits of OCR compared with LRU. Then we evaluate CCR. The simulation lasted for 36000 seconds and the results of the last 26000 seconds are collected.

In this section, we compared the performance of the OCR scheme with that of LRU. A single cache is considered under the server.

Case I: Impact of average user access rate in OCR

This experiment is set to investigate the performance by varying the average user access rate in the network. The cache size is 500M. The number of users is varied from 100 to 500. The average user access rate is N/7200.

Figure 3.6 shows the evaluation of the cache hit ratio versus the average user access rate. Increasing the average user access rate does not affect much on the hit ratio for both the OCR and LRU schemes. The OCR scheme can improve the hit ratio from 0.3 to 0.5, and hence greatly reduce the server load.

In Figure 3.7, the messages handled by the server with different user access rates are demonstrated. A server message is counted when an end user's request is solved by the server. It is obvious that higher frequent user access rate leads to more server messages. The OCR scheme performs better than the LRU scheme and the gain becomes higher as the average user access rate increases.

Cache II: Impact of cache size in OCR

In this case, we study the impact of different cache sizes, it varies form 50M to 450M, and the average user access rate is 300. Figure 3.8 demonstrates that the hit ratio improves

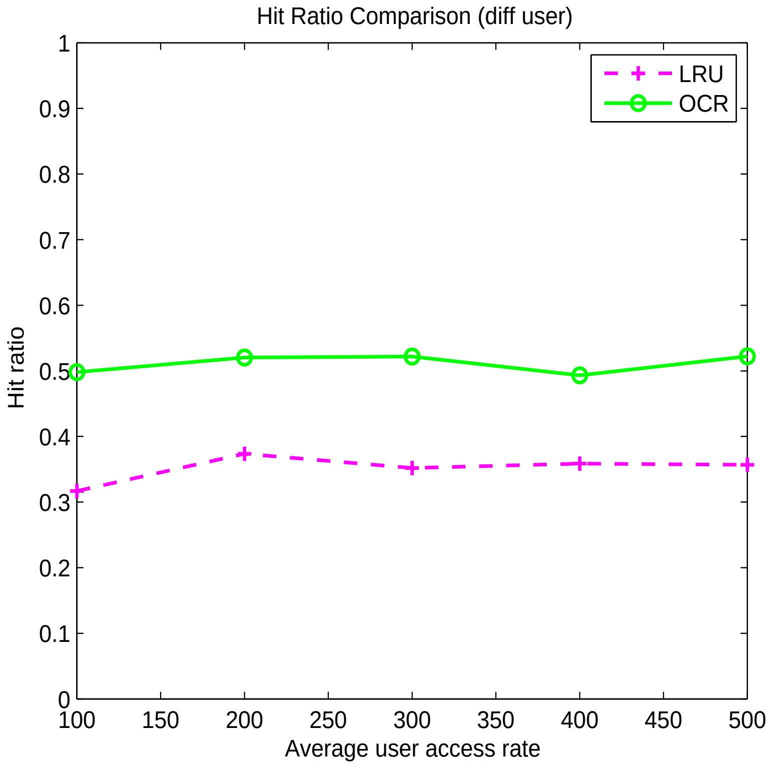


Figure 3.6: hit ratios with different users

for both schemes as the cache size increases. The performance of OCR is much better than that of LRU. For the small cache sizes (i.e. 100M, 250M), the hit ratio of OCR performs 2 times better than the LRU scheme.

From Figure 3.9, we can see that the server message of the LRU and OCR schemes by varying the cache sizes. For both schemes, the server message is reduced as the cache size increases. At cache size is at 350M, the OCR scheme can saves about 25% server messages compared to the LRU scheme.

Case III: Impact of average user access rate in CCR

To test the CCR scheme, two caches are simulated in this case. The performance of CCR, Cooperative LRU (CLRU) are conducted.

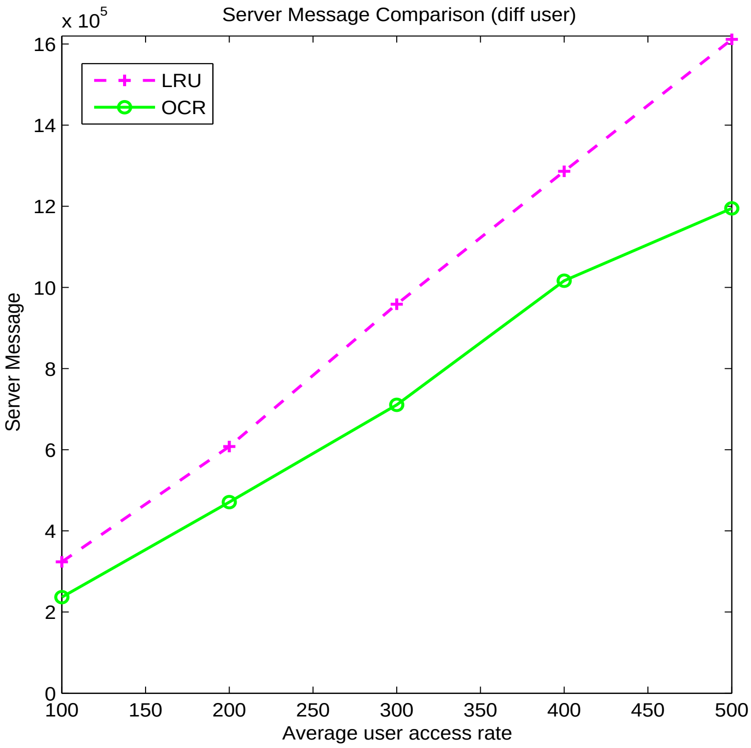


Figure 3.7: server Messages with different users

The cache size is set to 250M. The number of users varies from 100 to 500. At the same time, the mean arrival rate is set to N/7200.

From Figure 3.10, it is noticed that the hit ratios for all the schemes are almost stabilized at the same range as the increase of the average user access rate. This is because the increase of the average user access rate not only gives more hits but also affects the number of total requests. Intuitively, the number of users is not a factor that influences the hit ratio for each individual scheme. However, when compared the new scheme to the LRU scheme, we could see the benefit. CCR achieves the highest hit ratio, and then followed by the CLRU scheme. The hit ratio of CLRU is two times better compared with LRU, while OCR gives an increase of about one third.

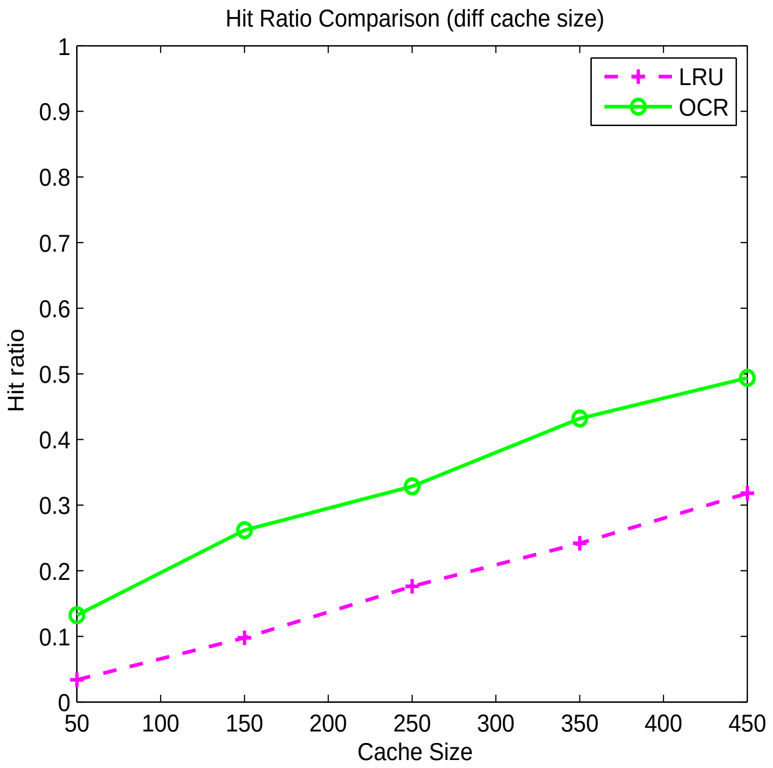


Figure 3.8: hit ratios with different cash sizes

Figure 3.11 shows the performance of the message of the server corresponding to the average user access rate. For all the schemes, the server messages increase with the raise of the average user access rate. However, the increase rate of the LRU scheme is much higher compared to the other three schemes. Among these schemes, CCR has the lowest increase rate.

Case IV: Impact of cache sizes in CCR

In this experiment, the impact of cache size for cooperative cache is evaluated. Same as the previous case, the size of cache is changed from 50M to 450M. The number of users is set to 300 under each cache and the average user access rate is 300/7200s.

Figure 3.12 demonstrates that the total hit ratio of these schemes. Obviously, all

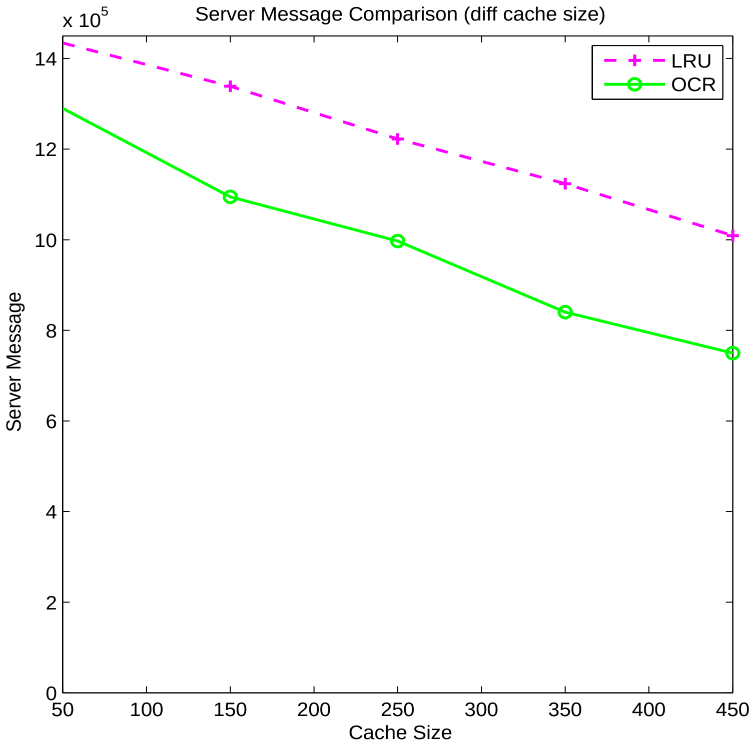


Figure 3.9: server Messages with different users

of them raise as the increase of the cache size. Similar to case I, CCR gives the best performance. Especially, the cache size at 250, CCR performs almost 1.5 times better than the CLRU scheme. At this situation, the hit ratio of CLRU is over 2 times of that in LRU and one half better compared with the CCR scheme. After a threshold, further increases the cache size only provide limited benefits. This indicates a relatively small cache could fit the acquirement from the users.

Figure 3.13 gives the performance of the server message versus the cache size. It is obvious that the LRU generates the highest number of the messages among the 4 schemes, while CCR achieves the best performance.

Case V: Impact of two videos

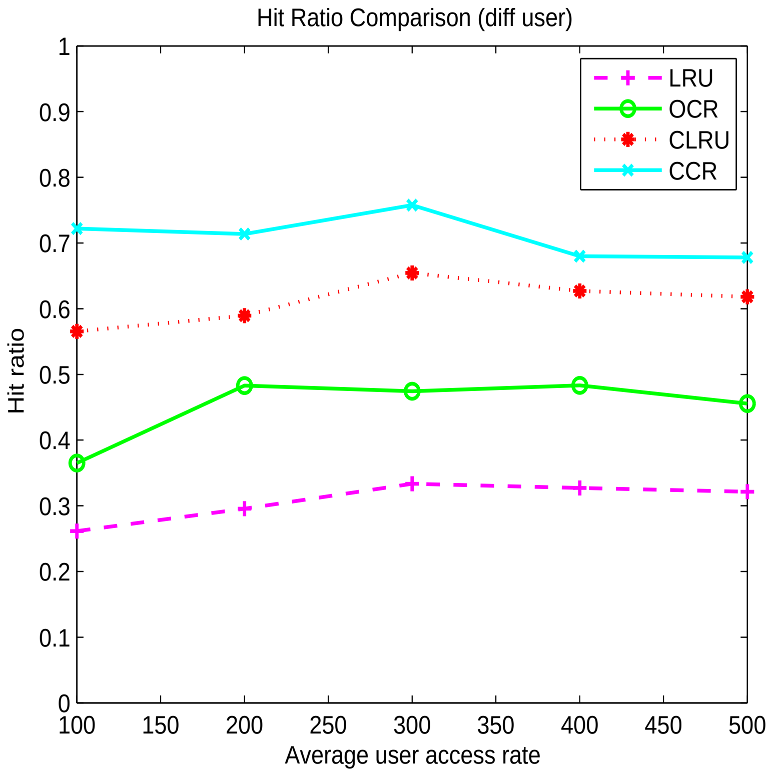


Figure 3.10: hit ratios with different users

In this case, we simulate two caches under a server, and 2 videos could be selected randomly by the users. The video length is 2 hours each. We compare the performance of CCR, CLRU, OCR and LRU.

In this simulation, we vary the average user access rate from 100/7200 to 500/7200 to study. The capacity of each cache is 250M. Figure 3.14 demonstrates that the hit ratio by varying the average user access rate. The hit ratio for CCR, CLRU, OCR and LRU all stay almost stable. Compared with 1 video case, CCR and OCR could gain much more benefits than these of CLRU and LRU.

From Figure 3.15 we noticed that the hit ratio increases as the increase of the cache sizes for all the schemes, among which CCR scheme have the best performance.

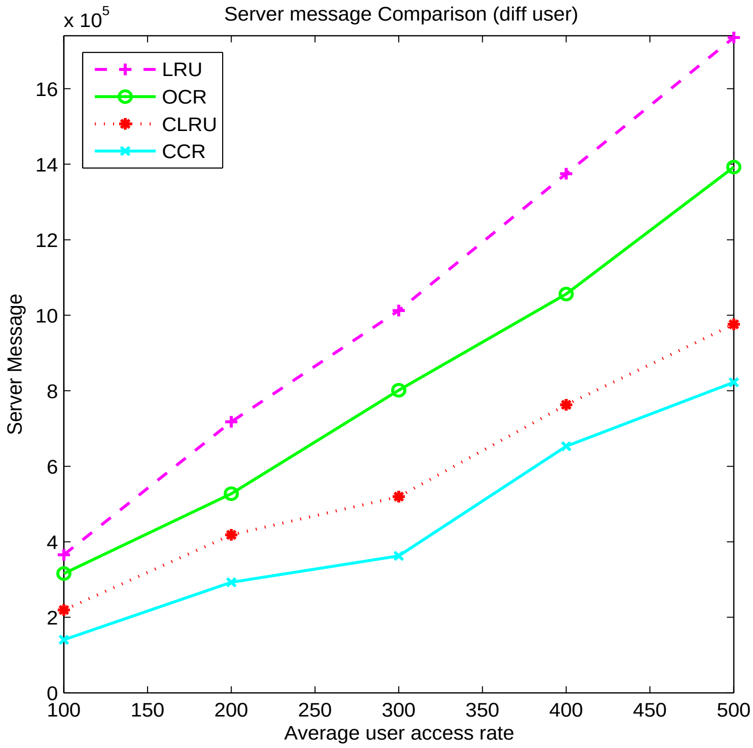


Figure 3.11: server Messages with different users

Figures 3.16 and 3.17 show the server message of the four schemes with different user access rates and cache sizes. It can be observed that the number of server messages reduces more significant for the proposed OCR and CCR schemes. Especially in Figure 3.17, when cache size is 450M, the OCR scheme can reduce about one third of the messages compared to the LRU scheme, while the CCR scheme can save almost two third of that in the LRU scheme.

3.4 Conclusions

In this chapter, we propose the OCR scheme and the CCR scheme for video stream. In the OCR and CCR schemes, the redundant chunks are abandoned and groups are

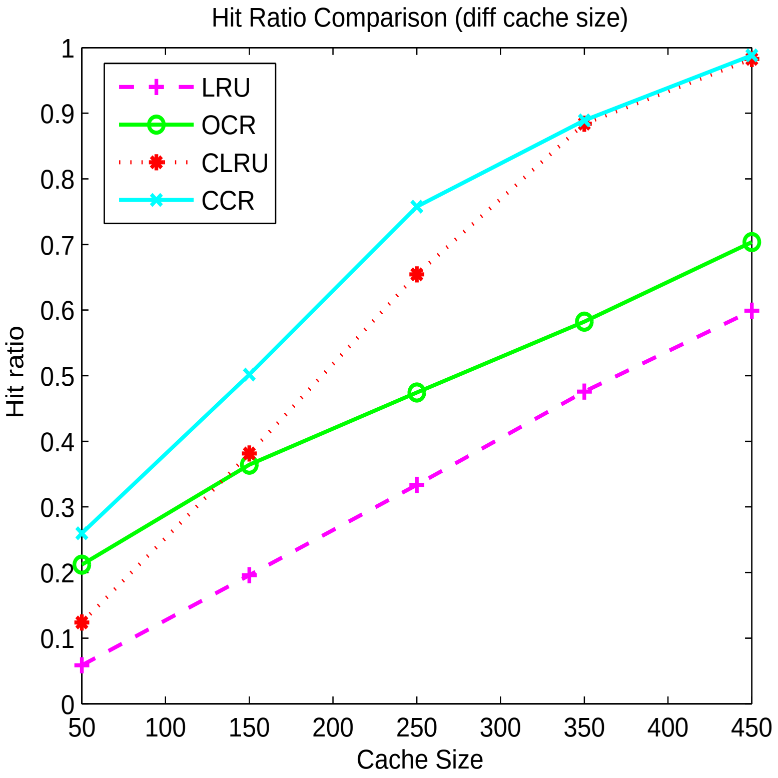


Figure 3.12: hit ratios with different cash sizes

built based on their group density, so as to maximize the hit ratio of the cache. CLRU and CCR benefit the performance of caches in a way that the miss in one cache may be assisted by another cache. With the new schemes, both the data access latency and bandwidth consumption on server can be reduced. The performances of the new schemes are compared with the LRU scheme, the results demonstrate a significant improvement on total cache hit ratio.

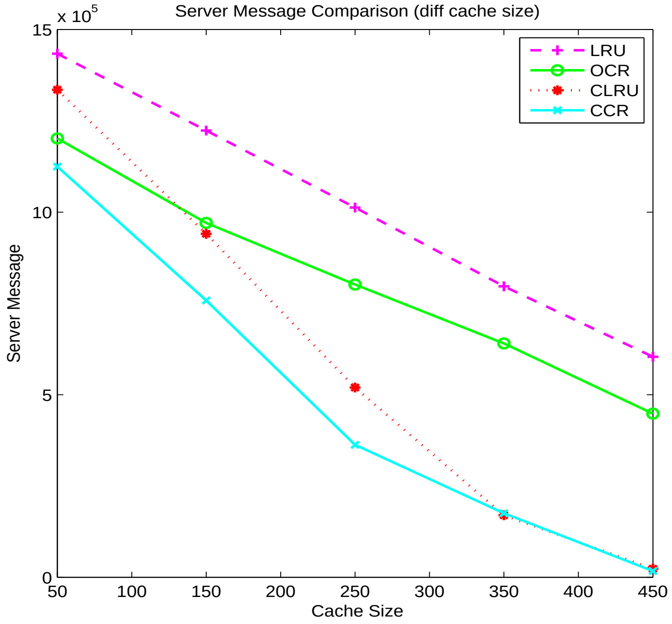


Figure 3.13: server Messages with different users

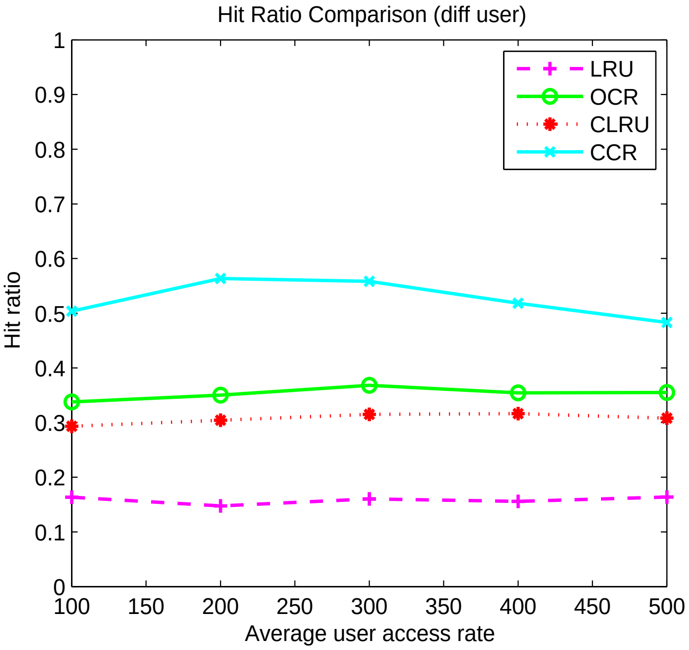


Figure 3.14: hit ratios with different users

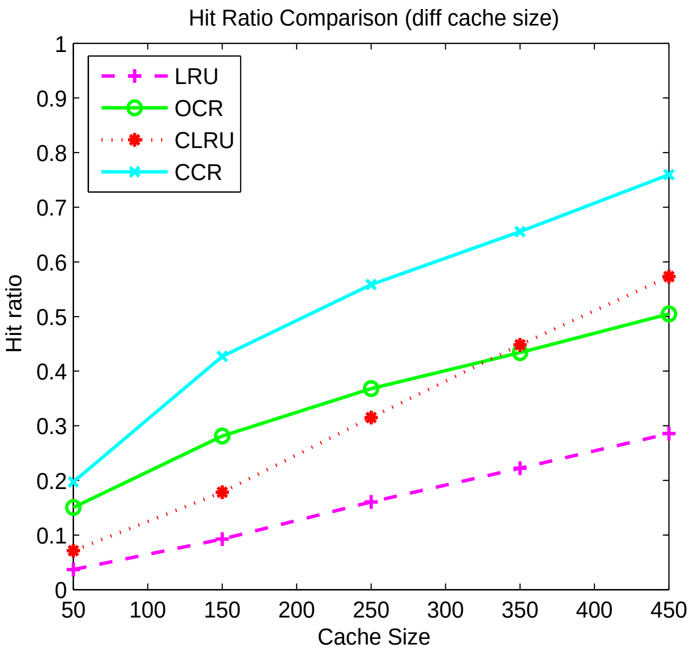


Figure 3.15: hit ratios with different cash sizes

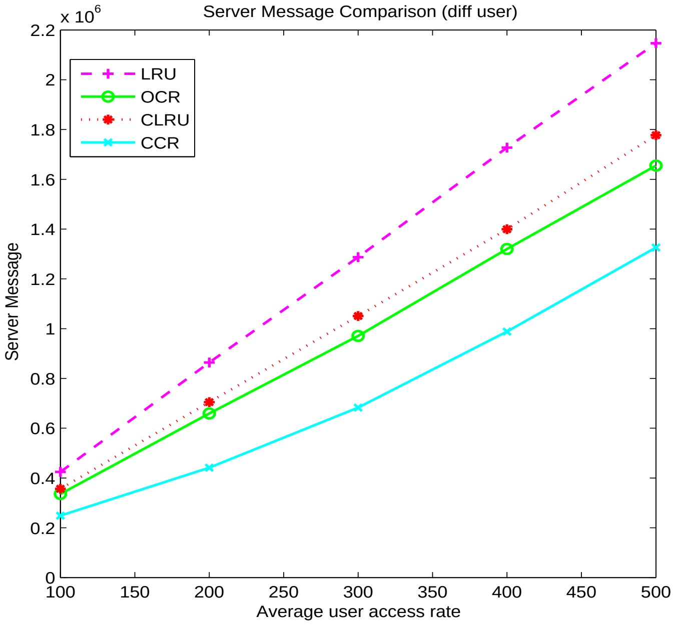


Figure 3.16: server Messages with different users

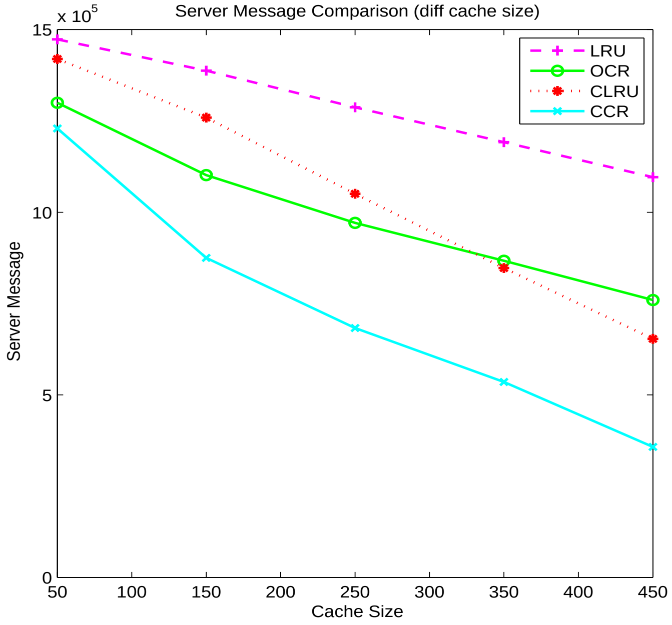


Figure 3.17: server Messages with different users

Chapter 4

An enabled Cache effect on Forwarding Table in Metro Ethernet

In this chapter, we describe an enabled Cache effect on Forwarding Table (CFT) in Metro Ethernet. The rest of this chapter is organized as follows: Section 4.1 gives the overview of this chapter. Section 4.2 presents the details of CFT. The performances are illustrated in Section 4.3. Finally, Section 4.4 concludes the chapter.

4.1 Overview

Ethernet is the dominant layer 2 communication technology. Currently, it has gained popularity for deploying in MANs. The traditional MANs are generally built based on layer 3 IP technology, usually using ATM as its layer 2 transport, whereas the LAN are dominated by layer 2 technologies, mostly based on Ethernet technologies. As a result, when a data frame is sent across a MAN, it may have to travel through multiple heterogeneous networks, and hence it may need to be re-encapsulated in different frame formats a number of times and even be split into multiple smaller frames/cells or merged with other frames/cells into a larger frame when it passes through the MAN. This not only makes the packet forwarding complex and inefficient, but also adds complexity and cost to the router/switch design, as well as network operation and management, in support of heterogeneous networks and interfaces. Clearly, deploying the same technology, such as Ethernet, in both MAN and LAN segments can potentially reduce the complexity and cost in network design and management, and improve packet forwarding performance. One of the main reasons that Metro Ethernet is highly attractive is due to its cost effectiveness. Ethernet provides the lowest per port cost, which increases sub-linearly as the port capacity increases. Besides, it requires minimal management and maintenance cost and is ubiquitously deployed. Network administrators are already familiar and comfortable with it. The third reason is ease of interoperability. Almost all the end systems support Ethernet interfaces. The fourth reason is convenience. The flexibility of creating virtual LAN segmentations enables enterprises that run geographically distant campuses, offices, and data severs could link distant campuses, offices and servers together, which brings significant value to those companies for effective information exchanges and data sharing.

Even though Ethernet has ample advantages, to be deployed in MANs, it has to solve the scalability issues. Ethernet was originally designed as a LAN technology that usually handles small number of users. Broadcast is frequently used to resolve unknown MAC addresses. Many protocols such as ARP and DHCP use the broadcast service as a service discovery mechanism. For example, to solve the MAC address for an IP address, an ARP request message is broadcast through the network so that the corresponding end user can receive and response an ARP replay message. The broadcast based address resolution schemes make Ethernet extremely convenient and easily accomplished. However, for MANs with millions of end users, high frequency broadcast messages waste a lot of bandwidth for address resolution. Specifically, as Ethernet uses a flat addressing scheme (i.e. non-hierarchical MAC addresses), Ethernet switches have the ability to automatically learn the location of end users by searching the frames they received and record the information in its forwarding table. Frequently broadcast frames accelerate the replacement of the table entries which may lead to forwarding table explosion and hence triggers excessive frame flooding. The MAC forwarding table at a node in a provider network needs to keep potentially a large number of MAC-to-port-mapping entries for frame forwarding. Moreover, every end user has to take resource to handle every broadcast message. VLAN technology is one way to reduce the broadcast message by logically segmenting the whole network into separate communication groups. However, even after segmentation the number of users located in one VLAN is still considerable in MAN. In this chapter, we propose a CFT in Metro Ethernet . In this scheme, we modified the forwarding table entries in the CE and PE nodes to learn both the IP and MAC addresses. By receiving a frame, the CE and PE nodes cache the IP-MAC address mapping carried in the frame. Once the mapping is remembered, the information can be served for the subsequent requests for the mapping. For protocols relying on broadcast to look for the MAC address of another end user, the designated CE and PE nodes could stop the broadcast and reply back directly with the help of the information stored in the new forwarding table. So, the CE and PE nodes not only could determine the next hope but also could handle the ARP request coming from the segments behind it. CFT is easy to be accomplished and fully backward compatible. We also combine CFT with EMiM encapsulation Scheme (details of EMiM is in next chapter). EMiM does MiM encapsulation by end users instead of the PE nodes, thus significantly reducing the PE node's forwarding table size.

4.2 Caching on Forwarding Table

Figure 4.1 shows a general Metro Ethernet including a provider network and multiple LAN segments. A provider network is composed of multiple PE nodes and CNs (switch or bridge). A LAN segment is composed of a CE node and multiple end users.

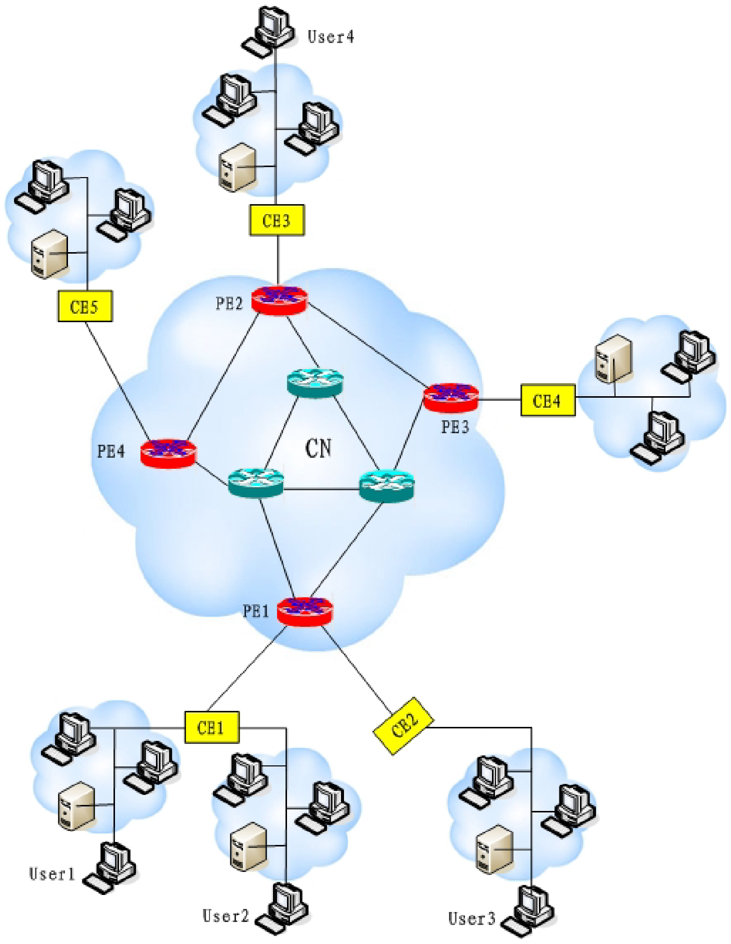


Figure 4.1: A Metro Ethernet

Ethernet nodes have the ability to dynamically learn the location of an end user by recording the port which the frame generated by the node comes from. Upon receiving a frame, a source address of the frame is learned in the forwarding table in an Ethernet node. A forwarding entry is in the format of <*MAC address*, *Port*, *Recordtime*, *age*>. The subsequent frames destined to this address passing through the node could be forwarded to that port. ARP is used to solve an unknown MAC address for an IP address. Once a user broadcast an ARP request frame to a network, all the other users attached to it in the same VLAN will receive this frame. Only the destination user with the corresponding IP address replies back its MAC address. The ARP reply is sent via unicast. Both users record the other's IP and MAC addresses. The other users received the ARP message(s) also can learn the IP and MAC address mapping and record them to its ARP table.

To enable the caching effect in the forwarding table, CFT adds a corresponding IP address to a forwarding entry. The IP address can be learnt when a forwarding entry is created. Note that the source IP address is associated with the ARP request and reply frame. Now a forwarding entry format in CFT is *<IP address*, *MAC address*, *port*, *recordtime*, *age>*. Here IP is the end user's IP address; *recordtime* is the time when the entry is created or refreshed; and *age* indicates if an ARP entry is valid or not. A modified forwarding entry caches its ARP entry in the forwarding table, and hence can be used to answer ARP request. we only deploy the CFT scheme in CE and PE nodes.

The address resolution and frame forwarding processes are as follows. When a CE or PE node receives an ARP request frame, it first checks its forwarding table. If the requested IP address is found in the forwarding table, a reply frame is sent back. The broadcast is stopped at this node. Otherwise, the request frame is forwarded. If the forwarding entry of the source MAC address is not in the table, the node creates a forwarding entry. When a CE or PE node receives an ARP reply frame, it records/updates the corresponding entry in the forwarding table and forwards the frame. The flow chart in Figure 4.2 demonstrates how an ARP frame is processed.

Now let us use Figure 4.1 as an example. Assume user 1 intends to start a data session with user 4 whose ARP information has not been known by user 1. User 1 broadcasts an ARP request to the network to solve the MAC address of user 4. When the broadcast frame reaches CE 1, CE 1 uses the IP address of user 4 to search its forwarding table. The broadcast frame is stopped once a corresponding entry is found, and then a reply is sent to user 1. If there is no corresponding entry in the forwarding table, the broadcast frame is forwarded to PE 1 as well as the LAN segment user 2 belongs to. Similarly, after PE 1 receives the broadcast frame, it first searches its forwarding table. PE 1 broadcasts the

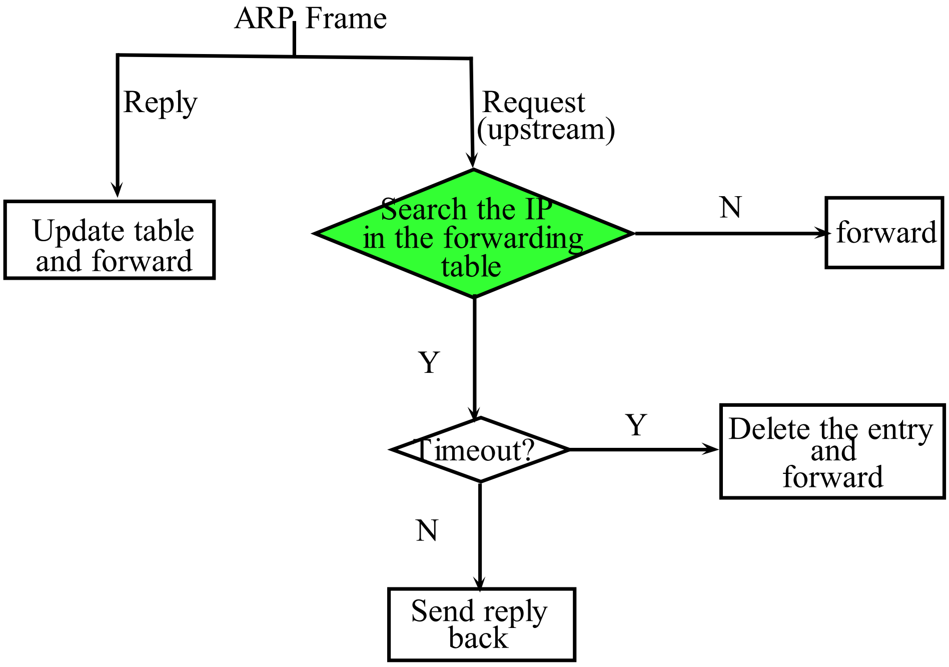


Figure 4.2: Process of a CE/PE handling an ARP frame

request to the core network if no entry is matched. Otherwise, PE 1 stops the broadcast and replies an ARP reply back to user 1, while both CE 1 and PE 1 add the entry of user 1 or update the *recordtime* of user 1 in its forwarding table . Other PE and CE nodes transmit the broadcast frame process the frame as the same way as in legacy VLAN scheme. If the request frame reaches user 4, it just sends out a reply frame back. All the CE or PE nodes receive the request or reply frame learn the corresponding entry. Finally, user 1 could communicate with user 4.

4.2.1 IP management in forwarding table

There are two ways to manage IP addresses. One way is to insert an IP address in each forwarding entry. Then an IP address is searched linearly when an ARP request comes. In this method, each entry only increases 4 bytes to store an IP address. A forwarding entry takes about 4/15 (4 bytes for IP address, 6 bytes for MAC address, 4 bytes for recordtime, 4 bytes for age and 1 byte for port number) more space. But the IP lookup time takes long. The second way is to divide the forwarding table into two parts. One part contains the index and the IP address which are sorted. The other part is the original forwarding table by adding an index for each entry. The two parts are united by an index. A binary search can be used to find an unsolved IP address. In this method, at most $\llcorner \log_2(N)+1 \lrcorner$ lookups are needed for an IP address matching. Here N is the number of IP addresses on the table. If the IP address has been found, the index is used to find the corresponding MAC address. The two tables are always updated at the same time. If the entry of a MAC address is outdated and deleted, the IP address with the same index will also be deleted. The table format is shown in Figure 4.3. This method can reduce the IP lookup time, but it costs two indexes space in the forwarding table. Assume that an index needs 2 bytes, then each entry needs 8/15 more spaces. Due to cache effect, a PE or a CE node handles less number of broadcast frames and hence the number of entries is reduced in CFT. The size of each entry is increased, but the total table is not increased as we will see later.

As the cache entry is timed out every 2 minutes, there is no problem on dynamic IP. If the user changes its IP within 2 minutes, some misleading problems may happen. The misleading frames will be dropped after some time and the user will notice this as no reply came back and start new sessions. But this should not be the frequent case.

4.3 Cache in EMiM

The EMiM encapsulation scheme [93] can reduce the forwarding table size in PE nodes. In EMiM, the MiM encapsulation is done by the end user instead of the PE node. To allow an end user to do MiM encapsulation, the PE node's MAC address is associated with an ARP entry. The associated PE node's MAC (PMAC) address allows an end user to do MiM encapsulation, and hence a PE node does not need to maintain the entries of

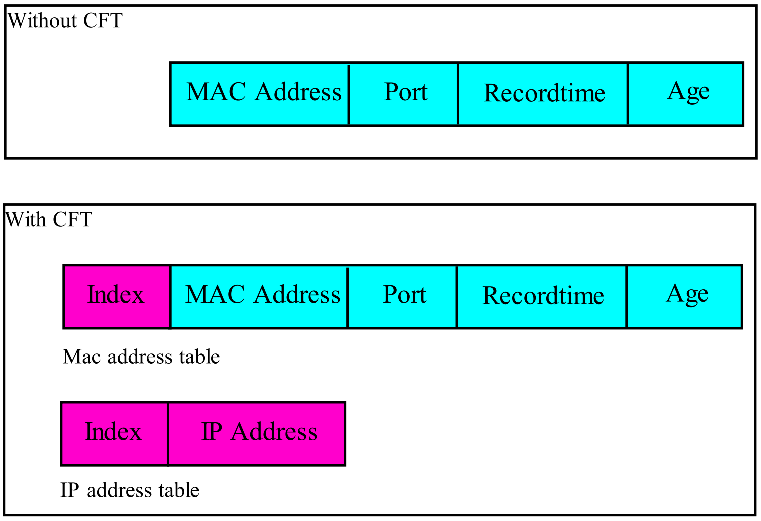


Figure 4.3: Forwarding Table Format

mapping End user's MAC address (EMAC) to PE node's MAC address, thus reducing the forwarding table size. The entry format of the ARP table is <*IP*, *EMAC*, *PMAC*, *recordtime*, *age*>.

In the EMiM scheme, the forwarding table has no end user's entry, and hence the forwarding table cannot be used to cache ARP entry (CE node still can use CFT to cache ARP entries). But a PE node can still learn an ARP entry from an ARP request/reply frame. Hence we use a cache table to store ARP entries. A cache entry includes the *MAC address*, *IP address*, *PMAC address* and *recordtime*. The cache size is fixed. A LRU replacement algorithm is used for entry management. This is because LRU can be easily implemented and an ARP entry is automatically timed out after some time. Besides LRU, a Modified LRU (MLRU) replacement algorithm is also introduced. In MLRU, the cache table is divided into two separate lists. List 1 is for entries that used more frequently, while list 2 is for entries with less frequency. When a new entry comes, it is stored in the head of list 2. When an entry is hit, it is put to the head of list 1. When list 1 is full, the entry in the tail is put to the head of list 2. If list 2 is full, the entry in the tail is deleted to release the space for new entries. Hence an entry has high hit ratio will stay longer than that in LRU, and hence MLRU can increase the hit ratio.

4.4 Hit ratio analysis

In this section, an analytical model is developed to examine the hit ratio in the CFT scheme at a PE node. The following assumptions are made in our model:

- The data access (to send frame) time of an end user follows Poisson distribution. The average data access time is T_a (one communication session per T_a seconds). R_i is the data access rate of user i.
- All the end users belong to one VLAN. Each user has probability to communicate with any other end users in the network.
- When an end user starts a data session, the destination is selected based on Zipf distribution with Zipf coefficient n.
- The forwarding tables are sufficient large to store all the entries and each forwarding table entry is timed out every T_o seconds.

Using Zipf's law, the probability of a user with rank i being selected is:

$$f(i, n) = (1/i^n) / \sum_{j=1}^{N} (1/j^n) \qquad (4.4.1)$$

where N is the number of users;

The rate a node being selected as the destination by all the nodes under a PE node is:

$$\lambda_i = f_i * \sum_{j=1}^{M} R_j \qquad (4.4.2)$$

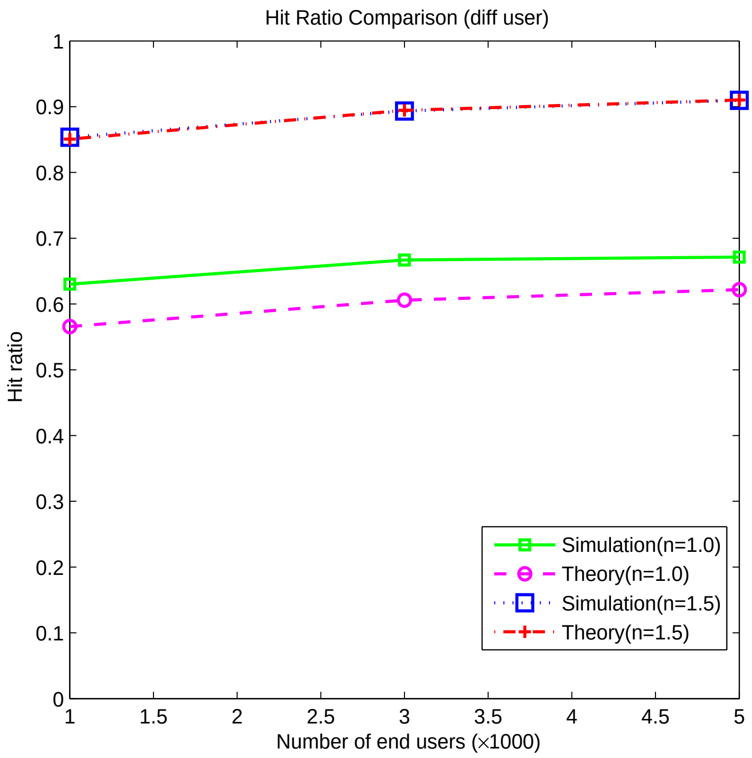


Figure 4.4: Hit Ratio Comparison (diff user)

here M is the number of users under the PE node.

When the entry of node i comes to the forwarding table, it has T_o time to be stored there. If it is selected as a destination node in less than T_o time by a node under the PE node, a hit is counted. Hence the hit ratio can be calculated as:

$$h_i = \int_0^{T_o} \lambda_i e^{-\lambda_i t} dt = 1 - e^{-\lambda_i T_o} \tag{4.4.3}$$

The total cache hit ratio for the PE node is:

$$h = \sum_{i=1}^{N} (f(i, n) h_i) \tag{4.4.4}$$

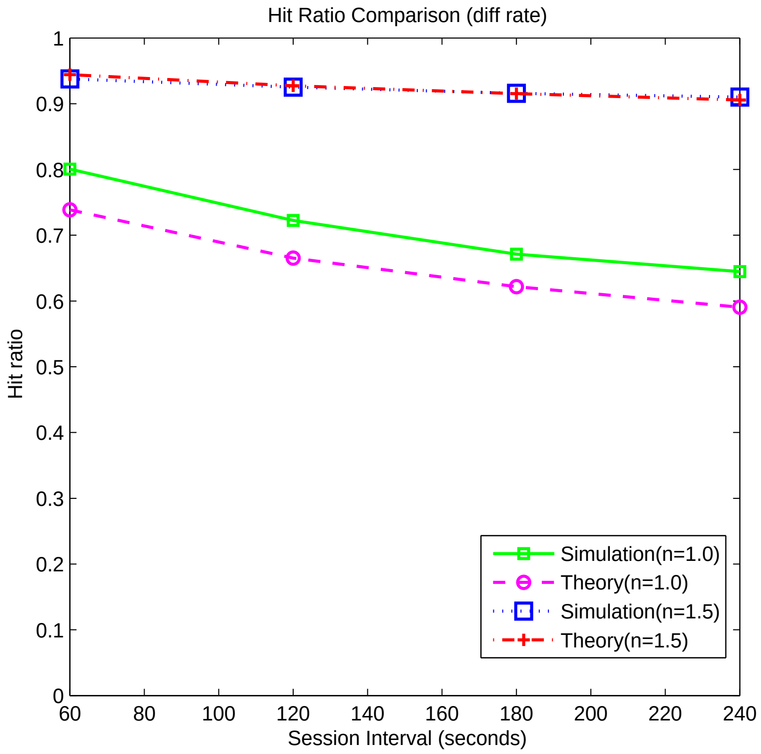


Figure 4.5: Hit Ratio Comparison (diff rate)

We compared the simulation results with the analytical results. The simulated Metro Ethernet have 4 PEs in the network. Every PE node has the same number of end users under it. We run 3600 seconds and the results of the last 2600 seconds are collected. The average data access rate is set to 180s. All the forwarding tables are set to be empty at start-up.

Figures 4.4 and 4.5 show the simulation and analytical modeling hit ratio with various value of N and R. From Figure 4.4, we know that the hit ratio increases as the increase of the number of end users. Figure 4.5 compares the results by varying the data access rate. The simulation with n=1.0 has 5 thousand end users while for n=1.5 has 6 thousand end users in the network.

The results show that the analytical results are well matched with the simulation results, particularly for n=1.5.

4.5 Evaluation

This section compares the performances of the proposed schemes and the legacy VLAN based scheme. Both the impact of session interval and number of users are considered. In this section, we simulated a Metro Ethernet with 50 thousand end users, 40 PEs, 150 CEs and 1000 VLANs in the network. Every PE node directly connects to at least 2 and at most 5 PE nodes. At least 2 and at most 6 CEs are behind a PE node. And every CE node can have up to 6 sites located under it. There are at least 8 and at most 256 end users behind a CE node. Each VLAN has at least 3 and at most 13 sites. Each user has probability to communicate with other end users in the same VLAN. The data access (to send frame) time follows Poisson distribution. The average data access time is T (one communication session per T seconds), and each data session lasts for average 20 seconds, randomly chosen between 1 and 39 seconds. When an end user starts a data session, it randomly picks a VLAN it belongs to, and then randomly chooses another end user in the same VLAN as the destination user. The destinations are selected based on Zipf distribution. That means the probabilities of session destined to few addresses are heavier than others, which is more suitable for the real world traffic situation [116]. Using Zipf's law, the frequency of a user of rank i being selected out of a population of N is:

$$f(i,n,x) = (1/i^n)/\sum_{j=1}^{N}(1/j^n) \tag{4.5.1}$$

Where: N is the number of users; i is their rank; n be the value of the exponent characterizing the distribution. The default forwarding table size is set to store at most

20000 entries, and each forwarding table entry is timed out every 120 seconds. In the simulation, we run 1600 seconds and the results of the last 600 seconds are collected. All the forwarding tables are set to be empty at startup. We compare the performances of the proposed scheme with the legacy VLAN based scheme. Three aspects are considered: the impact of session interval, the impact of Zipf exponent and the impact of cache size. To evaluate the impact of session interval, we varied the session interval from 60s to 240s in case 1. In case 2, we investigated the performance of the new scheme under different Zipf exponent. In case 3, we varied the cache table size in CFT-EMiM to see the impact of different cache table size. The data access rate is set to 60s for case 2 and case 3. In the simulation, PE node messages include the usual datagram messages PE node handles and the broadcast/multicast messages PE node generates to resolve unknown MAC addresses. User messages contain all the datagram messages and the broadcast/multicast messages users generate to resolve unknown destination MAC addresses.

Case 1: Impact of session interval

This case is set to investigate the performances of our proposed schemes and the legacy schemes as the increase of session rate.In this case, T varies from 60 to 240 seconds. Zipf coefficient n is set to 1.0. Figures 4.6 and 4.7 demonstrate the average number of messages a user handled per second and the average number of messages a PE node generated per second respectively. From the Figures, we can see that the CFT scheme can significantly reduce the messages handled by PEs or end users compared to the legacy VLAN scheme, particularly in heavier traffic case. Note that session interval is the reciprocal of data access rate. As the data access rate increases, the number of message grows more rapidly in legacy VLAN system. This is due to more broadcast frames needed to resolve unknown MAC addresses. But in the CFT scheme, more traffic results in high cache hit ratio and hence reduce more broadcast messages. Hence the number of messages increases much less than that in legacy VLAN scheme. At the session interval

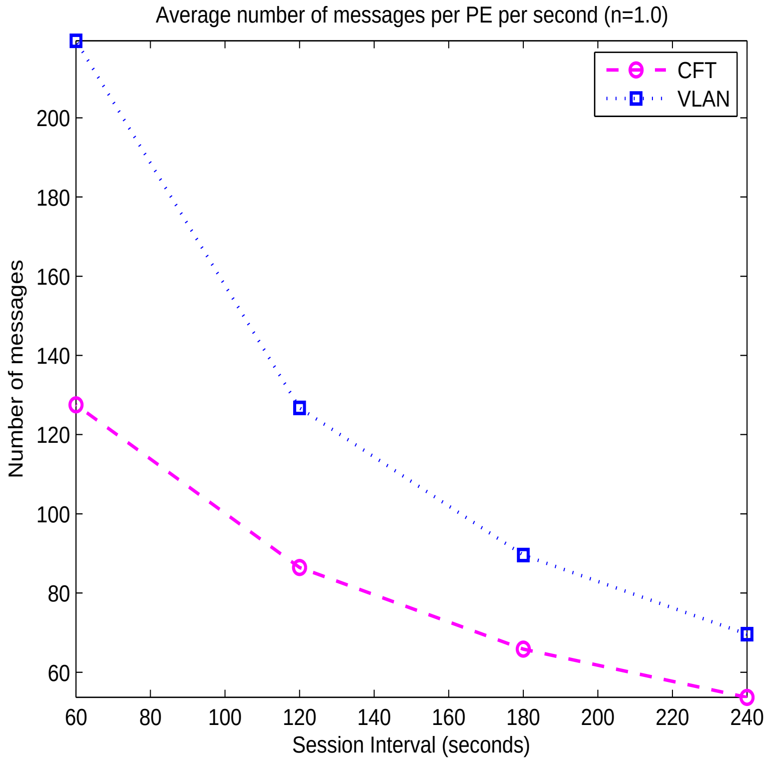


Figure 4.6: Average PE message

equals to 60s, around 50% PE messages can be reduced for CFT per PE node, and about 30% messages can be reduced for each end user.

Figures 4.8 and 4.9 illustrate the average table size and the maximum table size in a PE node. Figure 4.10 and Figure 4.11 show the average table size and the maximum table size in a CE node. Both the average and maximum table sizes increase as the decrease of session interval. This is due to shorter session interval results in high broadcast and hence more forwarding entries. The results clearly show that the CFT scheme can reduce both average and maximum table size in PE and CE tables. As we pointed out that the CFT scheme needs more space for each forwarding entry, but the maximum table size is reduced and hence the total memory space is also equal.

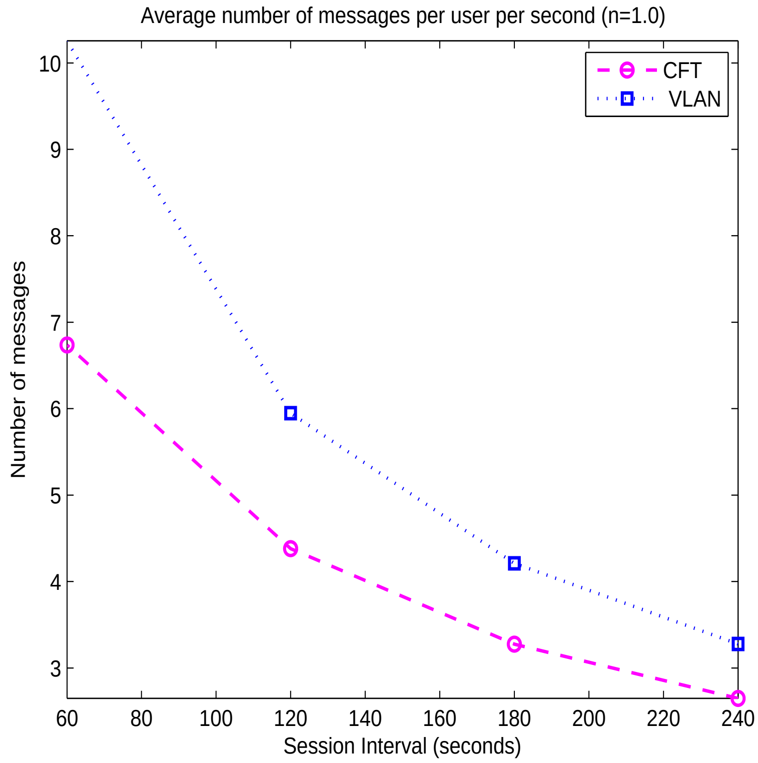


Figure 4.7: Average User message

The hit ratio of CFT is plotted in Figure 4.12. The hit ratio is computed as the number of ARP requests answered by the cache divided by the number of ARP requests received by a PE or CE node. The result shows that the hit ratio is reversely proportional to the session interval time. This is due to less session interval time corresponding to high broadcast requests, and hence high probability to be answered by cache. When the data interval time is at 60 seconds, the hit ratio reaches about 45%, this can significantly reduce the broadcast messages.

Case 2: Impact of zipf exponent

In this section, we vary the Zipf coefficient from 0.5 to 1.5. A high Zipf coefficient stands for the destination users are more focused on the popular users.

The average message number a PE node (a user) handled per second is demonstrated

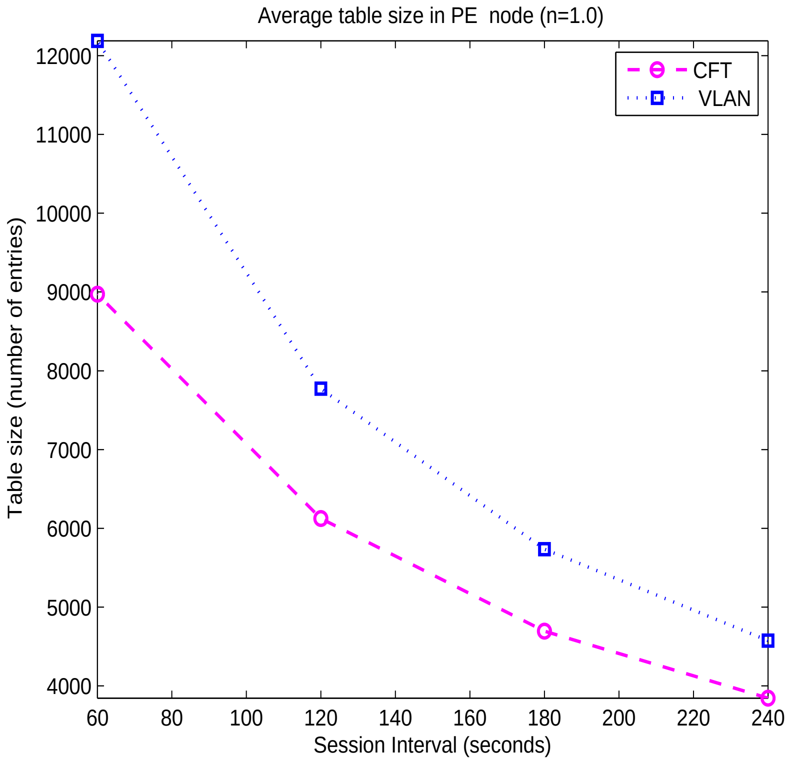


Figure 4.8: Average PE table size

in Figures 4.13 and 4.14. The CFT scheme has better performance for larger Zipf coefficient. This is because the destination users are more focused on the popular ones for a larger coefficient system than that for a smaller coefficient system. At *n* =1.5, more than 60% PE messages can be reduced for average PE node.

Figure 4.15 and Figure 4.16 report the average and max PE table size of the schemes. VLAN based Ethernet requires more state than CFT scheme because of the nature of broadcasting and inherent source address learning. The benefit of the proposed scheme becomes more obvious with bigger zipf coefficient. When Zipf coefficient is 1.5, the maximum table size is reduced about 40% in CFT. The average and max table size in CE node are shown in Figure 4.17 and Figure 4.18. Similarly, the CE table size is reduced as the Zipf coefficient increases. A larger Zipf coefficient results in fewer destination users

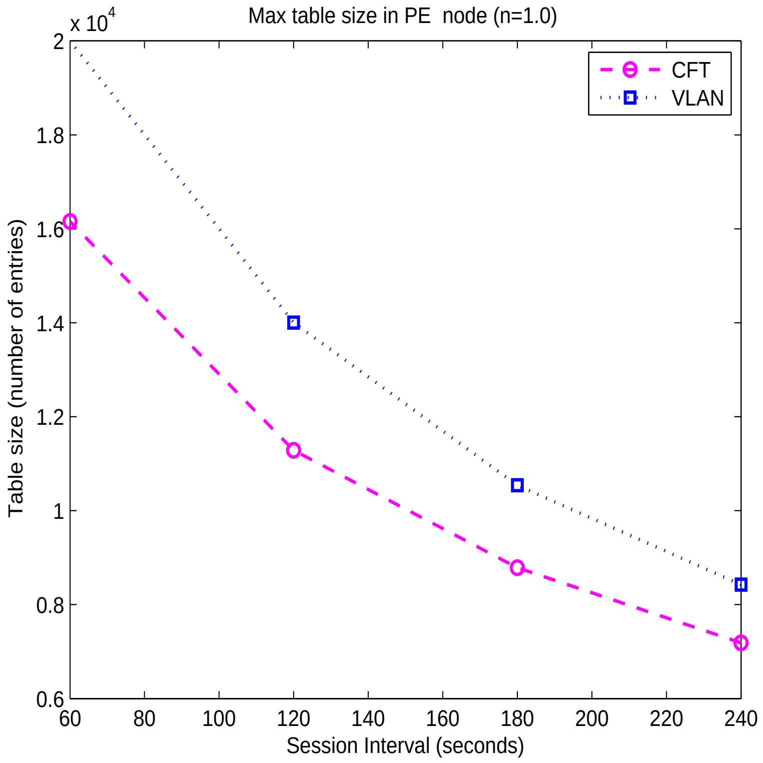


Figure 4.9: Max PE table size

and hence fewer number of forwarding entries in the forwarding table.

The explanation for the improvement is: When the zipf coefficient increases, the chance users start sessions with the same destination increase. The replacement of the table entries that are used more often are refreshed frequently. This means fewer table entries will be deleted as the utilization frequency for individual table entry is increased. This in turn improves the overall traffic load as more broadcast traffic can be intercept by the CE and PE nodes.

Figure 4.19 shows the hit ratio varying with the Zipf coefficient. When the Zipf coefficient increases, the chance users start sessions with the same destination increase, and hence the hit ratio increases. At n =1.5, the hit ratio can be over 60%. Hence the CFT scheme is more efficient for a system with some popular users.

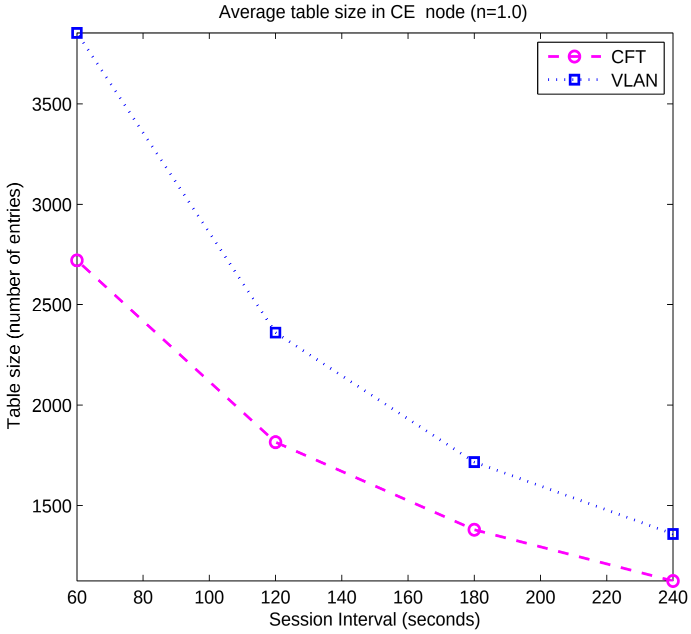


Figure 4.10: Average CE table size

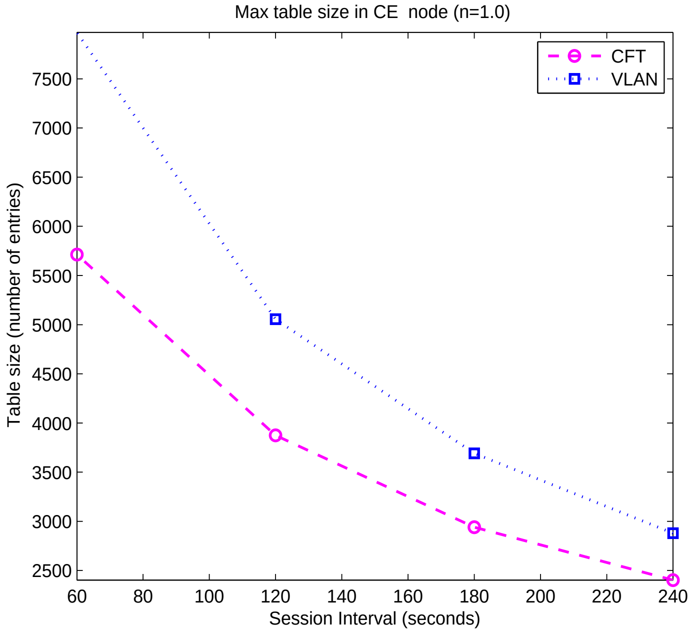


Figure 4.11: Max CE table size

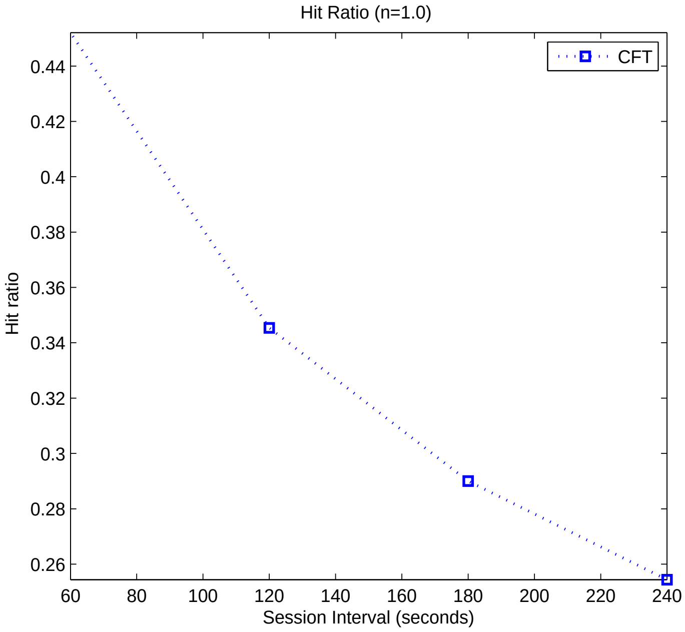


Figure 4.12: Hit Ratio

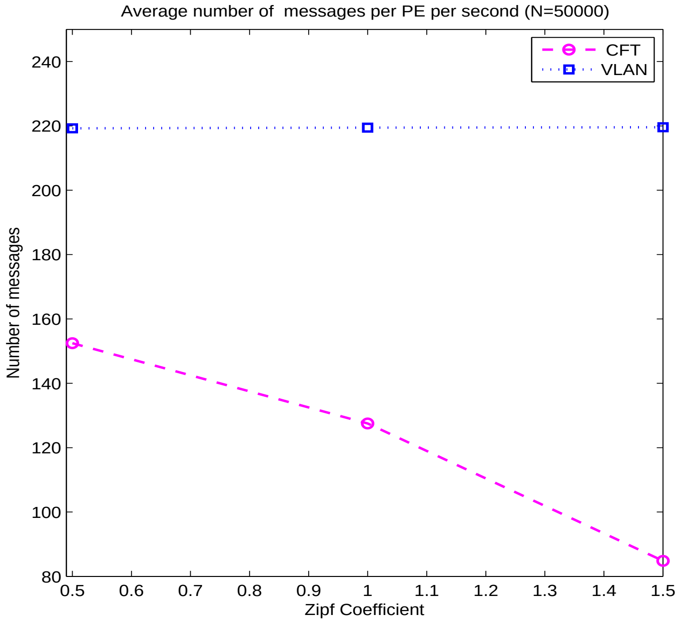


Figure 4.13: Average PE message number

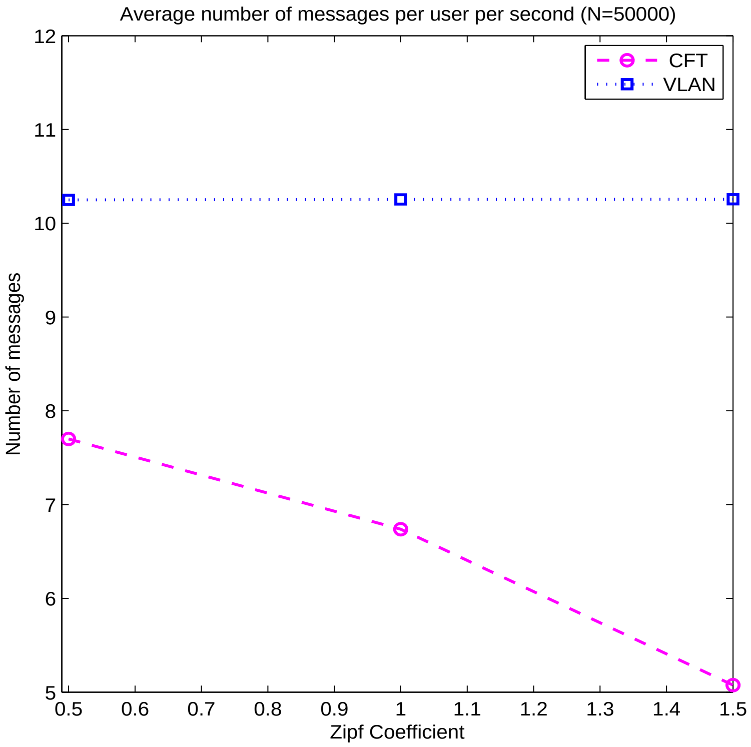


Figure 4.14: Average User message

Case 3: Impact of cache table size

In this case we study the impact of varying the CFT-EMiM ARP cache table size. In this experiment, the network has 50000 users and the destination is selected based on zipf distribution with 1.0 exponent. The table sizes in CFT-EMiM include both the forwarding table size and the added cache table size. The data in VLAN scheme does not change along with the increase of the cache size as there are no caches in this scheme.

Figure 4.20 and Figure 4.21 illustrate the average number of messages handled by every PE node and every end user per second respectively. The decrease rate of CFT-EMiM becomes slower along with the increase of the cache size. For PE messages, from 500 to 1500, more than 12% message can be further reduced compared with the VLAN scheme, while from 1500 to 2500, only less than 3% more messages could be saved. The

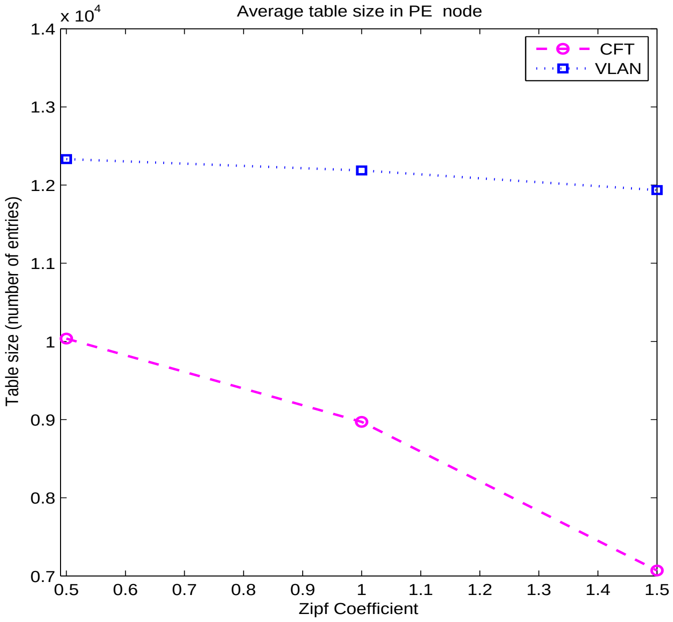


Figure 4.15: Average PE table size

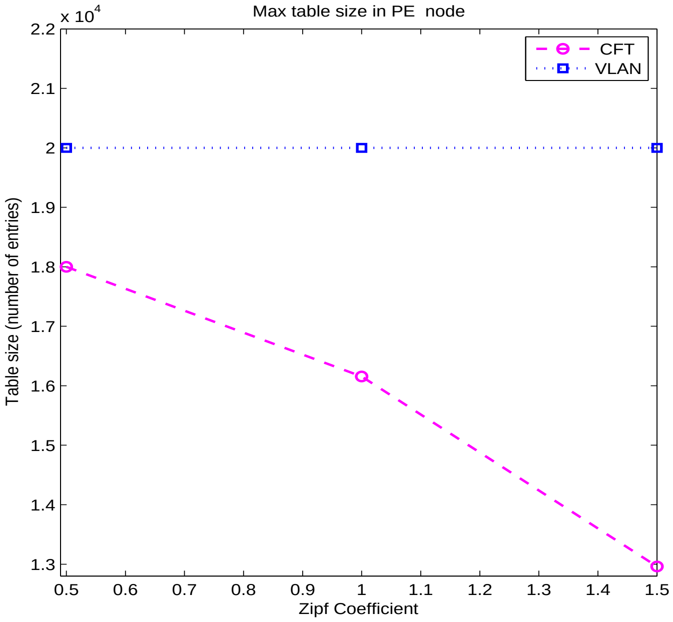


Figure 4.16: Max PE table size

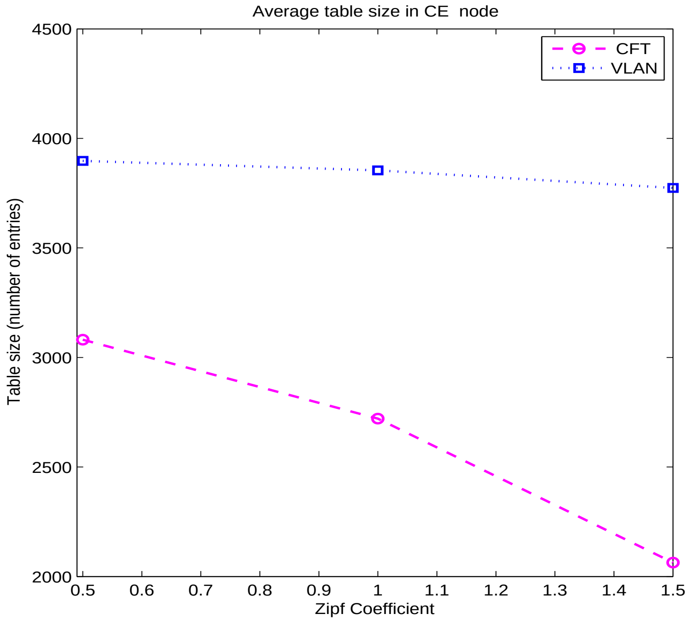


Figure 4.17: Average CE table size

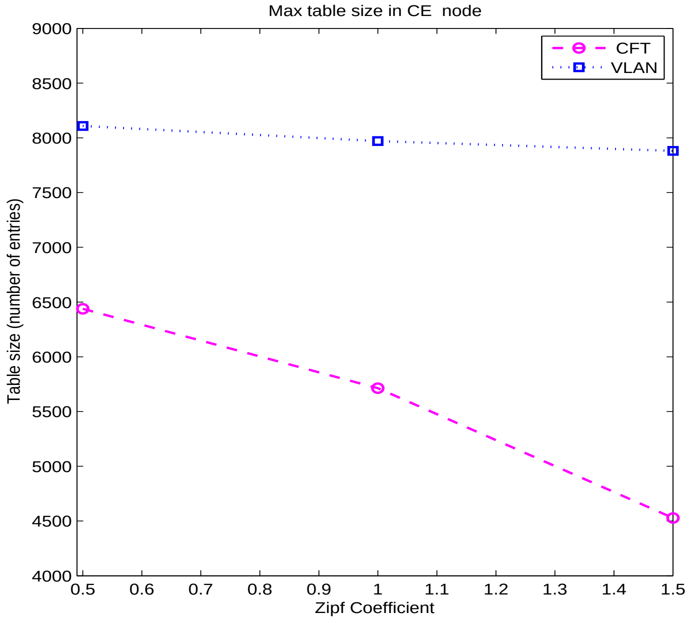


Figure 4.18: Max CE table size

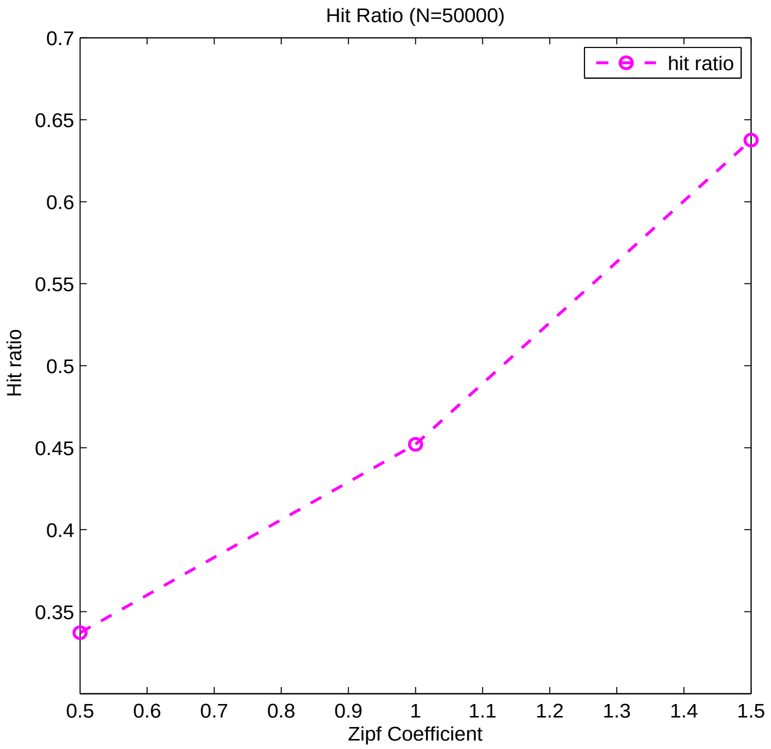


Figure 4.19: Hit Ratio

increase becomes less and less impressive. The similar tendency also appears in Figure 4.21. The increase of cache size does not offer unlimited benefit in reducing messages.

The average and maximum table sizes in PE nodes are shown in Figures 4.22 and 4.23. It is obvious that even adding the cache table, the total table size in CFT-EMiM is significantly smaller compared with the VLAN scheme. The average table size does not grow linearly with increased cache size. This is expectable as the following explanation illustrated. Let α be the cache size for one PE node to store all the entries needed in our simulated network and β be the cache size constraint. When β less than α, that means this PE node does not have enough space, an existing entry may be pushed out of the cache by a new entry. So cache miss occurs and this in turn triggers more broadcast which heavies the message number. When β increases, less cache entries are replaced and thus

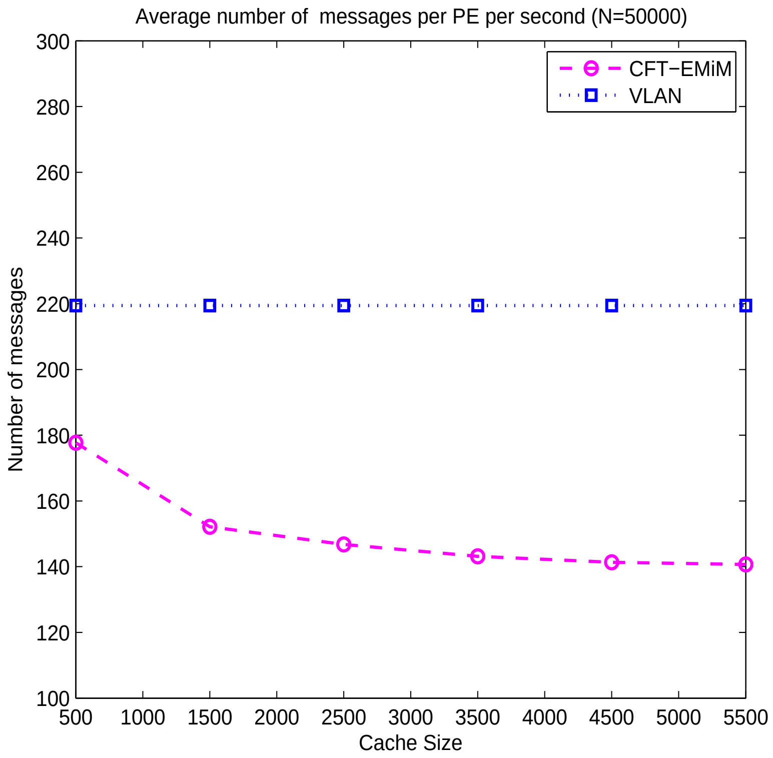


Figure 4.20: Average PE message number

less broadcast happens. However, this improvement is bounded at β equals α. Passing this threshold, as the cache size increases further, the table size in this PE node does not be affected due to it has already reach its own maximum.

Figure 4.24 and Figure 4.25 plot the average and maximum table sizes in CE node. The table sizes almost stay stable after cache size reaches 2500 entries, as in a network with 50000 end users and Zipf distribution, this cache size is sufficient for CE nodes to store all the entries.

Figure 4.26 demonstrate the cache hit ratio in PE and CE node when varying the cache size from 500 to 6000 entries. From these figures, it is noticed that a comparable small table size can perform excellent in hit ratio.

Case 4: Impact of number of users

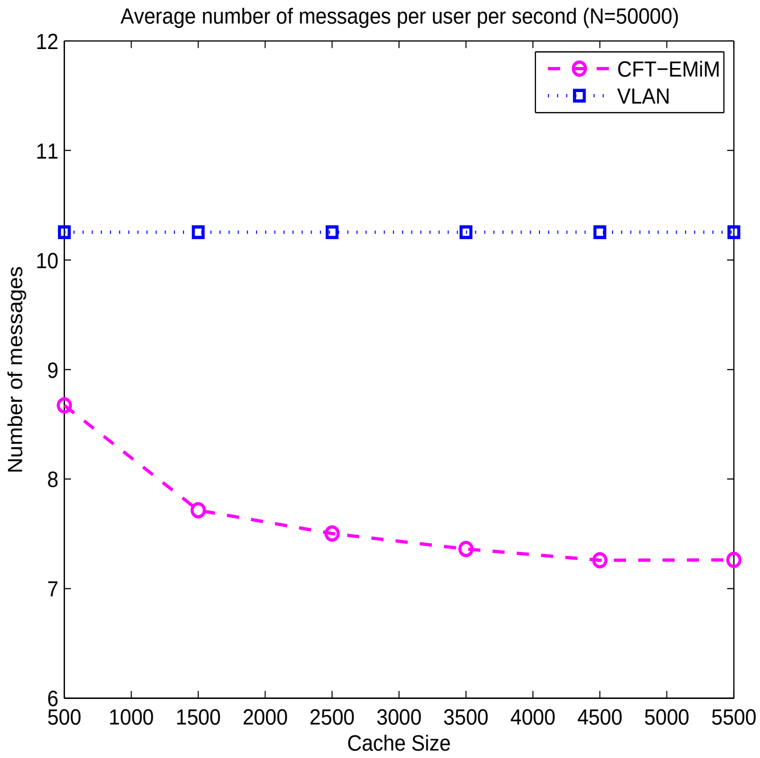


Figure 4.21: Average User message

In this case, we vary the number of users from 10,000 to 35,000 to study the impact of the number of users. The data access rate is set to 120 seconds and the Zipf coefficient is set to 1.0. We compare the performance of the CFT scheme combing with EMiM and the VLAN scheme. The cache size in EMiM is large enough to record all the entries.

Figures 4.27 and 4.28 show the number of messages a PE node and an end user handled per second respectively. The number of messages increases as the number of users increases. As the number of users increases, the number of broadcasts increase and hence a PE node and a user need to handle more messages. The number of messages for all the three schemes increases linearly. But the VLAN scheme has larger slope, i.e., increase rate higher than the other two schemes. Caching in EMiM has similar slope as that of the CFT scheme, but a little bit more messages than that of the CFT scheme.

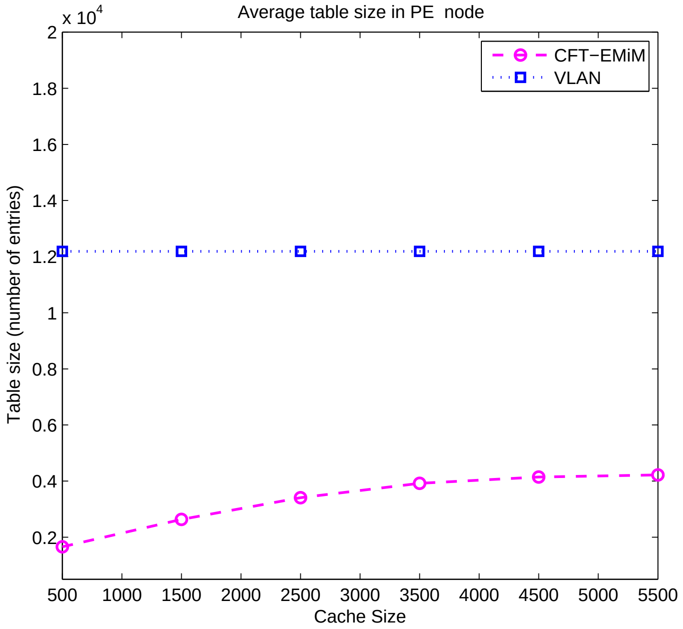


Figure 4.22: Average PE table size

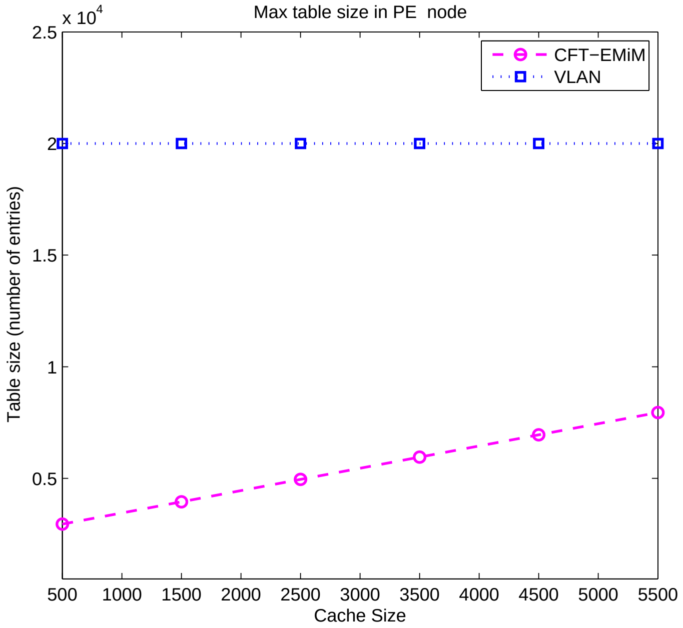


Figure 4.23: Max PE table size

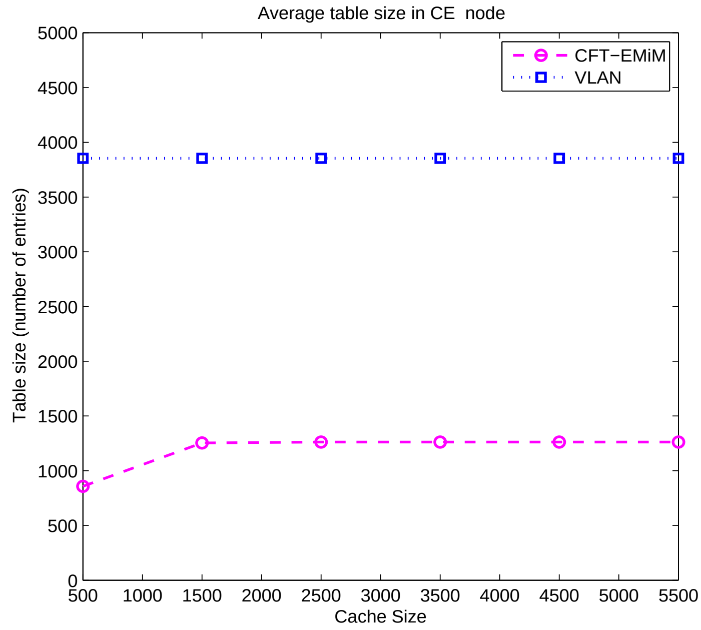


Figure 4.24: Average CE table size

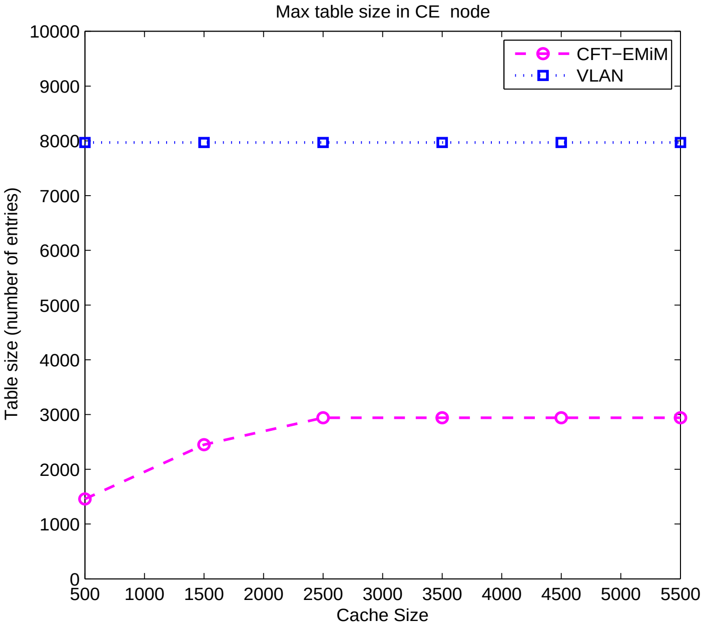


Figure 4.25: Max CE table size

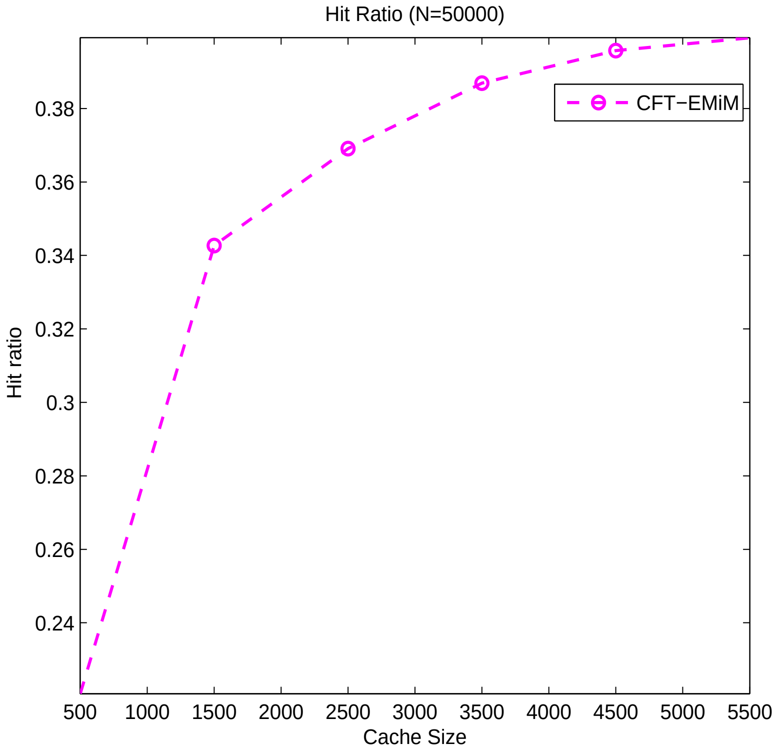


Figure 4.26: Hit Ratio

From Figure 4.29, we observe that the hit ratio increases along with the increase of the number of users when the number of users is less than 30,000. However, when the number of users reaches 30,000, the hit ratio is almost a constant by further increase the number of users. When the number of users is small, increasing the number of users can enlarge the probability of the most popular users to be selected as the destination nodes, and hence increase the hit ratio. But when the number of users is over a threshold, more and more unpopular users have been selected and hence make the hit ratio nearly a constant.

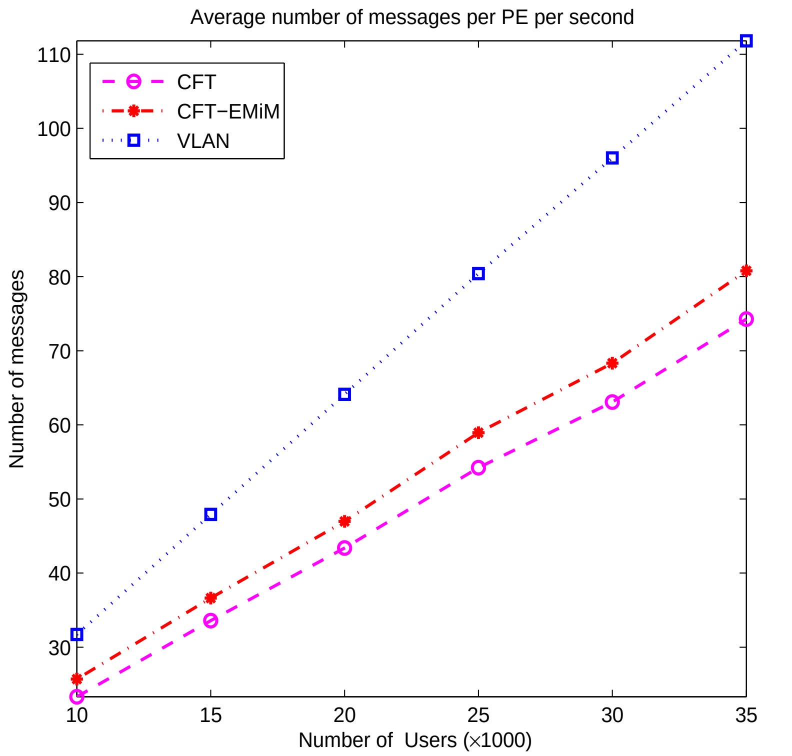


Figure 4.27: Average PE message per second

4.6 Summary

This chapter proposes an efficient and scalable Metro Ethernet architecture. The CFT scheme learns the IP and MAC mapping in a frame and eliminates the subsequent broadcast frames looking for this mapping. Combined with EMiM which allows an end user to encapsulate both the destination users MAC address and its PE nodes MAC address in the frame, the forwarding table sizes of PE node can be further reduced. The simulation results show the proposed schemes can save overhead messages for address resolution and reduces the forwarding table size in PE nodes. Moreover, it inherits the plug-and-play setup and self-configuration nature of Ethernet.

In next chapter, an EMiM encapsulation scheme is proposed to reducing the forward-

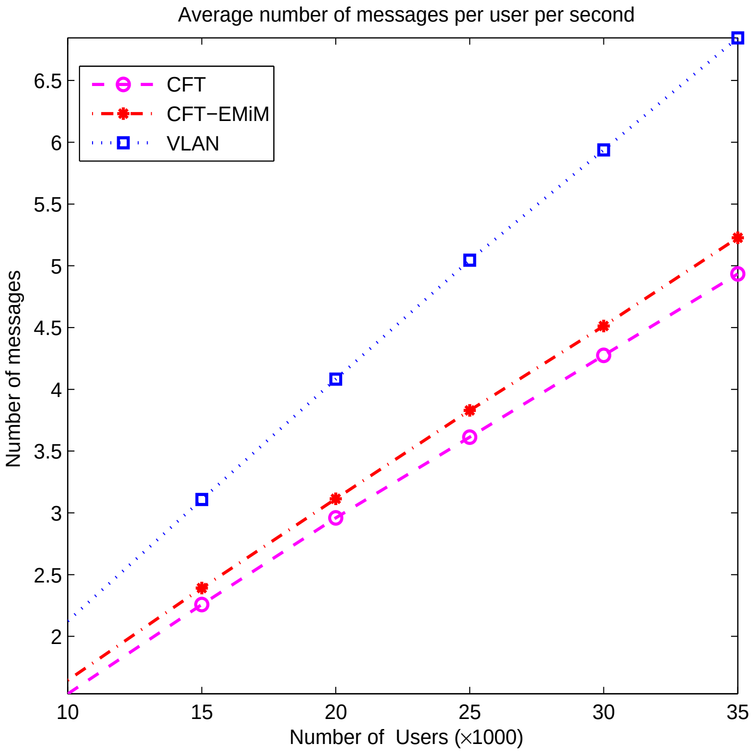


Figure 4.28: Average User message per second

ing table size in PE nodes.

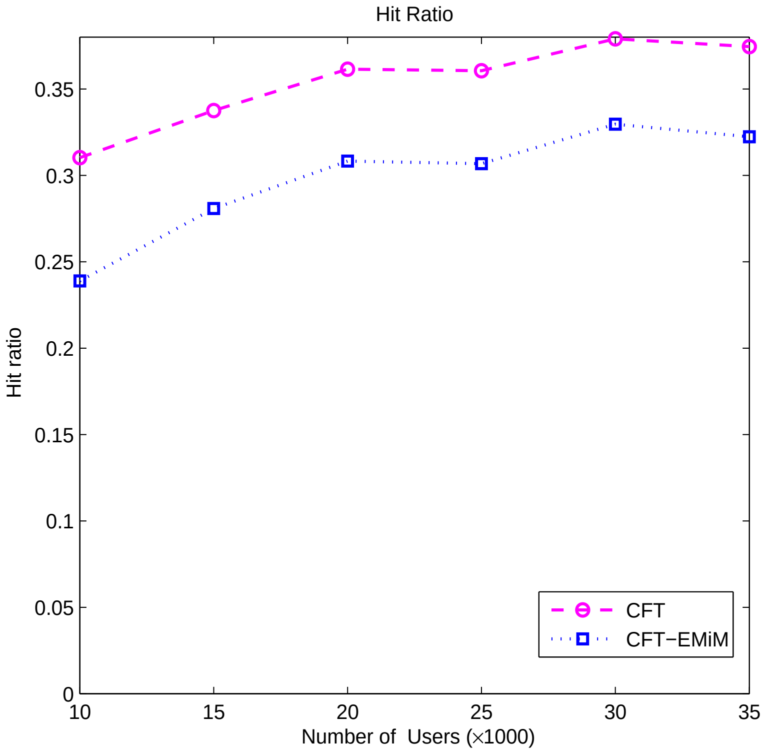


Figure 4.29: Hit Ratio

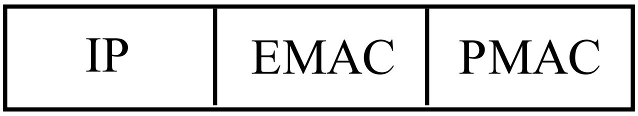
Chapter 5

End user enabled MiM encapsulation

In this chapter, we describe the proposed End user enabled MiM (EMiM) encapsulation scheme in details. The rest of this chapter is organized as follows: Section 5.1 describes the overview of this chapter. The proposed EMiM encapsulation scheme is presented in Section 5.2. The performance evaluation of the proposed scheme is given in Section 5.3. Finally, Section 5.4 concludes this chapter.

5.1 Overview

A general Metro Ethernet includes a provider network and multiple LAN segments. A provider network is composes of multiple PE nodes and CNs (switch or bridge). A LAN segment composes of a CE node and multiple end nodes. In the Metro Ethernet, a PE node and a CN may need to maintain a potential large number of forwarding entries for frame forwarding. MiM encapsulation can be used for frame forwarding in the provider network so that the CNs only need to maintain PE nodes' MAC addresses, thus significantly reducing their learning table size. However, each PE node needs to keep an end user's MAC address to its PE node's MAC address mapping table for MiM encapsulation. Hence the PE nodes may still have learning table explosion problem. To solve this problem, we propose an EMiM encapsulation scheme. In the proposed scheme, the MiM



IP	EMAC	PMAC

Figure 5.1: ARP entry in the proposed scheme

encapsulation is done by an end user instead of a PE node. Hence the PE nodes only need do MAC addresses swapping, thus removing the destination node's MAC address to its PE node's MAC address mapping table.

In the proposed scheme, the PE node's MAC address is inserted to a node's ARP entry, i.e., the ARP entry of an end user composing of the node's IP address, End User's MAC (EMAC) address and its PE node's MAC (PMAC) address, as shown in Figure 5.1. In MiM encapsulation, two pairs of addresses are encapsulated in the frame. Due to an ARP entry includes both EMAC and PMAC addresses, a node has ability to do MiM encapsulation, i.e., adding the Destination EMAC (DEMAC), Destination PMAC (DPMAC), Source EMAC (SEMAC), and Source PMAC (SPMAC) in a frame. In the following, we discuss on how an intermediate node to learn the forwarding entry and how a PE node to encapsulate the MAC addresses for frame forwarding.

5.2 EMiM encapsulation scheme

In the proposed scheme, the MiM encapsulation is done by the end user instead of the PE node, and hence the PE nodes do not need to maintain the forwarding entries mapping end user's MAC address to PE node MAC address and avoid the forwarding table overflow.

In the EMiM encapsulation scheme, the PE node's MAC address is inserted into an end user's ARP entry, i.e., the ARP entry of an end user including the end user's IP address, the EMAC address and the end user's PMAC address. A frame is composed of two pairs of addresses: DMAC1-SMAC1-DMAC2-SMAC2. Here DMAC1 and SMAC1

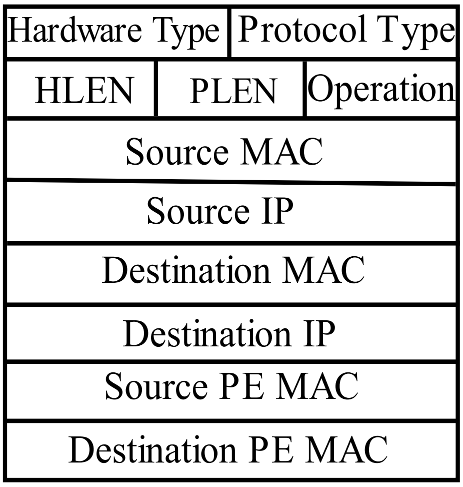


Figure 5.2: ARP Frame Format

are used as the destination and the source MAC addresses for frame forwarding; while DMAC2 and SMAC2 are used for possible address swapping. Because an ARP entry is composed of both EMAC and PMAC addresses, an end user has ability to do MiM encapsulation, i.e., adding DEMAC, DPMAC, SEMAC, and SPMAC in a frame. The ARP request/reposne frame format, shown in Figure 5.2, is modified by attaching the destination and source PE node's MAC addresses.

5.2.1 ARP request frame format

When a user needs to resolve an unknown MAC address, an ARP request frame is broadcasted. The VLAN technology can be used to limit the broadcast messages. When an end user needs to solve a MAC address of a IP address, it can broadcast an ARP request into its VLAN. We do not discuss how to implement the VLAN technology here, because our scheme changes nothing on it. In the broadcast frame, the two pairs of MAC addresses are inserted. The addresses are the broadcast address (denoted as FF), SEMAC, FF, and SPMAC, as shown in Figure 5.3(a). Here SEMAC is the end user's own MAC address and SPMAC is its PE node 's MAC address. If the end user does not know its PE node's MAC address, just inserts the null address (all 0s). When the sender's PE node receives the frame, it sends the frame to CNs by swapping the SEMAC and SPMAC (if SPMAC is null, the PE node just inserts its own MAC as the SPMAC), the frame format is shown in Figure 5.3(b). The PE node also forwards the frame to the other LAN segments behind it without any address change. Hence when the ARP request frame travels in the LAN segments behind the same PE node, only the first pair of MAC addresses are used for frame forwarding. But when the frame travels in the provider network or other LAN segments behind other PE nodes, the user's MAC address has been hidden, while its PE node's MAC address is used as the source MAC address. The other PE nodes receive the broadcast frame just insert their own MAC as the DPMAC, as shown in Figure 5.3(c), and then forward the frame to the LAN segment behind them.

In the above description, we assume the end user having its own PE node's MAC address. In case that the user has no such address, it simply sets the SPMAC as null. The sender's PE node can add its MAC address in the SPMAC field when it forwards the frame to other LAN segments. The end user can learn its own PE node's MAC address when it receives an ARP reply frame. Besides this learning method, the ARP frame format shown in Figure 5.3 indicates that an end user can also learn its own PE node's MAC address by processing an ARP request frame. In the ARP request frame, if the third MAC address is not a broadcast address, the address is its own PE node's MAC address. Otherwise, the forth MAC address is its own PE node's MAC address. This is the case that the ARP requester is behind the same PE node.

5.2.2 ARP response and data frame format

When a user receives an ARP request frame asking for its MAC address, it sends a response frame back to the requester. If the requester is behind the same PE node, the first

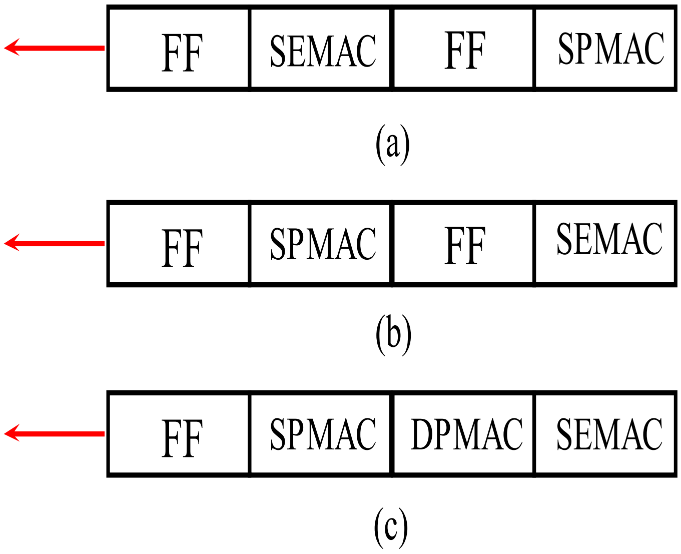


Figure 5.3: ARP broadcast frame format sent by: (a) the sender to the LAN segment; (b) the sender's PE node to the core network; and (c) the receiver's PE node to its LAN segments.

pair of addresses are set to DEMAC and SEMAC, and the second pair of addresses are set to SPMAC and SPMAC. The PE node just forwards the frame without any address swapping. If the requester is in a LAN segment behind another PE node, the frame formats are shown in Figure 5.4. Figure 5.4(a) shows the frame format sent by the sender to the LAN segment. The destination address is set as the receiver's PE node's address; Figure 5.4(b) gives the frame format sent by the sender's PE node to the CNs, and the source address is set as the sender's PE node's address; and Figure 5.4(c) presents the frame format sent by the receiver's PE node to its LAN segment, the address of the sender's PE node is set to be the source address. Note that here SEMAC is the response's MAC address and DEMAC is the ARP requester's MAC address. The data frame has the same format as the ARP response frame.

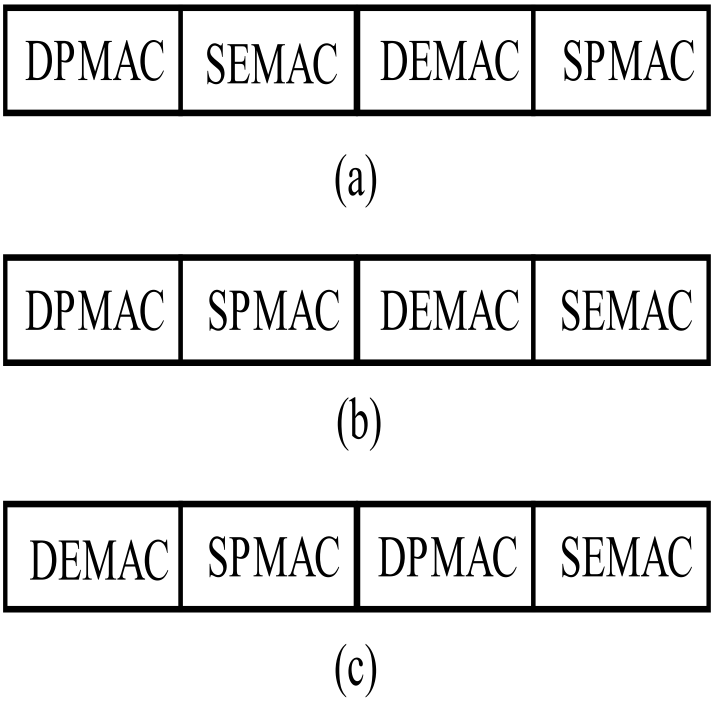


Figure 5.4: ARP response and data frame MiM format sent by: (a) the sender to the LAN segment; (b) the sender's PE node to the CNs; (c) the receiver's PE node to its LAN segment.

5.2.3 Illustration on frame forwarding and forwarding entry learning

Ethernet is a plug and play network. Our EMiM sustains this feature and only modifies the ARP entry format, while the PE nodes may need to swap the addresses. Now we use an example to illustrate how the intermediate nodes to forward the frame and to learn the forwarding entry.

We use the Metro Ethernet given in Figure 4.1 as the example network. Assume the MAC addresses for users 1, 2 , 3 and 4 are $e1$, $e2$, $e3$ and $e4$; and for PE1 and PE2 are $p1$ and $p2$, respectively. The IP addresses of users 1, 2, 3 and 4 are IP1, IP2, IP3 and IP4, respectively. We discuss how user 1 to solve the MAC addresses of user 2, 3, and 4, and how user 1 to encapsulate data frame sending to user 2, 3 and 4.

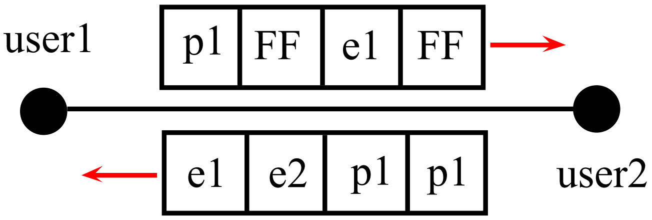


Figure 5.5: Addresses in the frames between user 1 and user 2 - ARP process

Communication between user 1 and user 2

Assume user 1 needs to send data to user 2. If the ARP entry of user 2 is valid in the ARP table, the data frame can be sent directly. Otherwise, user 1 needs to resolve the MAC address of user 2. Now user 1 broadcasts an ARP request to the whole VLAN, and the two pairs of MAC addresses are FF-$e1$-FF-$p1$. The ARP entry of user 1 including IP1, $e1$, and $p1$ is attached to the ARP request frame.

User 2 receives the frame coming from CE1. CE1 learns the forwarding entry of $e1$. After user 2 receives the ARP request frame, it checks the frame and knows that user 1 is behind the same PE node. So it sends an ARP response frame to user 1 with format $e1$-$e2$-$p1$-$p1$. The ARP response frame contains the ARP entry of user 2, i.e., IP2, $e2$, and $p1$. When CE1 receives the response frame, it can forward the frame to user 1 due to it learned the forwarding entry of $e1$, while CE1 learns the forwarding entry of $e2$. After user 1 receives the ARP response frame, it is ready to send data frames. The data frame format is $e2$-$e1$-$p1$-$p1$. The data frame format from user 2 to user 1 is the same as that of the ARP response frame, i.e., $e1$-$e2$-$p1$-$p1$. The addresses in the frames are given in Figure 5.5 and Figure 5.6.

For communication between user 1 and user 3 which are behind the same PE node but different CEs, all the frame format for ARP request, ARP response and data are the

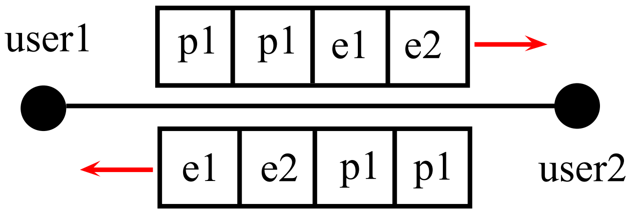


Figure 5.6: Addresses in the frames between user 1 and user 2 - data process

same as these in the communication between user 1 and user 2. Hence when the source and destination users are behind the same PE node, the frame forwarding is similar to that in the case without MiM encapsulation scheme, but a pair of PE node's MAC addresses are hidden (not used for frame forwarding) in the frame.

Communication between user 1 and user 4

Now let us look at the frame forwarding in the communication between user 1 and user 4 which are behind different PEs.

The ARP request frame sent by user 1 is the same as that in the previous case, i.e. FF-$e1$-FF-$p1$. When PE1 gets the frame, it forwards the frame to the CNs by swapping the addresses $e1$ and $p1$, i.e. FF-$p1$-FF-$e1$, so the CNs can learn the forwarding entry of $p1$. When the frame reaches PE2, it just inserts its MAC address to the frame, and forwards the frame to the LAN segments behind it. The frame format is FF-$p1$-$p2$-$e1$. CE3 learns the forwarding entry of $p1$ (from PE2). After user 4 receives the ARP request, it knows that user 1 is behind another PE node, so a response frame with format $p1$-$e4$-$e1$-$p2$ is sent back. CE3 can forward the frame to PE2 because it has learnt the forwarding entry of $p1$. When PE2 gets the frame, it swaps its address with user 4's address, the format now is $p1$-$p2$-$e1$-$e4$. The CNs forward the frame based on $p1$ and learn the forwarding entry of $p2$. When PE1 gets the frame, it swaps its address with user 1's. The format is

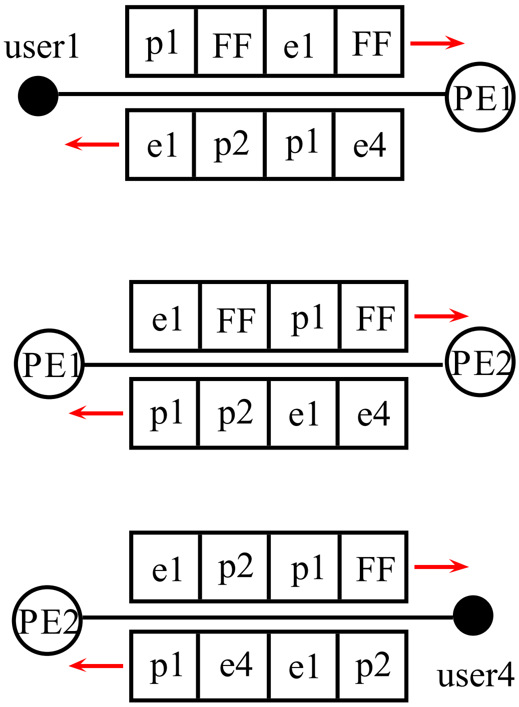


Figure 5.7: Addresses in frames between user 1 and user 4 - ARP process

$e1$-$p2$-$p1$-$e4$. CE1 learns the forwarding entry of $p2$. These formats are shown in Figure 5.7.

When user 1 sends a data frame to user 4, the format is $p2$-$e1$-$e4$-$p1$. PE1 changes the frame format to $p2$-$p1$-$e4$-$e1$. When the frame arrives at PE2, it swaps the destination addresses, the format is $e4$-$p1$-$p2$-$e1$. The data frame from user 4 to user 1 has the same format as that of the ARP response frame. These formats are illustrated in Figure 5.8.

From the above descriptions, we know that the proposed scheme only needs to insert the PE node's MAC address in the ARP entry for encapsulation in the sender. The PE nodes only need to do address swapping. Hence the proposed scheme is highly scalable.

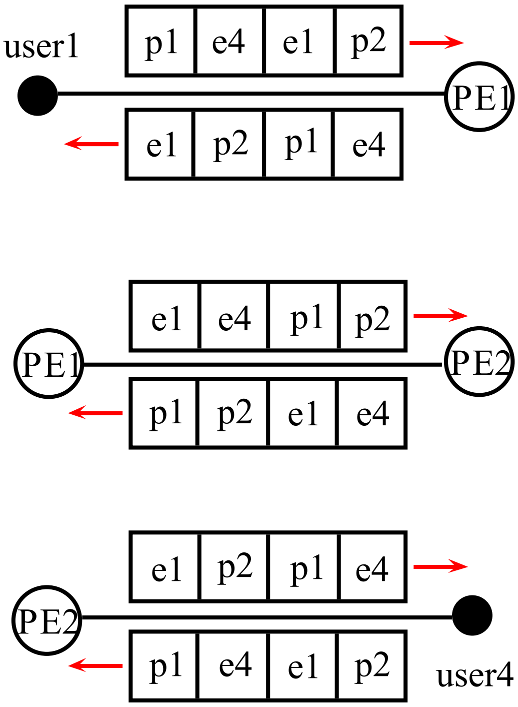


Figure 5.8: Addresses in frames between user 1 and user 4 - data process

5.2.4 Maximum forwarding table size

From the above descriptions, we know that the CNs only need to maintain the forwarding entries for PE nodes. A PE node and any intermediate nodes in a LAN segment may need to maintain the forwarding entries of all the PE nodes and all the end users behind the PE node. Assuming the number of users in LAN segment behind a PE node is N and the number of PEs is M, the maximum forwarding table size in the proposed scheme is $N-1+M$, which is much smaller than the number of entries $N*M$ in traditional VLAN based scheme. For example, for a Metro Ethernet with 100 PE nodes and 5000 user behind each PE node, our scheme needs only 5099 forwarding entries, while the traditional scheme needs 50 thousand entries. Hence the proposed scheme is very scalable. Moreover, the proposed scheme does not introduce any redundancy communication message while sustains Ethernet's plug and play feature.

5.2.5 Solution for a Metro Ethernet Composed of Both End Users with and without EMiM

The EMiM scheme is easy to be implemented in a Metro Ethernet with all end users having the knowledge of EMiM. But it may be difficult to update the software for all end users and hence it is necessary to support both end users with and without EMiM knowledge.

For a Traditional end User (TU) without EMiM knowledge, it cannot do MiM encapsulation in the frame and also has no its PE node's MAC address in its ARP entry. It also cannot decapsulate the MiM frame. Hence when an end user needs to solve an unknown MAC address, it broadcasts an ARP request frame without MiM head in the frame no matter the end user is TU or an EMiM User (EU). The only difference is that the EU attaches a pair of PE node's MAC addresses in the frame, as shown in Figure 5.2. In the following, we discuss the communications for two end users in four cases (a) TU to TU; (b) TU to EU; (3) EU to TU; and (4) EU to EU.

(a) TU to TU: in this case, the ARP request and response as well as the data frame have the same frame formats as those in the traditional MiM frame format (i.e., no MiM encapsulated from the end user/PE node to PE node/end user). The PE nodes can use MiM scheme for frame forwarding in the provider network, and hence the PE nodes need to maintain the mapping entry of such end users' MAC address to their PE node's MAC address.

(b) TU to EU: the TU sends an ARP request frame with traditional frame format and the EU knows that the requester is a TU (no PE node's MAC address in the request frame) and hence sends the response back using traditional ARP response frame format. The following data frame has the same format as the traditional one.

(c) EU to TU: the EU sends an ARP request frame by attaching a pair of PE node's

MAC addresses in the frame. The TU only learns the first pair of MAC address (i.e., the end user's MAC address), and sends an ARP response frame back to the requester. The EU knows the responser is a TU and hence uses the traditional frame format for data frame.

(d) EU to EU: the EU sends an ARP request frame by attaching a pair of PE node's MAC addresses in the frame. The responser knows that the requester is an EU (due to the attached pair of PE node's MAC adress), and hence it sends a response frame with EMiM frame format. The requester can know the responser is an EU and the data frame is encapsulated in EMiM format.

In the above cases, a PE node needs to record the entry of mapping the end user's MAC address to its PE node's MAC address if an ARP frame is not encapsulated in EMiM format. So the proposed EMiM scheme is easily to be implemented in Metro Ethernet composed of both end users with and without EMiM scheme.

5.3 Performance Evaluation

The performance of the proposed scheme is evaluated by simulation in this section. We study the average and maximum forwarding table sizes in the PE nodes by comparing with the existing VLAN technology. In the simulations, we set a Metro Ethernet with 40 PEs, 500 CEs and 40 CNs. Each PE node directly connects to at least 2 and at most 5 PE nodes. At least 2 and at most 25 CEs can be behind each PE node. There are at least 8 and at most 256 end users behind a CE node. We set 1000 VLANs in the network. Each VLAN has at least 3 and at most 13 CEs. Each user has probability to communicate with other end users in the same VLAN. The data session interval time (the time interval to send data frame) follows the Poisson distribution. The average data session interval time is T (i.e., one communication session per T seconds), and each data session lasts average

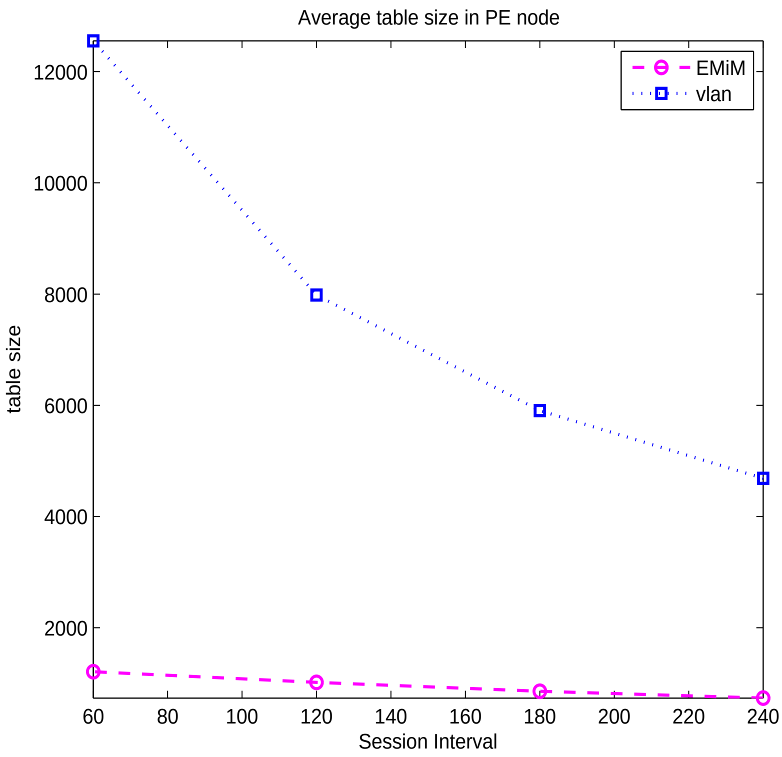


Figure 5.9: Average forwarding table size in PE node

20 seconds, randomly chosen between 1 and 39 seconds. When a user starts a data session, it randomly picks a VLAN it belongs to, and then randomly chooses another user in the same VLAN as the destination user. The ARP entry is timed out every 120 seconds. At the beginning, all the forwarding tables are set to be empty.

Case 1: Impact of data session interval

We set 50 thousand end users in this case. Note that the size of the forwarding table at the PE node is the same no matter if the MiM scheme is used or not for the traditional VLAN technology. Figures 5.9 and 5.10 show the simulation results of the proposed scheme (denoted as EMiM) and the traditional VLAN scheme (denoted as vlan). From the Figures, we know that the average and maximum forwarding table sizes at the PE

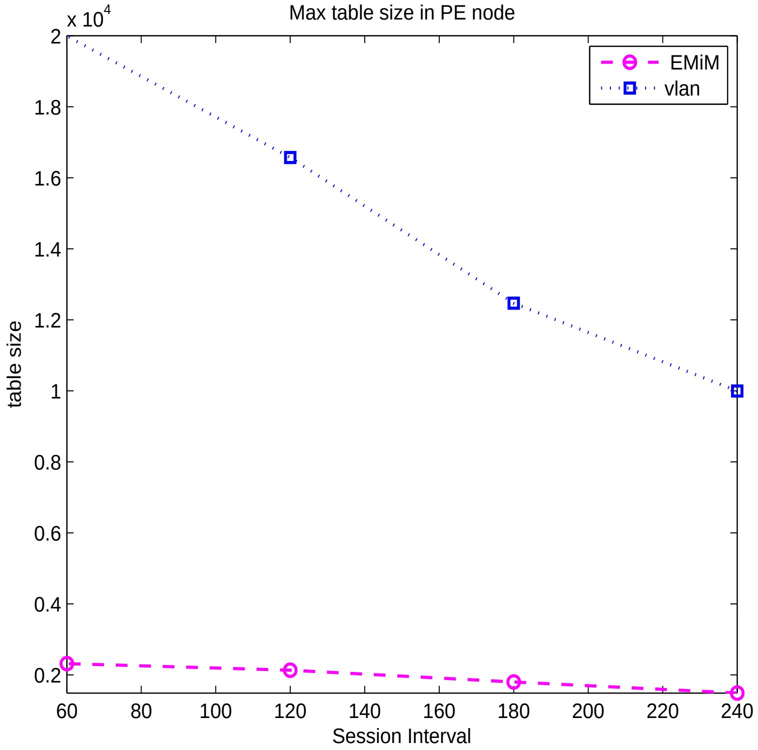


Figure 5.10: Maximum forwarding table size in PE node

nodes reduce as the data access interval time T increases. This is due to the number of concurrent communication sessions decreases as T increases. Remember that a forwarding entry is timed out every 120 seconds, and hence the total number of valid forwarding entries decreases. From the results, we also know that the proposed scheme can reduce around 90% for both average and maximum forwarding table size at T=60 seconds. These results show that the proposed scheme can significantly reduce the forwarding table size in the PE nodes. Moreover, our scheme only modifies the ARP entry without changing any other existing communication protocols.

Case 2: Impact of number of users

We study the performance impact of number of users. We vary the number of users

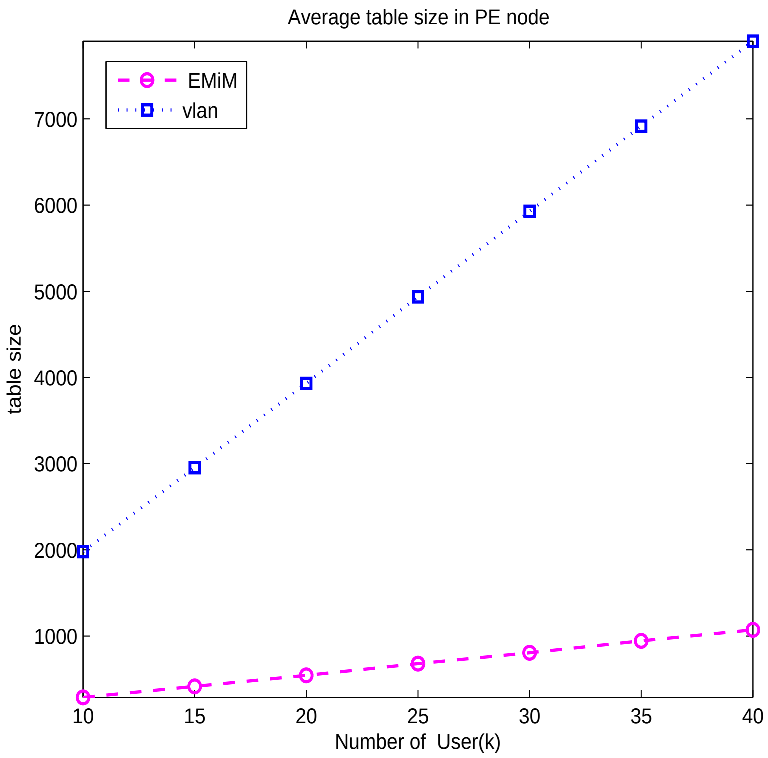


Figure 5.11: Average forwarding table size in PE node

from 10,000 to 40,000. Here the data access interval time is set as T=120s.

The results shown in Figure 5.11 and Figure 5.12 indicate that the average and maximum table sizes grow as the number of users increases in both schemes. However, the increase for the traditional VLAN technology is much faster than that in EMiM. In all the range of the number of users, EMiM can save more than 50% forwarding entries. This shows that the proposed scheme has much better scalability than the traditional VLAN technology.

Case 3: Impact of number of end users for Metro Ethernet composed of both TU and EU

Case 3 is set to investigate the performance of a Metro Ethernet composed of both EUs

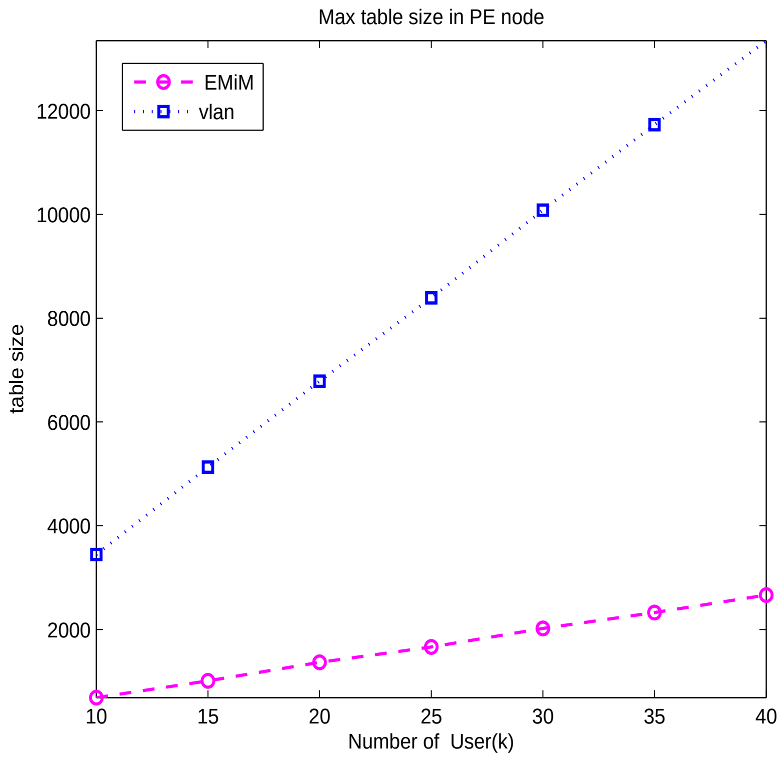


Figure 5.12: Max forwarding table size in PE node

and TUs (called mixed network for short). In this case, the simulated network contains 50,000 end users, the same as that in case 1. The end users contain 50% EUs and 50% TUs. Both EUs and TUs are randomly distributed. Communications between EUs and TUs are conducted at random. Demonstrated in Figures 5.13 and 5.14, the table size in the mixed network is over 4 times higher than that in a network with all EUs, but it is about 30% lower than the network with all TUs. The results indicate that EMiM is useful for mixed networks.

Case 4: Impact of EU ratio

In this case, the simulated network contains 10,000 end users, including both EUs and TUs. Both EUs and TUs are randomly distributed. The percentage of EU in the whole network varies from 0% to 100%. As shown in Figure 5.15 and Figure 5.16, the

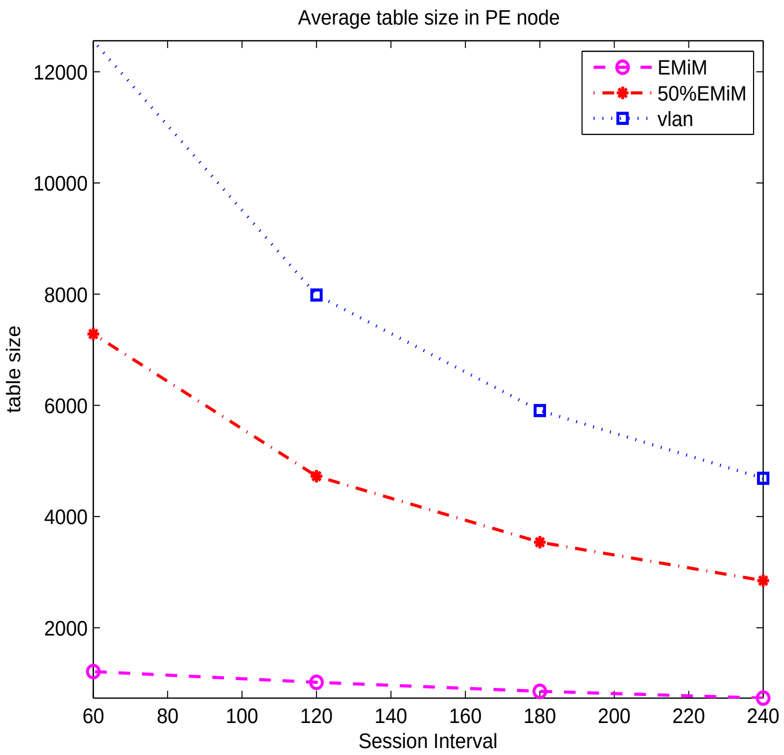


Figure 5.13: Average forwarding table size ratio in PE node

average and maximum table sizes in PE nodes decrease dramatically as the number of EUs increases. The average table size with 0% EU is 6.6 times as big as that with 100% EUs, and the Max table size changes from around 3700 to 700, a decrease of 81%. This is because for the EMiM scheme, PE node only needs to store the MAC addresses of end users that are under itself instead of these in the whole network.

These results show that the proposed EMiM encapsulation scheme is highly scalable and easily implemented in Metro Ethernet.

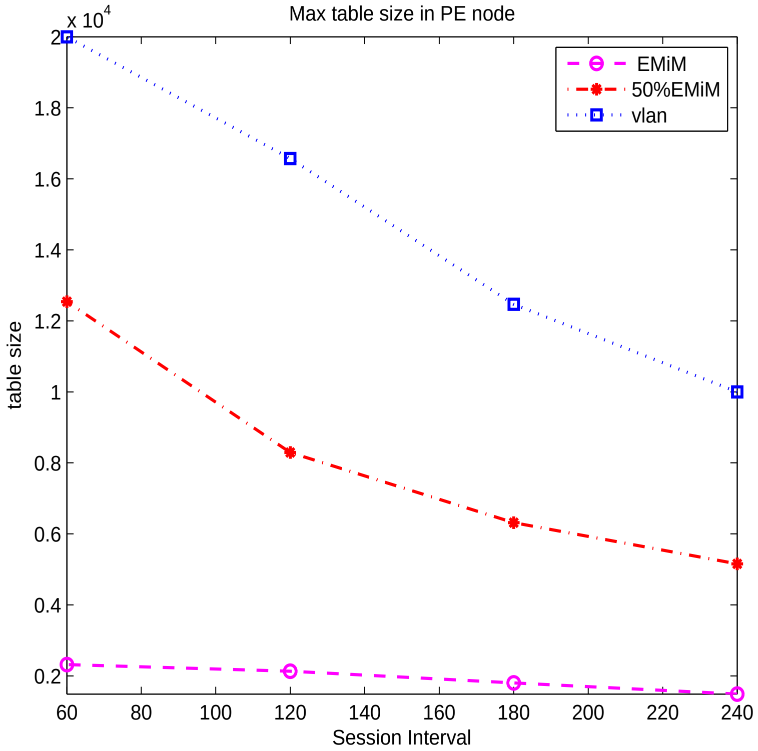


Figure 5.14: Max forwarding table size ratio in PE node

5.4 Summary

In this chapter, we have proposed an EMiM encapsulation scheme for Metro Ethernet. Unlike the traditional MiM encapsulation method, our scheme not only reduce the forwarding table size in the CNs, but also in the PE nodes. The proposed scheme can work with any existing Metro Ethernet protocols by only modified the ARP entry format. Hence the proposed scheme is highly scalable.

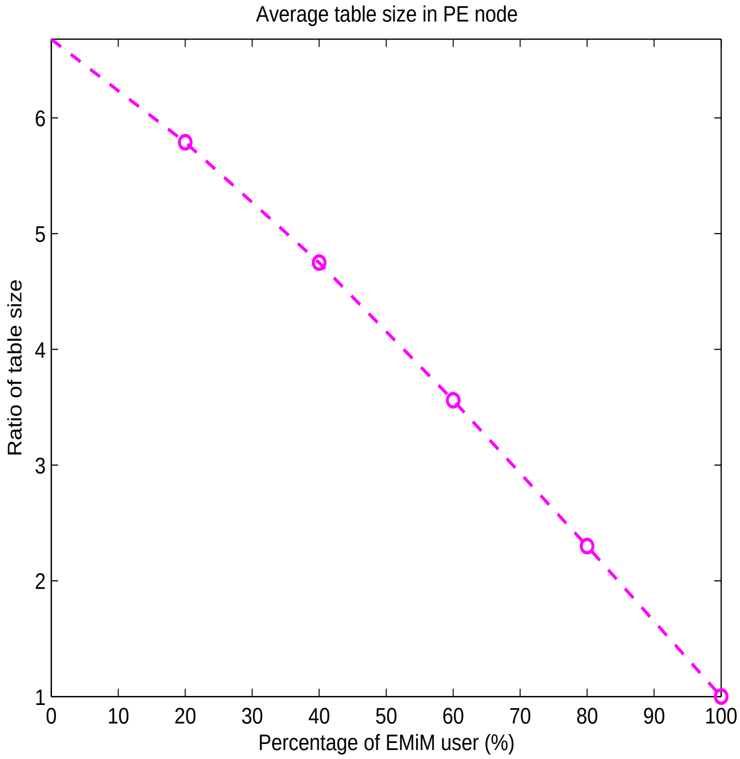


Figure 5.15: Average forwarding table size ratio in PE node

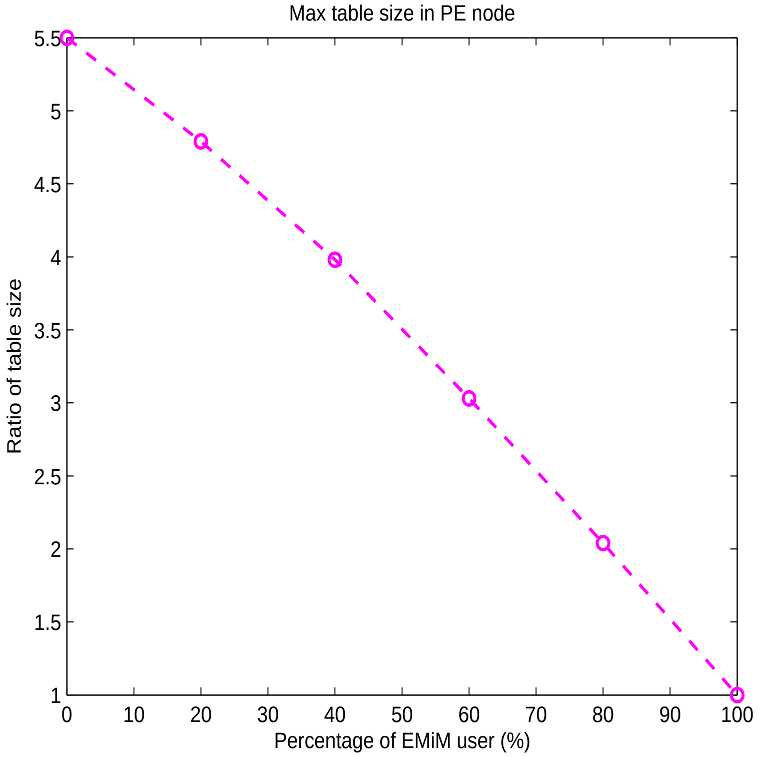


Figure 5.16: Max forwarding table size ratio in PE node

Chapter 6

Distributed Registration address resolution Protocols

This chapter describe the proposed distributed Registration address resolution Protocol (RP). The rest of this chapter is organized as the follows: Section 6.1 introduces the overview of this chapter. Section 6.2 gives the detailed architecture of Provider edge based distributed Registration address resolution Protocol (PRP). Section 6.3 presents the details of the proposed Customer edge based distributed Registration address resolution Protocol (CRP) architecture. The performance comparisons are shown in Section 6.4. Finally, Section 6.5 concludes this chapter.

6.1 Overview

In this charpter, we propose an efficient and scalable Metro Ethernet architecture. The proposed architecture includes two RPs. In the RP, multiple ARP registers are allocated to support address resolution. Each IP address has a home register which stores its ARP entry. When an end user moves to another location but keeps its IP address, its current PE or CE node is considered to be its foreign register. A foreign register temporally caches the ARP entry for an immigrated user and is in charge of the ARP entry updating in the home register. The IP address is used as an index to locate the corresponding home register through unicast, thus eliminating the broadcast to solve an unknown address. We combine the EMiM scheme into the architecture and also discuss how the proposed schemes coexist with end users using traditional ARP technology.

6.2 Provider edge based distributed Registration address resolution Protocol (PRP)

As shown in Figure 4.1, a general Metro Ethernet including a provider network and multiple LAN segments. A provider network is composed of multiple PE nodes and CNs (switch or bridge). A LAN segment consists of a CE node and multiple end users. Our goal is to develop a scalable and efficient Ethernet architecture which allows any pair of end users to communicate with each other.

6.2.1 Home and foreign ARP register

In the proposed PRP, each PE node is considered to be an ARP register, and hence multiple ARP registers are allocated to provide address resolution. The IP address is used as an index to locate its corresponding ARP register so that no broadcast service is needed in the provider's core network.

A PE node is in charge of all of the ARP entries belonging to some IP prefixes. The PE node is called as the home ARP register for any ARP entry whose IP address belongs to any of these IP prefixes. The PE node is called as the foreign ARP register for an IP address which does not belong to any of these IP prefixes. A PE node also has a table storing the mapping of IP prefix to home register for all the IP prefixes in the whole Metro Ethernet. When a user joins to the LAN segment under its home register, it needs to register its ARP entry immediately so that it can communicate with other users.

When a user moves to a LAN under another PE node (as the foreign PE node), it needs to register its ARP entry to its foreign register immediately. The foreign register adds the entry and immediately forwards the entry to its home register. Each user needs to periodically send its ARP entry to home/foreign register. The home register updates the ARP table whenever the ARP entry is changed. When a user moves from one foreign register to another, a new ARP entry will be sent to the home register which in turn asks the previous foreign register to remove the ARP entry. Hence each ARP entry needs to be stored at most twice in the network, and the maximum ARP table size in a PE node is at most the twice of the number of users in the LANs directly connected to it.

In summary, the entries in an ARP table in a PE node include:

(1) all the IP addresses using the PE node as their home ARP register;

(2) one ARP entry for each IP prefix in the network;

(3) the IP addresses of end users in all LANs directly connected to the PE node, but the PE node is not their home ARP register.

6.2.2 PRP combining with EMiM encapsulation Scheme

The RP scheme eliminates the broadcast messages for address resolution and the forwarding table sizes by associating PMAC addresses in the ARP entry. The entry format of the ARP table is <IP address, EMAC, PMAC, recordtime,age>, EMAC is the MAC address of an end user; PMAC is the MAC address of the PE node (called local PE node) that directly connects to the LAN the user belongs to; recordtime is the time when this ARP is created or refreshed; and age indicates if the ARP entry is valid or not.

6.2.3 Address Resolution Process and Data Frame Format

The address resolution and frame forwarding process are as follows. When a new user joins in a LAN, it first needs to register its ARP entry in the local PE node through DHCP or directly broadcast its <IP, EMAC, NULL> to the LAN, and the CE node of the LAN forwards the frame to the local PE node. When the local PE node gets an ARP registration message, it inserts its MAC address as the PMAC, and adds the entry to the ARP table. If the PE node is not the home ARP register of that IP address, it needs to send the ARP entry to its home ARP register to add/update the corresponding ARP entry. After the registration, each user needs to periodically send message to its local PE node (foreign or home ARP register) as well as its home register to refresh the ARP entry. In case of a user moves from one site to another site belonging to a different PE node, it needs to immediately register to its new local PE node as its foreign ARP register for communication. The foreign register forwards its ARP entry to its home register.

When a user needs to resolve a MAC address for a destination, it directly sends a message to its local PE node. The ARP entry is returned if the entry is found there. Otherwise, the PE node uses the IP address as the index and sends an ARP request message to its home ARP register to get the corresponding ARP entry. After the home register gets the ARP request message, it sends the ARP reply back to the local PE node. The local PE node then forwards the ARP entry to the end user. The end user record the destination's IP address, EMAC and PMAC in its ARP entry. Then it generates the communication frame to start a data session.

The EMiM encapsulation scheme is used to transmit data session frames after an end user knowns the MAC address of the destination. The data frame format is as following: If the destination is behind the same PE node, the first pair of addresses are set to Destination EMAC (DEMAC), and Source EMAC(SEMAC), and the second pair of addresses

DPMAC	SEMAC	DEMAC	SPMAC

(a)

DPMAC	SPMAC	DEMAC	SEMAC

(b)

DEMAC	SPMAC	DPMAC	SEMAC

(c)

Figure 6.1: ARP response and data frame EMiM format sent by: (a) the sender to the LAN segment; (b) the sender's PE node to the CN node; (c) the receiver's PE node to its LAN segment.

are set to Source PMAC (SPMAC) and SPMAC. In case both the destination and the source are located behind the same CE node, this frame can be forwarded by CE node or other switches noticed the destination MAC addresses. This frame transits through the standard Ethernet forwarding path of the switches. It is not necessary for the PE node to handle this data frame. If the destination and the source nodes located in different CE node behind the same PE node, when the frame reaches the PE node, the PE node just forwards the frame without any address swapping. If the destination is in a LAN segment behind another PE node, the frame formats are shown in Figure 6.1. Figure 6.1(a) shows the frame format sent by the sender to the LAN segment. The destination address is set as the receiver's PE node's address; Figure 6.1(b) gives the frame format sent by the sender's PE node to the CNs, and the source address is set as the sender's PE node's address; and Figure 6.1(c) presents the frame format sent by the receiver's PE node to its LAN segment, the address of the sender's PE node is set to be the source address.

6.2.4 Illustration of ARP entry learning and frame forwarding

We use the Metro Ethernet given in Figures 4.1 as the example network. Assume the MAC addresses for user 1, 2, 3 and 4 are *e*1, *e*2, *e*3 and *e*4; and for PE 1, 2, 3 and 4 are *p*1, *p*2, *p*3 and *p*4, respectively. The subnet IP gate way PE 1, 2, 3, 4 handle are 61.1.1.1, 62.1.1.1, 63.1.1.1, 64.1.1.1 respectively with subnet mask 255.255.0.0. And the IP addresses of user 1, 2, 3 and 4 are 61.1.1.2, 62.1.2.2, 63.1.3.3, 64.1.4.4. IP addresses here are only used as the index to find the MAC addresses and the local register of the end user whereas for propagation purpose. End user can keep their IP addresses regardless the location restriction in legacy scheme. We discuss both the conditions that PE2 is and is not the Home Register of user 4.

(1) ARP registration and resolving process

PE2 is the Home Register of user 4

Assume user 4 joins in the LAN segment for the first time, it broadcasts a request to ask for the IP address of its own if no IP address is given to it previously. After the DHCP server receives the frame, it assigns IP address 64.1.4.4 to host 4. In the meantime, a registration frame is sent to PE2 as the new IP address is under the charge of PE2. The ARP entry < 64.1.4.4, e4, *p*2> is created in the PE2. When user 1 needs to send frames to user 4, it can send the data frame directly if its forwarding table entry for user 4 is valid. Otherwise, user 1 generates a unicast ARP request message and directs it to PE1. PE1 uses IP address 64.1.4.4 to search its ARP entry. The IP address belongs to the subnet 64.1.0.0 which is behind PE2. That is PE2 is the Home Register of user 4. So PE1 sends the ARP request message to PE2. After PE2 gets the ARP request message, it looks for its ARP table and replies the corresponding ARP entry back to PE1. PE1 then forwards

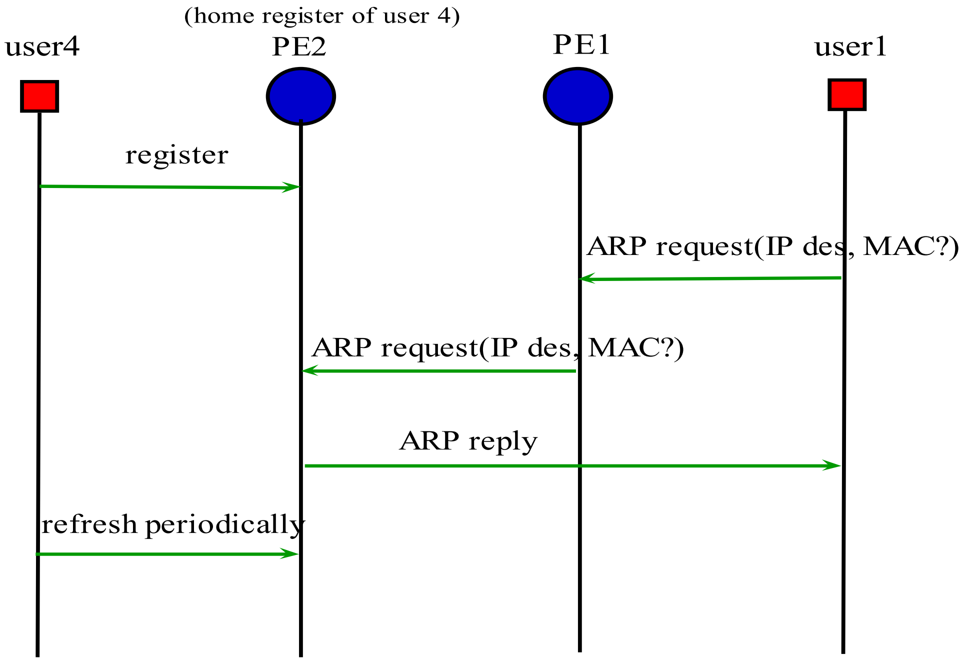

(home register of user 4)
user4
PE2
PE1
user1
register
ARP request(IP des, MAC?)
ARP request(IP des, MAC?)
ARP reply
refresh periodically

Figure 6.2: ARP registration and resolving process case 1: end user locates under its home register

this response to user 1. User 1 adds < IP4, *e*4, *p*2> in its ARP table and generates the data frame directly destined to user 4. Now communication between user 1 and user 4 can start. The process is illustrated in Figure 6.2. It is similar to the case that two end users reside under the same PE node.

PE2 is not the Home Register of user 4

Now let us consider the case that user 4 is an immigrant from PE3. Assume PE3 is the home register of user 4 and PE2 is the foreign register. And the IP address of user 4 is derived from 63.1.0.0 whose designated register is PE3. Based on our RP, one roaming host should deliver a registration frame right after it immigrated. So, user 4 should send an ARP registration frame to PE2 immediately after the movement to announce its existence. When CE3 receives this frame, it stops the broadcast and sends this frame to PE2 directly. After checking the ARP table, PE2 knows that it is not the home register of user 4, in other words, PE2 knows it is the foreign register of user 4. Consequently, PE2 adds the entry of user 4 to its ARP table. In this way, the following ARP request intended for user 4 can find this proximal entry instead of the subnet entry destined for PE3. Meanwhile, PE2 also transmits this registration message to PE3, the home register of user 4. This can make sure the entry in the home PE node always has the most recently ARP entry. Then, PE3 updates the ARP entry of user 4 in its ARP table. The new location is recorded. When user 1 needs to resolve the MAC address of user 4, it generates a unicast ARP request message to PE1 in case its forwarding entry of user 4 is invalid. PE1 sends this frame to PE3 after checking its ARP table. PE3 responses to PE1 indicating that the current register of user 4 is PE2. Then PE1 sends the reply back to user 1. User 1 adds the entry in its forwarding table and transports data frames to user 4 directly. Note in this case, when user 4 roams out of PE2, PE2 deletes the entry of user 4 after noticing of the movement(i.e. the roaming is notified by PE3) as it is not the home register of user 4. Figure 6.3 shows the process in this case.

In the other case, if user 4 moves from PE1 to PE2 and its IP address is 61.1.1.4. That is, PE1 is the home register and PE2 is the foreign register of user 4. In this situation, the registration process is the same as the previous one. Specially, when user 1 wants to communicate with user 4, after it sends request to PE1, it knows the the local PE node of user 4 is PE2 directly.

(2) Data communication process

Communication between user 1 and user 2

Assume that user 1 needs to send data to user 2. If the ARP entry of user 2 is valid in the ARP table, the data frame can be sent directly. Otherwise, user 1 needs to solve the MAC address of user 2. Instead flooding to the whole VLAN, user 1 sends a unicast

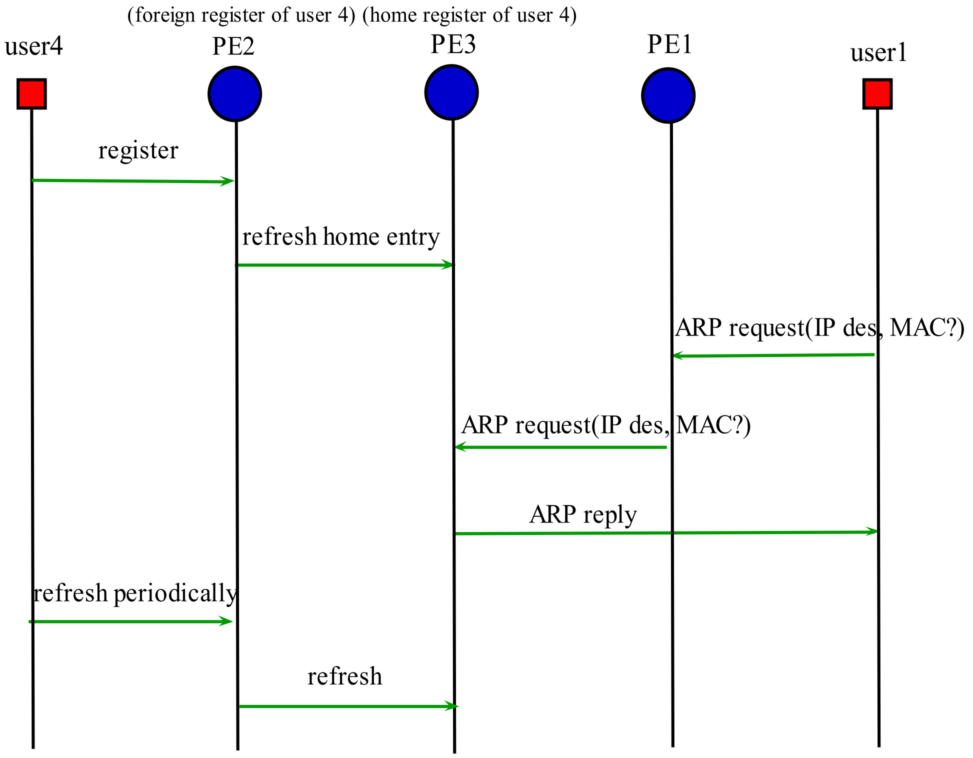

(foreign register of user 4) (home register of user 4)
user4
PE2
PE3
PE1
user1
register
refresh home entry
ARP request(IP des, MAC?)
ARP request(IP des, MAC?)
ARP reply
refresh periodically
refresh

Figure 6.3: ARP registration and resolving process case 1: end user locates under a foreign register

ARP request to PE1. As user 2 attaches to the LAN that belongs to PE1, an ARP reply can be sent back directly from PE1. The ARP response frame contains the ARP entry of user 2, i.e., IP2-$e2$-$p1$. After user 1 receives the ARP response frame, it is ready to send data frames. The data frame format is $e2$-$e1$-$p1$-$p1$. The data frame format from user 2 to user 1 is the same as that of the ARP response frame, i.e., $e1$-$e2$-$p1$-$p1$. The addresses in the frames are given in Figure 6.4.

For communication between user 1 and user 3 that are behind the same PE node but different CEs, all the frame format for ARP request, ARP response and data are the same as those in the communication between user 1 and user 2. Hence when the source and destination users are behind the same PE node, the frame forwarding is similar to the case without MiM encapsulation scheme, but a pair of MAC addresses are hidden (not used

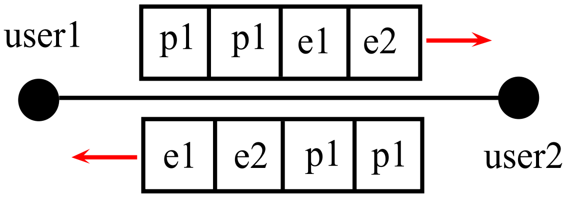


Figure 6.4: Addresses in the frames between user 1 and user 2 - data process

for frame forwarding) in the frame.

Communication between user 1 and user 4

Now let us look at the frame format in the communication between user 1 and user 4 which are behind different PEs.

In case that user 1 has the ARP entry of user 4 in its table, it can start a session based on the information. If it does not know, it should ask from the register. When PE1 gets the ARP request frame from user 1, it forwards the frame to the home register of user 4 based on the IP prefix. There are two possibilities. User 4 can be roamed from other LAN belongs to a different PE node. Or, it can be a original local resident of PE 2. In both situations, the frame should be redirected to the home register of user 4. As the home PE node always keep an entry of user 4, it can reply the information back. When the home register receives the frame, after checking its ARP table, it sends back a response frame to tell the MAC address of $ip4$ is $e4$, the format is IP4-$e4$-$p2$. PE1 transit this frame to user 1.

Now user 1 can send data frames to user 4, the format is $p2$-$e1$-$e4$-$p1$. PE1 changes the frame format to $p2$-$p1$-$e4$-$e1$. When the frame arrives at PE2, it swaps the source addresses, the format is $e4$-$p1$-$p2$-$e1$. The data frame from user 4 to user 1 are illustrated in Figure 6.5.

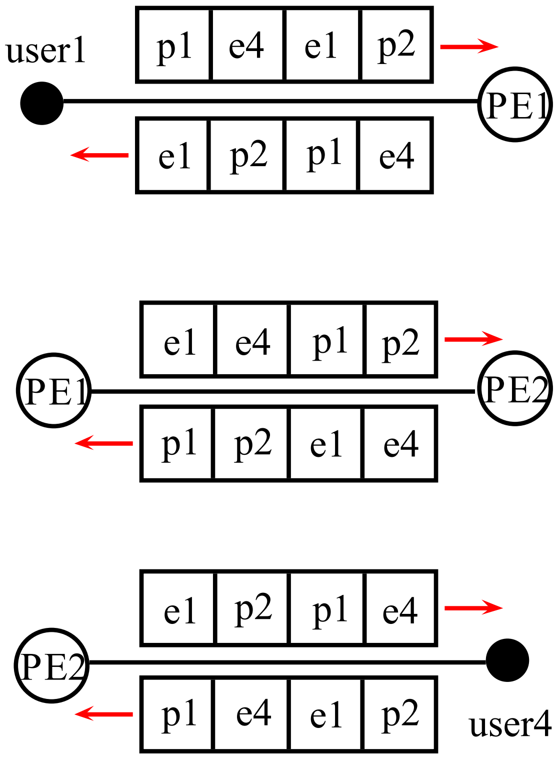


Figure 6.5: Addresses in frames between user 1 and user 4 - data process

The proposed PRP replaces the ARP broadcasting or multicasting with unicasting and allows any user to communicate with any other user in the network. The scheme does not need any configurations and hence maintains Ethernet's plug-and-play setup and self-configuration capability.

6.2.5 Performance analysis

The proposed architecture can greatly reduce the forwarding table size and save the communication message for address resolution. We estimate the performance of the proposed architecture in metrics of maximum forwarding table size and number of communication messages per unit time.

The following assumptions are made in a Metro Ethernet:

- there are N_p PE nodes
- there are N_l LANs, each VLAN has n_l end users
- there are total N_u end users
- a user has a data session every t_d time
- forwarding entry timeout time is t_t
- ARP refresh time in PRP is t_r

We use S_{vlan}^{max} and S_{new}^{max} to represent the maximum forwarding table size for the VLAN scheme and the proposed scheme, M_{vlan} and M_{new} to represent the number of messages for address resolution per unit time for the VLAN scheme and the proposed scheme.

The maximum forwarding table size in VLAN is N_u, i.e., $S_{vlan}^{max} = N_u$. This is the worst case that all the end users are in communication within a time period t_t. The maximum table size of the proposed scheme is the number of end users behind a PE node plus the number of PE nodes, it can be estimated as $S_{new}^{max} = N_u/N_p + N_p$. The average number messages for address resolution in VLAN is estimated as $M_{vlan} = N_u n_l/t_d$, and for the proposed one is $M_{new} = N_u/t_d + N_u/t_r$.

As an example, suppose $N_p = 20$, $n_l = 200$, $N_u = 600,000$, $t_r = t_t = 120$ seconds, $t_d = 50$ seconds, then $S_{vlan}^{max} = 600,000$, and $S_{new}^{max} = 30,020$. The maximum table size is reduced by more than 90%. The number of messages for address resolution per unit time in the network is $M_{vlan} = 2,400,000$, and $M_{new} = 17,000$. 99% messages can be saved.

6.2.6 Coexisting with legacy users

Backward compatibility should be mentioned. Assume a user using legacy ARP attaches to the network for the first time, it broadcasts a DHCP message. The DHCP server assigns an IP address to this user. If this message passes through the local PE node, it will record the forwarding entry. This entry can be used for future communication. Not like RP, this entry can not be refreshed periodically. It can only be refreshed when this user is processing a data session. So this user entry may be timed out and no longer usable. When the user moves to another PE node, again broadcast is used to send the DHCP message and a new IP address is assigned to this user. In the following, three scenarios are considered for the frame forwarding between different types of users.

The sender is using legacy ARP

If the destination MAC address is stored in the forwarding table of this user, a communication can start directly. Otherwise, this user broadcasts an ARP request to the whole VLAN. When the local PE node receives this broadcast message, it stops broadcasting and uses the IP address as index to find the corresponding MAC address. If the destination is located in the LAN area connected to this PE node, it will send a reply back directly. Otherwise, it will send an ARP reply message including the MAC and IP address mapping of the destination to the sender. The destination's MAC address can be resolved either in this local PE node or in its home PE node. Then, the sender can use the information contained in this ARP reply to start the session.

The destination is using legacy ARP

As the source user has knowledge of PRP, it sends a unicast message to its home PE node to query the MAC address of the destination. The home PE node lookup its ARP table based on the destination IP address. There are two conditions. The destination may locate in the same PE node. If the detailed table entry is found, the home PE node can reply the mapping directly to the source. Otherwise, the home PE node informs the source user the mapping did not find. Then the sender floods an ARP request to the network. As mentioned above, home PE node can stop transmitting the broadcast messages.

The destination user replies to the sender, and a session can begin. The destination may belong to another PE node. In this situation, the home PE node sends the request to the corresponding PE node. This PE node can reply back immediately if the entry is available. Otherwise, it broadcasts a message to resolve the destination MAC address. Once this PE node receives the reply from the destination user, it will send the message back to the home PE node of the sender. After the source user receives the ARP reply frame from the home PE node, the communication can start.

Both the sender and the destination are using legacy ARP

This case is the combination of the two cases mentioned above. The broadcast messages are delivered in the LAN area. Whereas, there are no broadcast happens in the core network as PE node can always stop broadcast and use unicast to communicate with another PE node.

6.3 Customer edge based distributed Registration address resolution Protocol (CRP)

In the above section, we described PRP. In order to further lower the table size in PE nodes, we also consider to use Customer edge based distributed Registration address resolution Protocol (CRP). Following are the details of this design.

The basic principles of CRP and PRP are similar. The mainly difference is that the CE node instead of PE node is considered as the ARP register in CRP. In CRP, the entries in the ARP table in a PE node include:

(1) One ARP entry for each IP prefix in the network. The mapping of IP prefix and the PE node that originally owns this IP prefix is contained in this entry;

(2) One ARP entry for each IP prefix under this PE node. The mapping of IP prefix to

the CE node that resides under the PE node and originally owns the IP prefix is contained in this entry.

The entries in the ARP table in a CE node include:

(1) All the IP addresses of the end users that use the CE node as their home ARP register;

(2) The IP addresses of end users in all LANs directly connected to the CE node, but the CE node is not their home ARP register.

The registration and address resolution process is similar to that in PRP. The end user registers to the local CE node (local register) when it firstly resides in this network and periodically sends message to this local CE node (foreign or home ARP register) to refresh the ARP entry. If the local register is not their home register, a message is also sent to its home register to keep the ARP entry in its home register up to date. A user can send a request to its local CE node to ask for an unknown MAC address. A reply will come back if the entry can be found in this CE node. Otherwise, a further request message will be delivered to the home register of the destination based on the IP prefix. A reply will send back to the source requester from this home register.

The EMiM encapsulation scheme is still used to transmit data session frames. In order to cooperate with CRP, the entry format of the ARP table is modified as <IP address, EMAC, CMAC,PMAC, *recordtime*,*age*>, EMAC is the MAC address of an end user; CMAC is the MAC address of the CE node (called local CE node) that directly connects to the LAN the user belongs to ; PMAC is the MAC address of the PE node that directly connects to the CE node. The data frame format is different from that in PRP. There are three situations following: (1), the source and destination are behind the same CE node. The first pair of addresses is set to DEMAC and SEMAC, the second pair of addresses are set to Source CEMAC (SCMAC) and SCMAC, and the third pair of addresses are

DCMAC	SEMAC	DEMAC	SCMAC	SPMAC	SPMAC

(a)

DCMAC	SCMAC	DEMAC	SEMAC	SPMAC	SPMAC

(b)

DEMAC	SCMAC	DCMAC	SEMAC	SPMAC	SPMAC

(c)

Figure 6.6: case 2 ARP response and data frame EMiM format sent by: (a) the sender to the LAN segment; (b) the sender's CE node to the PE node and PE node to the receiver's CE node; (c) the receiver's CE node to its LAN segment.

set to Source PMAC (SPMAC) and SPMAC; (2), the destination and the source users locate in different CE nodes behind the same PE node. The frame format delivered by the end user to the LAN segment is presented in Figure 6.6(a). Figure 6.6(b) shows the format sent by the sender's CE node to the PE node, and the frame format sent by the PE node to the receiver's CE node has the same format. The destination address is set as the receiver's CE node's address; Figure 6.6(c) gives the frame format sent by the receiver's CE node to its LAN segment, and the source address is set as the sender's CE node's address; (3), the destination is in a LAN segment behind another PE node. The frame formats are shown in Figure 6.7. Figure 6.7(a) shows the frame format sent by the sender to the LAN segment. The destination address is set as the receiver's PE node's address; Figure 6.7(b) gives the frame format sent by the sender's CE node to the PE node, and the source address is set as the sender's CE node's address; and Figure 6.7(c) gives frame format sent by the sender's PE node to the CN node, and the source address is set as the sender's PE node's address; Figure 6.7(d) presents the frame format sent by the receiver's PE node to the receiver's CE node; The frame format delivered by the receiver's CE node to its LAN segment is shown in Figure 6.7(e).

DPMAC	SEMAC	SCMAC	DCMAC	DEMAC	SPMAC

(a)

DPMAC	SCMAC	SEMAC	DCMAC	DEMAC	SPMAC

(b)

DPMAC	SPMAC	SEMAC	DCMAC	DEMAC	SCMAC

(c)

DCMAC	SPMAC	SEMAC	DPMAC	DEMAC	SCMAC

(d)

DEMAC	SPMAC	SEMAC	DPMAC	DCMAC	SCMAC

(e)

Figure 6.7: case 3 ARP response and data frame EMiM format sent by: (a) the sender to the LAN segment; (b) the sender's CE node to the PE node; (c) the sender's PE node to the CN node; (d) the receiver's PE node to the receiver's CE node; (e) the receiver's CE node to its LAN segment.

6.3.1 Illustration on ARP entry learning and frame forwarding

We use the Metro Ethernet given in Figures 4.1 again as the example network to illustrate the ARP entry learning and frame forwarding process in CRP. Assume the MAC addresses for user 1, 2, 3 and 4 are $e1$, $e2$, $e3$ and $e4$; and for CE 1, 2, 3, 4 and 5 are $c1$, $c2$, $c3$, $c4$ and $c5$, respectively; and for PE 1, 2, 3 and 4 are $p1$, $p2$, $p3$ and $p4$, respectively. IP addresses are used as the index to find the MAC addresses. End user can keep their IP addresses regardless the location restriction in legacy scheme. We discuss both the conditions that user 4 is under its home register and user 4 is under a foreign register.

(1) ARP registration and resolving process

User 4 is under its home register

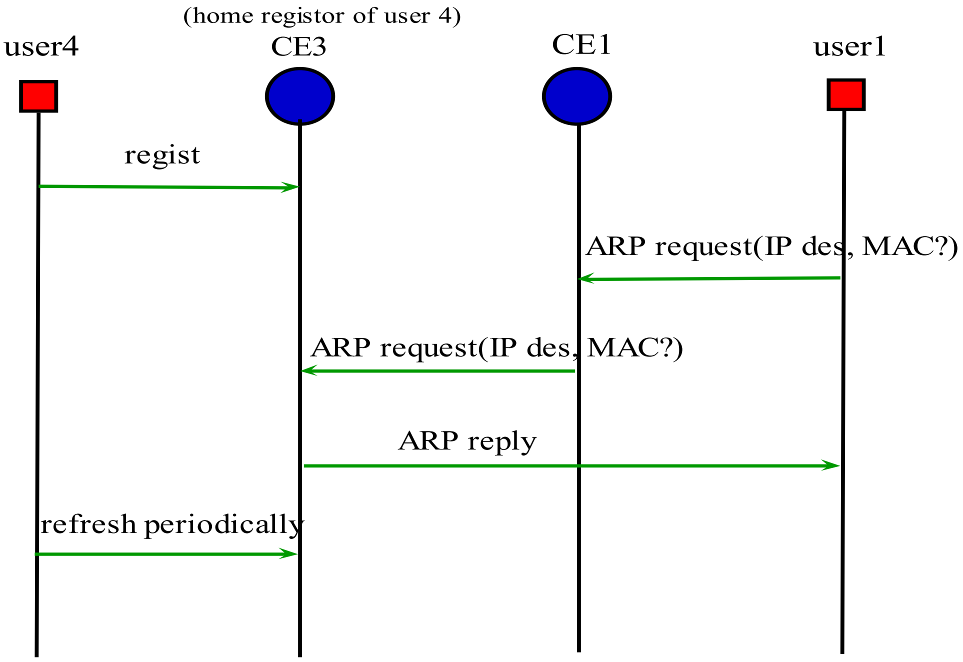

(home registor of user 4)
user4
CE3
CE1
user1
regist
ARP request(IP des, MAC?)
ARP request(IP des, MAC?)
ARP reply
refresh periodically

Figure 6.8: ARP registration and resolving process case 1: end user locates under its home register

Figure 6.8 demonstrates the ARP registration and resolving process in this case. When user 4 connects to the network, a registration message is sent to CE3. CE3 creates the ARP entry containing the IP address and MAC address mapping of user 4 as it's the home register of user 4 based on the IP subnet assigned. If another user (e.g. user 1) needs to know the MAC address of user 4, user 1 can deliver a query frame to its local register CE1. CE1 looks up the IP address in its registration table and find the corresponding register of this IP. As the IP address belongs to the subnet matching to CE3, CE1 directs the request to CE3 and CE3 replies the MAC address and its location (under CE3) back. User 4 refreshes its registration entry periodically.

User 4 is under a foreign register

In this case, after CE3 (the foreign register of user 4) receives the registration frame from user 4, CE3 stores the location (port) and the addresses information in its registration

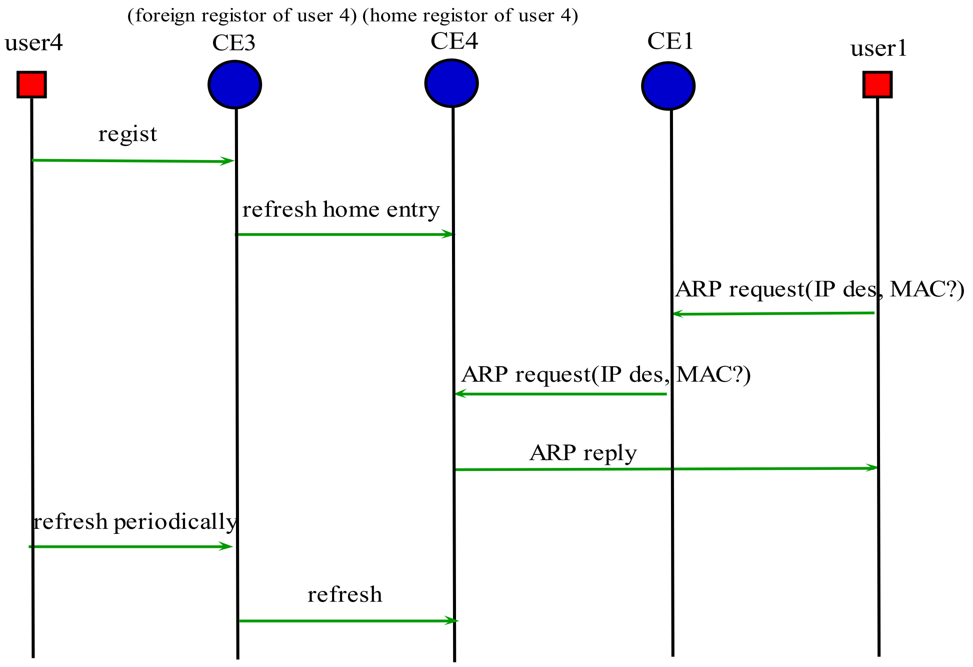

(foreign registor of user 4) (home registor of user 4)
user4
CE3
CE4
CE1
user1
regist
refresh home entry
ARP request(IP des, MAC?)
ARP request(IP des, MAC?)
ARP reply
refresh periodically
refresh

Figure 6.9: ARP registration and resolving process case 2: end user locates under a foreign register

table and notifies CE4 the new location of user 4 at the same time. Therefore, CE4 (the home register of user 4) can keep the entry up to date. When CE1 (the local register of user 1) gets the request from user 1 to resolve the MAC address of user 4, it forwards this request to CE4 after checking the IP subnet registration table. CE4 can reply the location (under CE3) and the MAC address back to user 1. The process is shown in Figure 6.9.

Every time user 4 updates its registration table, both the foreign registration and home registration table will be refreshed. Once CE3 notices user 4 roamed away (e.g. notified by user4's home register), it can delete this entry. But the entries in the home register will not be deleted in this case.

(2) Data communication process

Figure 6.10 shows the communication process between user 1 and user 2 that are under the same CE node. Figure 6.11 illustrated the communication process between user

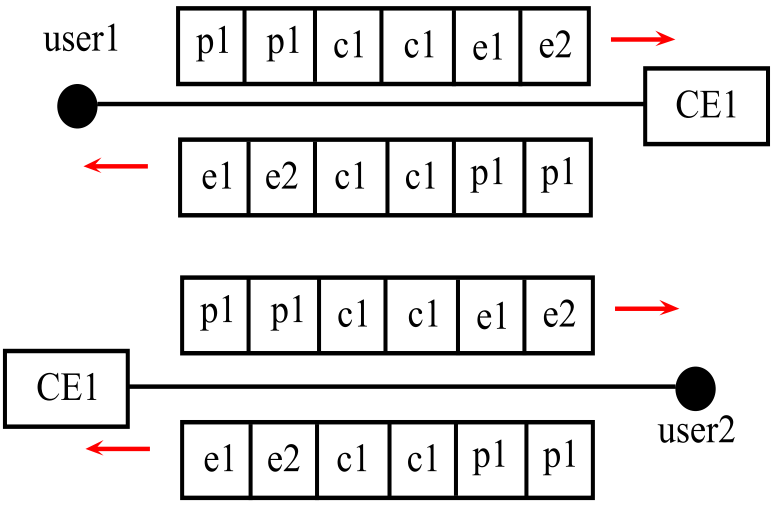


Figure 6.10: Addresses in the frames between user 1 and user 2 - data process

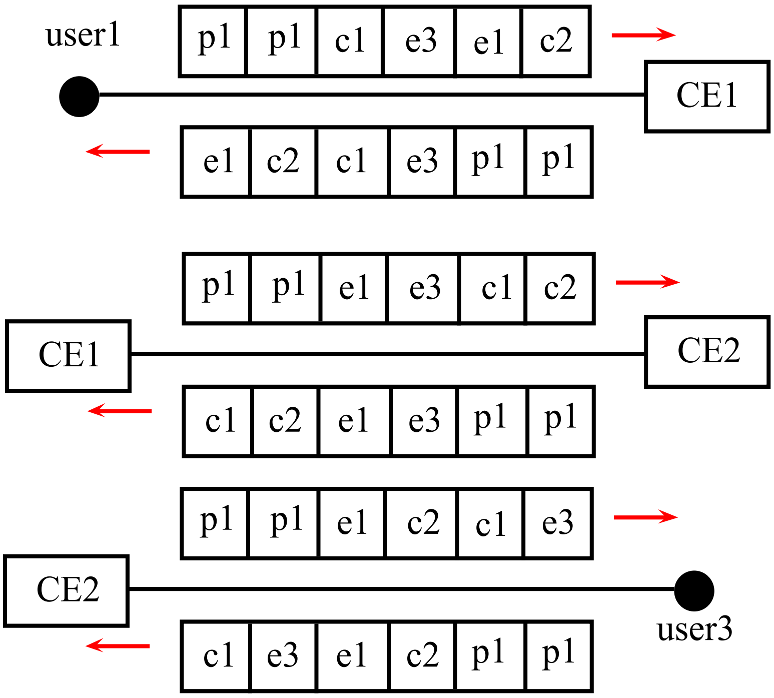


Figure 6.11: Addresses in the frames between user 1 and user 3 - data process

1 and user 3 that are under the same PE node but different CE nodes. The communication process between user 1 and user 4 that are under different PE nodes is demonstrated in Figure 6.12.

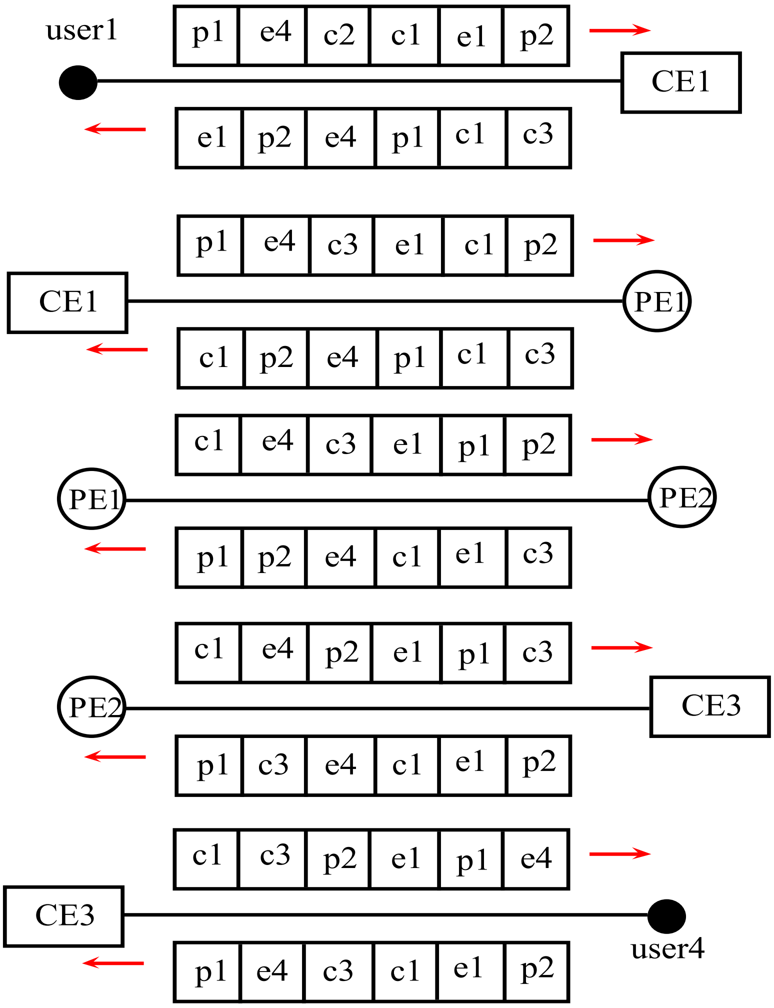


Figure 6.12: Addresses in the frames between user 1 and user 4 - data process

6.4 Evaluation

This section compares the performances of the proposed schemes and the legacy VLAN based schemes. Both the impact of data session interval and number of users are considered. Assume that there are *N* end users per VLAN and *M* PE nodes in the Metro Ethernet, when a user floods to resolve an unknown destination, all the other *N*-1 users in the same VLAN receive the broadcast message, and *M*-1 messages are transited through the provider network (not considering the reply message).The devices transmitting these broadcasting frames have to record the source MAC address. This is the situation happened in the legacy VLAN based scheme. In RP, an unknown MAC address is solved through unicast. Up to 3 messages may need to pass through the provider network depending on the location of the destination.

In our simulation, three cases are studied. The parameters are shown in Table 1. In PRP and CRP, each user can communicate with any other end users without the restriction of VLAN, and every user sends message to its registers to refresh its ARP entry in every 2 minutes. For the legacy VLAN based scheme, PE node messages include the broadcast/multicast frames transmitted by PE nodes to resolve unknown MAC addresses. User messages contain all the broadcast/multicast frames generated by the user to resolve unknown destination MAC addresses. In PRP and CRP, the PE and CE node messages consist of:

(1) all the ARP frames transmitted by PE and CE nodes.

(2) the messages that a foreign register sends to a home register to refresh the ARP entry when a user roams.

(3) the messages that a foreign register sends to a home register when a user sends messages to refresh its ARP entry.

The user messages in both registration based schemes include:

1) all the ARP frames the users generated/handled.

2) the messages users sent to PE nodes or CE nodes for registration or refreshing.

Case 1: Impact of session interval

This case is set to investigate the performances of our proposed schemes and the legacy VLAN based scheme along the increase of session rate. In this case, *T* varies

Table 6.1: Simulation parameters

	Number of PE nodes	Number of CE nodes	Number of VLANs	Number of end users	Data session interval	User roaming interval
Case I	40	150	1000	50k	60 s to 240 s	no roaming
Case II	30	90	800	10k to 40k	120 s	no roaming
Case III	30	90	800	10k	120 s	600 s to 4200 s

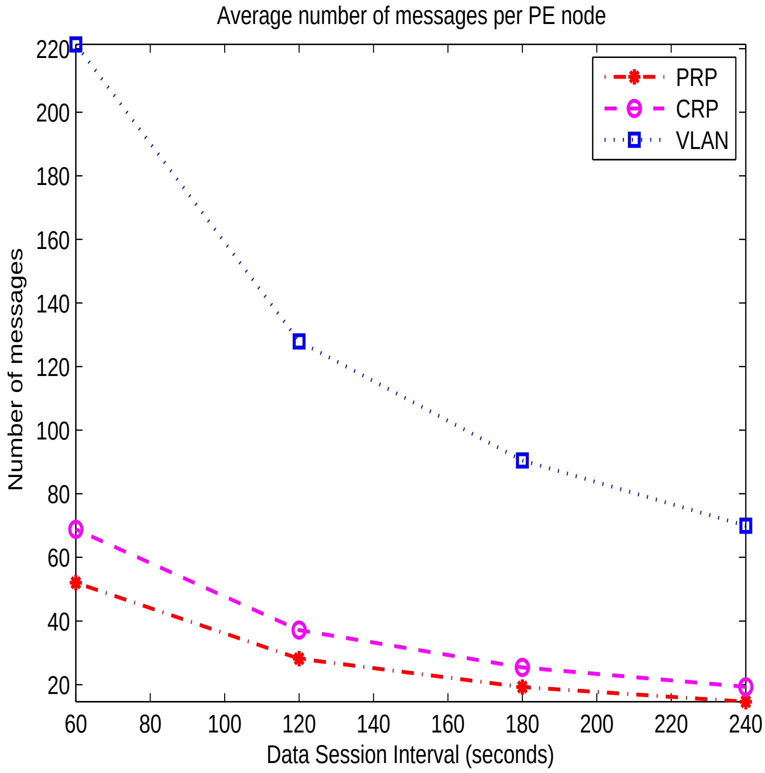


Figure 6.13: Average PE message

from 60s to 240s. Both the traffic load and table size are measured.

Figure 6.13 and Figure 6.14 show the average number of messages a PE node handled per second and the average number of messages a user handled per second. From the Figure, we can see the performance for the proposed schemes are much better than the legacy VLAN based one. Note that session interval T is the reciprocal of data session

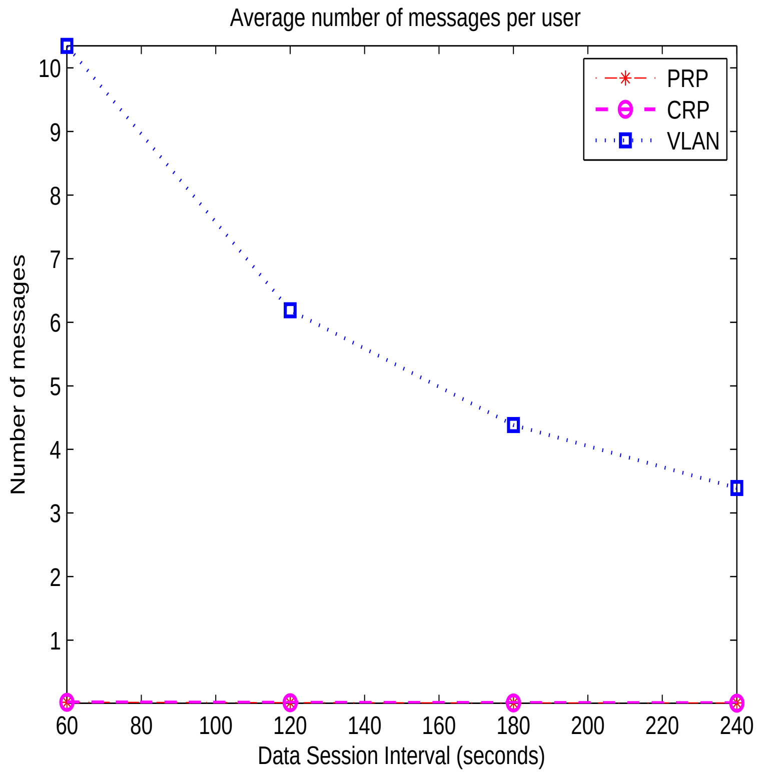


Figure 6.14: Average User message

rate. As data session rate increases, the number of messages grows rapidly in traditional systems. This is due to using broadcast to resolve unknown MAC addresses. Whereas, the new ones are much slower as the unknown MAC addresses can be resolved by sending a unicast request message to its PE node.

Figure 6.15 and Figure 6.16 illustrate the average table size and the max table size of PE node for all the schemes. Figure 6.17 and Figure 6.18 show the average table size and the max table size of CE node for all the schemes. In PRP and CRP, the table of PE/CE node includes the sum of the entries in the registration ARP table and routing tables. Both table sizes increase linearly with increasing load. However, the table size of the traditional vlan based one grows much more rapidly compared with the new schemes. Though the registration entries have to be stored, the table size is much smaller in both

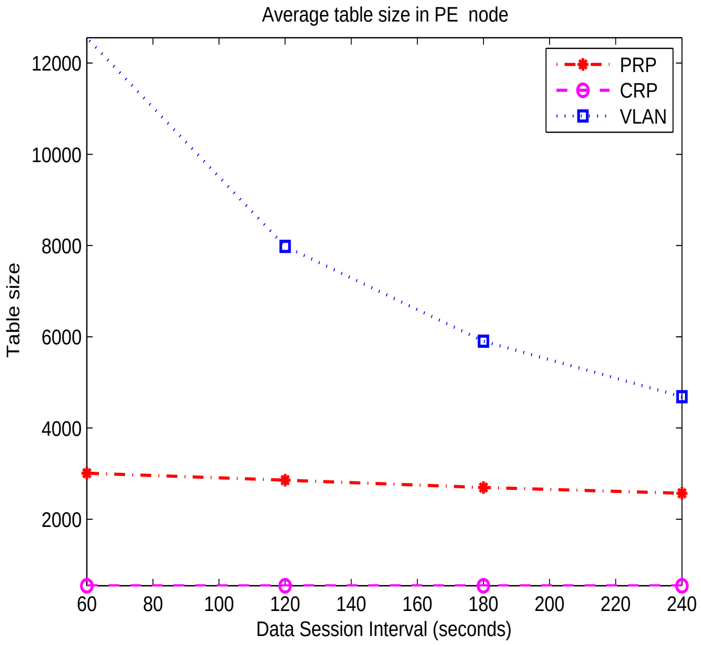


Figure 6.15: Average PE table size

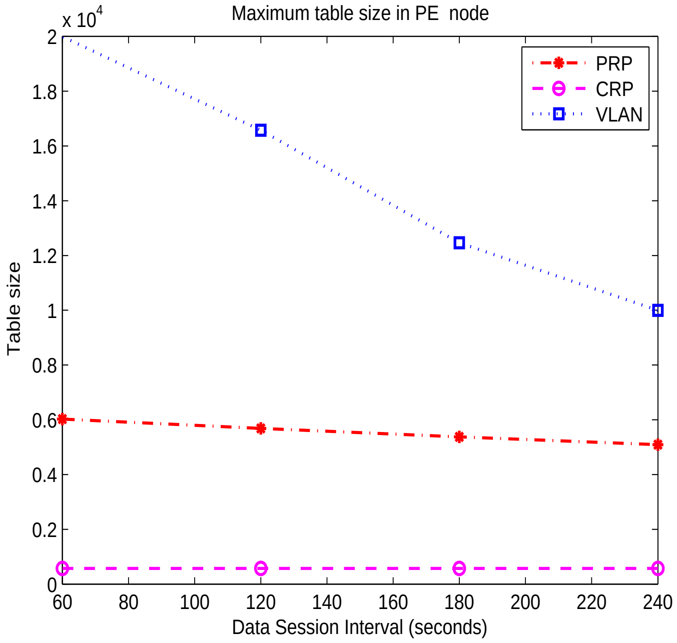


Figure 6.16: Max PE table size

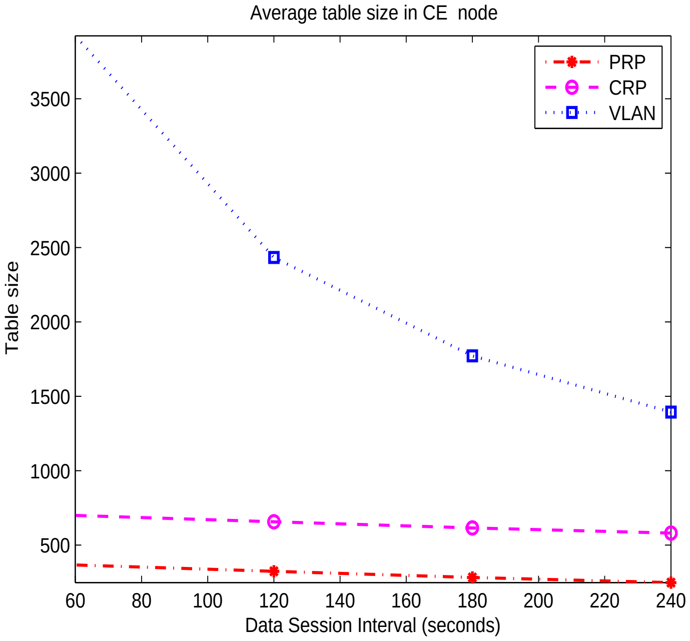


Figure 6.17: Average CE table size

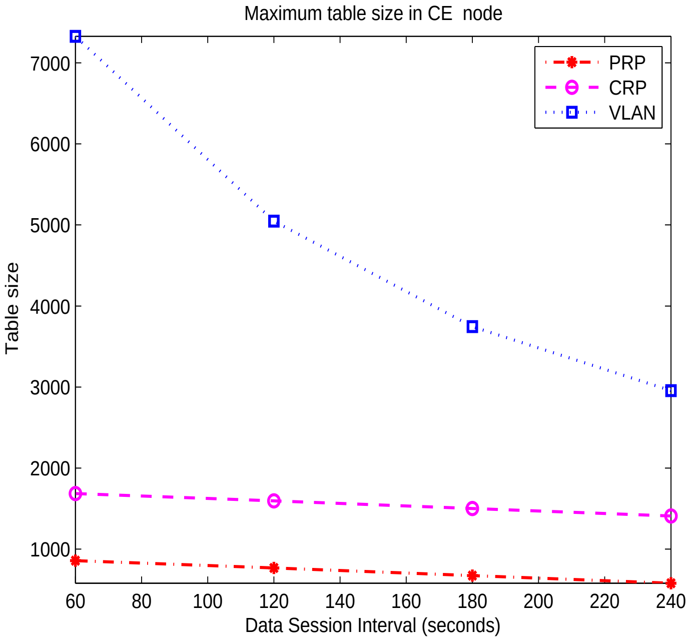


Figure 6.18: Max CE table size

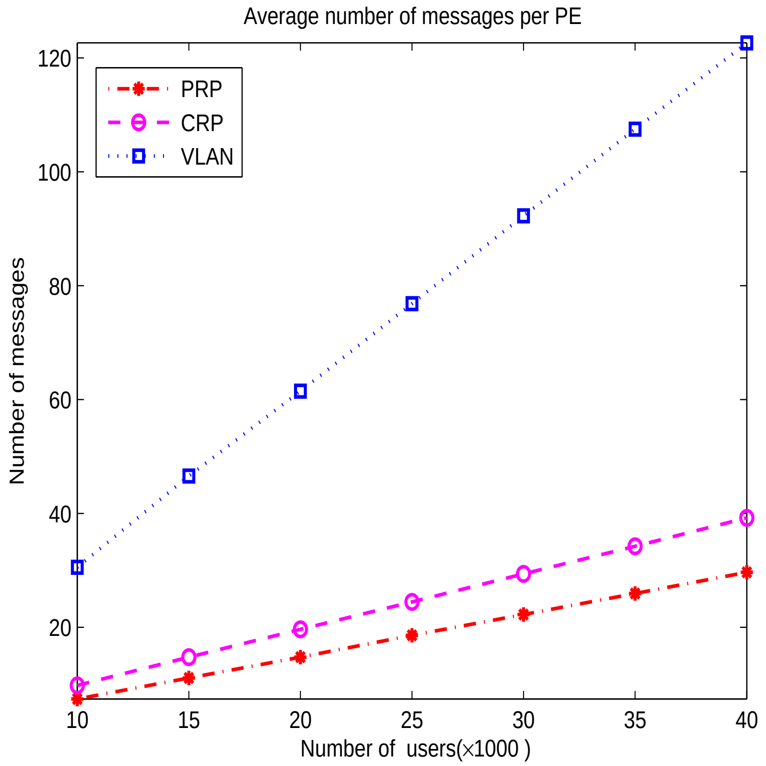


Figure 6.19: Average PE message number

PRP and CRP. The legacy scheme performs even worse along with the increase of data access rate λ. In this situation the routing table of the traditional VLAN based one grows rapidly whereas the table size of the new schemes are not so sensitive to the increasing of the traffic load. This is because the size of the registration ARP table is not affected by the growth of the traffic load.

Case 2: Impact of number of users

We explore the performance impact of the number of end users, which we vary from 10,000 to 40,000 with T=120s.

The results of average message amount a PE node generated in a second are demonstrated in Figure 6.19. At 10k users, PRP and CRP can reduce more than 20 average

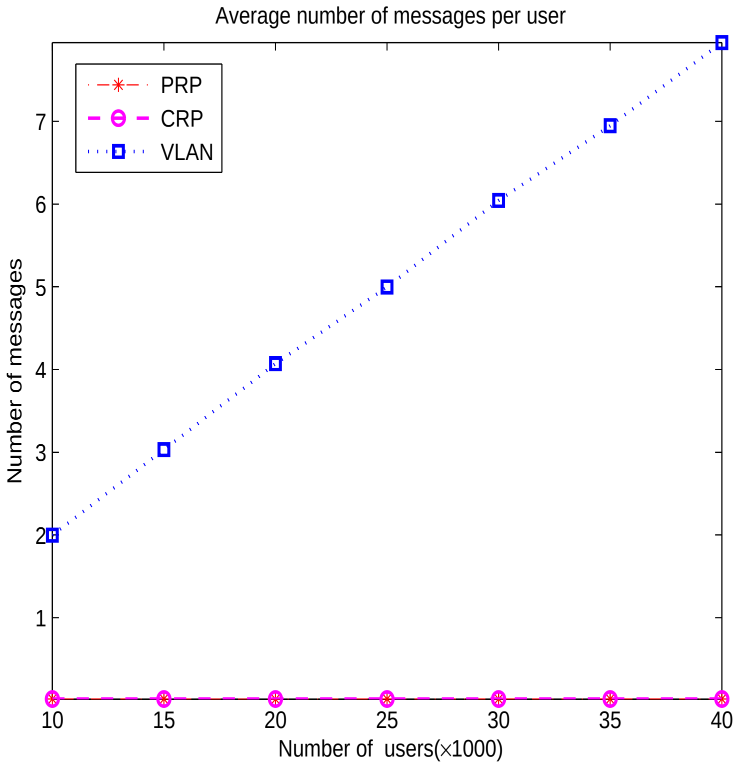


Figure 6.20: Average User message

messages. As the end users increase, the performances get more and more progressive as we expected. At 40k users, about 90 messages can be reduced for average PE node. Compared with CRP, PRP generates less PE messages. This is because in CRP, a PE node needs to send extra request to the CE node to ask for the registered ARP information of the end users.

The average messages a user generated increase along with the increase of end users for the traditional scheme, while in registration based ones (both PRP and CRP scheme), the average user messages almost stay at the same level around 0.016 which is much less compared with the traditional scheme. This can be seen in Figure 6.20. The explanation is: as session interval T is fixed, the number of active sessions a user generated does not increase. So, few changes happen in the number of user messages in our proposed

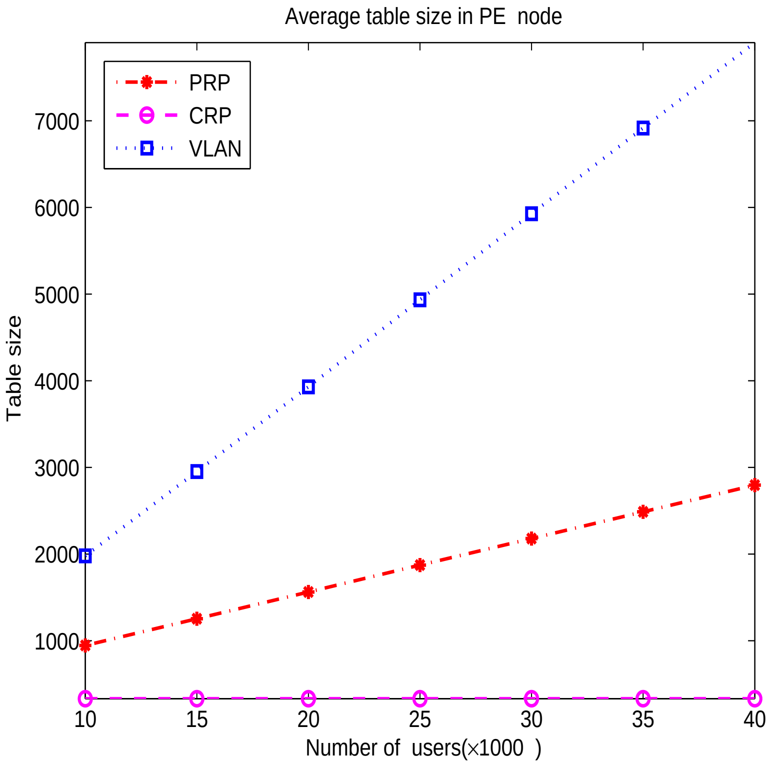


Figure 6.21: Average PE table size

schemes. However, when there are more end users, the chance a user starts a session with a previously unknown user is increased. So more broadcast messages have to be generated to find the MAC address of the unknown destinations in VLAN based scheme. Beside this, note that the table entry is timed out every 120s. This means more table entries will be deleted as the utilization frequency for individual table entry is decreased. This in turn triggers an increase in broadcast traffic, that heavies the overall traffic load.

Figure 6.21 and Figure 6.22 show the average and max PE table size of all the three schemes. VLAN based scheme requires more space than the registration based schemes because of the nature of broadcasting and inherent source address learning. The benefit becomes more obvious when more end users join in the network. The performance for PE node is better in CRP as it in this scheme do not have to maintain the end users

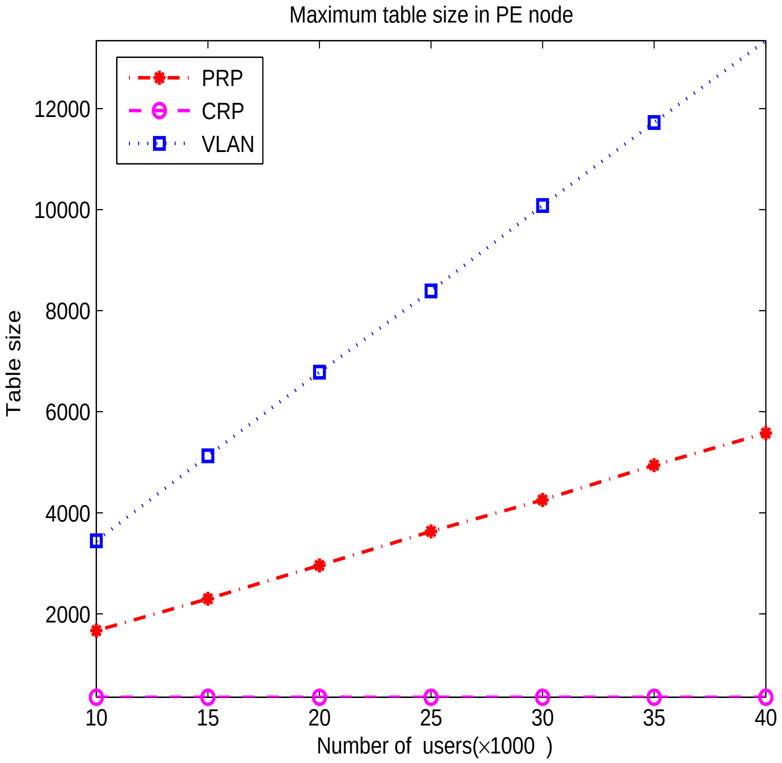


Figure 6.22: Max PE table size

registration table.

The average and max table size in CE node are shown in Figure 6.23 and Figure 6.24. Registration based schemes significantly reduce the CE node table size as compared with the legacy scheme as the same reason we mentioned previously. The reason caused the difference between PRP and CRP is that the CE node in CRP needs to store the end user registration entries while the CE node in PRP does not need.

Case 3: Impact of roaming

This case investigates the user roaming impact on the performance. Every end user in this case has the ability of roaming. The roaming interval varies from 600s to 4200s, this is the time interval of an end user roams from its current site to another site. As end users do not do registration in the legacy VLAN based scheme, its performance is not affected

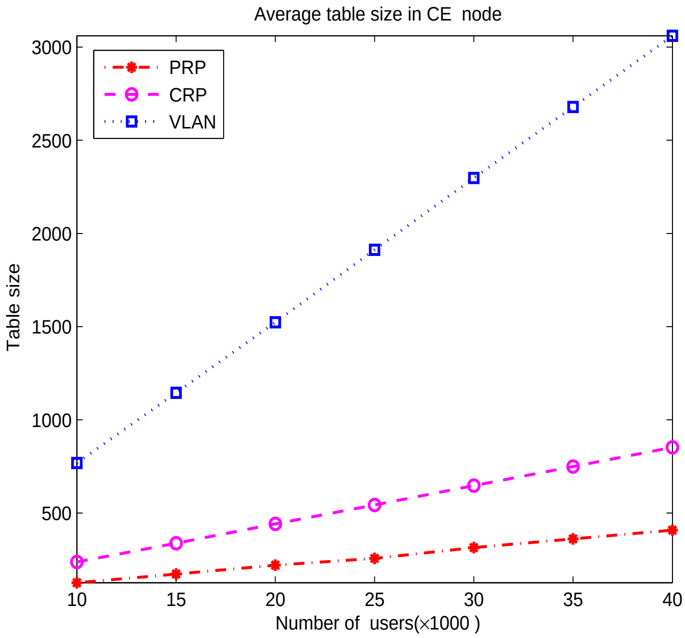


Figure 6.23: Average CE table size

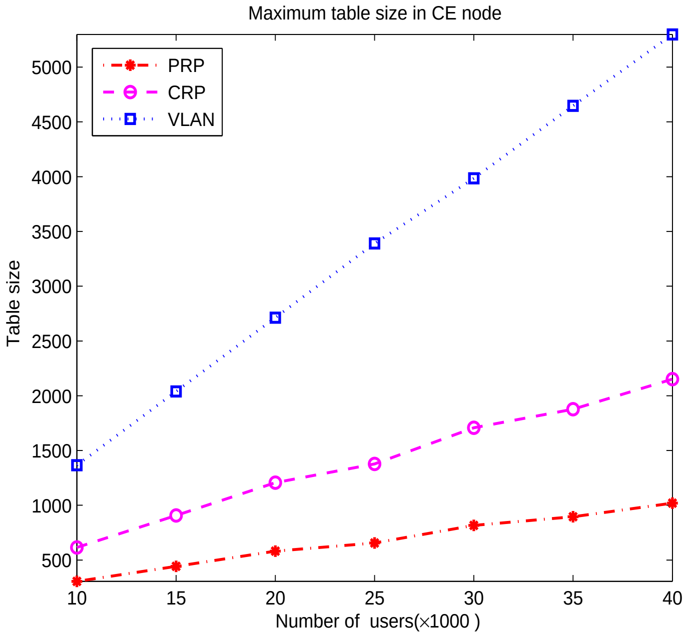


Figure 6.24: Max CE table size

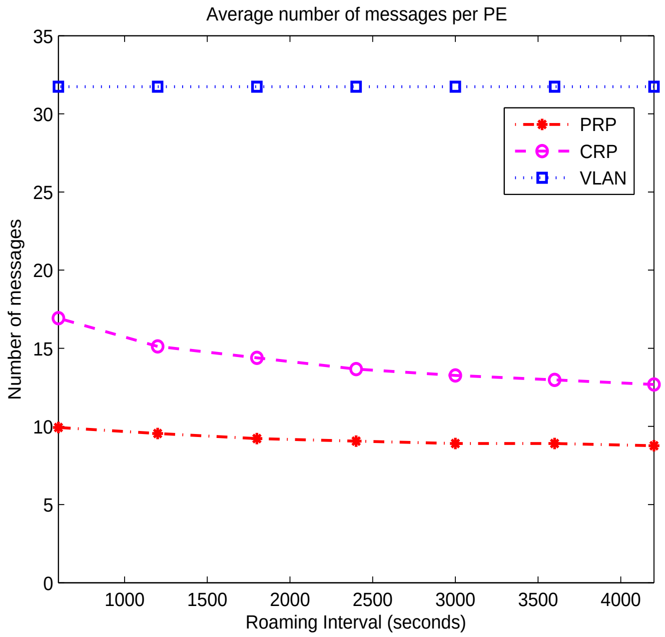


Figure 6.25: Average number of messages per PE

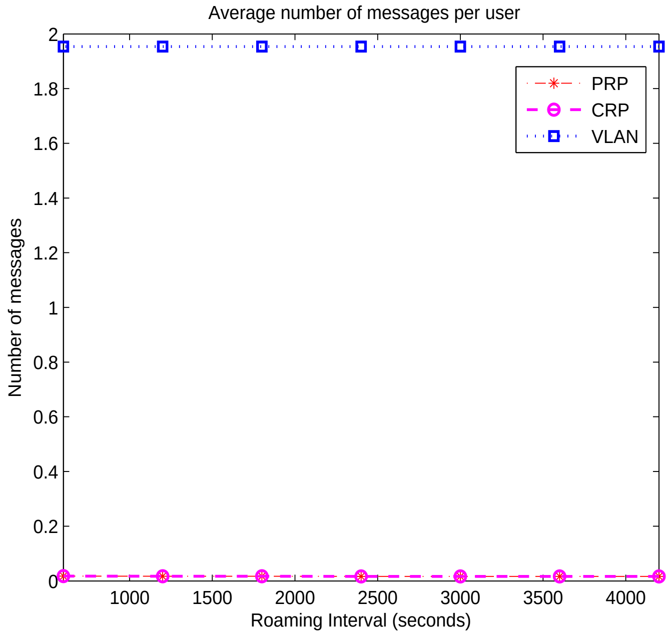


Figure 6.26: Average number of messages per user

by roaming.

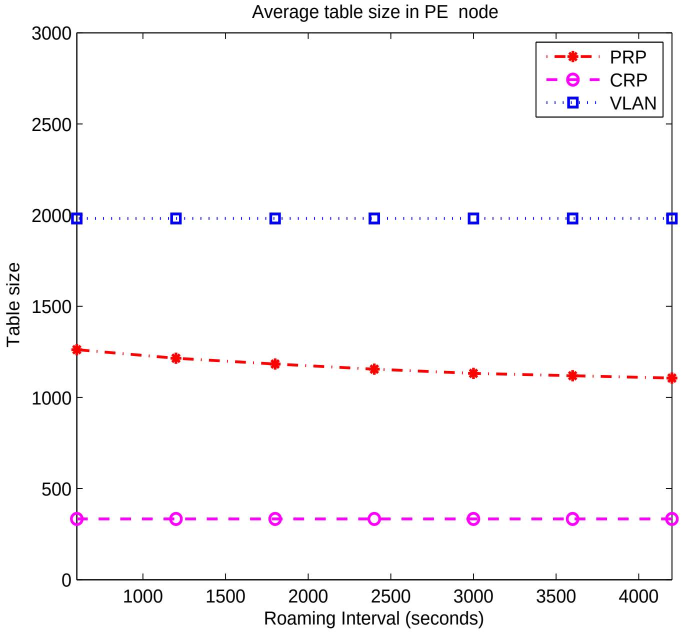


Figure 6.27: Average table size in PE node

Figure 6.25 and Figure 6.26 illustrate the average number of messages handled by a PE node and an end user per second respectively. For PRP and CRP, the number of messages handled by a PE node is almost the same, which are slightly heavier compared to that without roaming. The number of messages, however, is still much less than that of the legacy VLAN based scheme. Besides, the number of messages handled by PE nodes in RP reduces as roaming interval increases.

The maximum and average table sizes in PE nodes are shown in Figures 6.27 and 6.28. As seen in the Figures, the PE table sizes in CRP and the legacy VLAN based scheme are stable with the variance of the roaming interval, while those in PRP have a tendency of decrease. For PRP, the PE table size is larger than that without roaming. However, it is still smaller than that of the legacy VLAN based scheme.

Figure 6.29 and Figure 6.30 plot the average and maximum table sizes in CE node. The table sizes are greater in both PRP and CRP with roaming than those without roam-

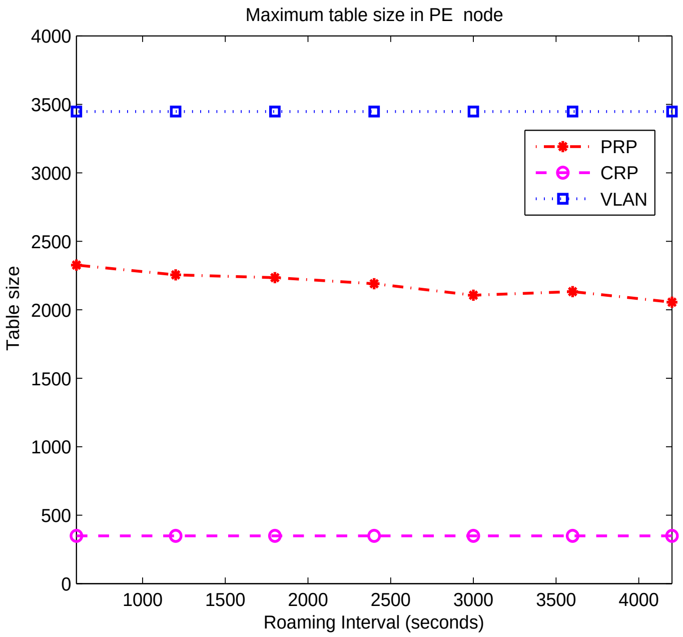


Figure 6.28: Max table size in PE node

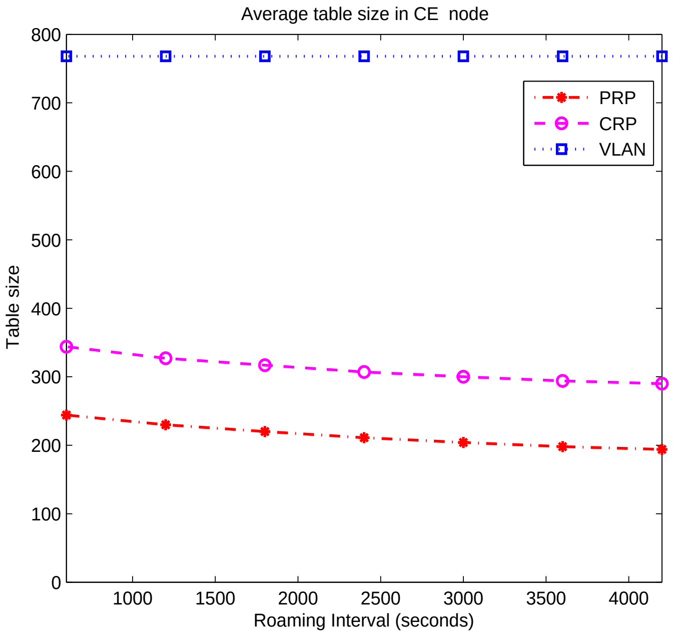


Figure 6.29: Average table size in CE node

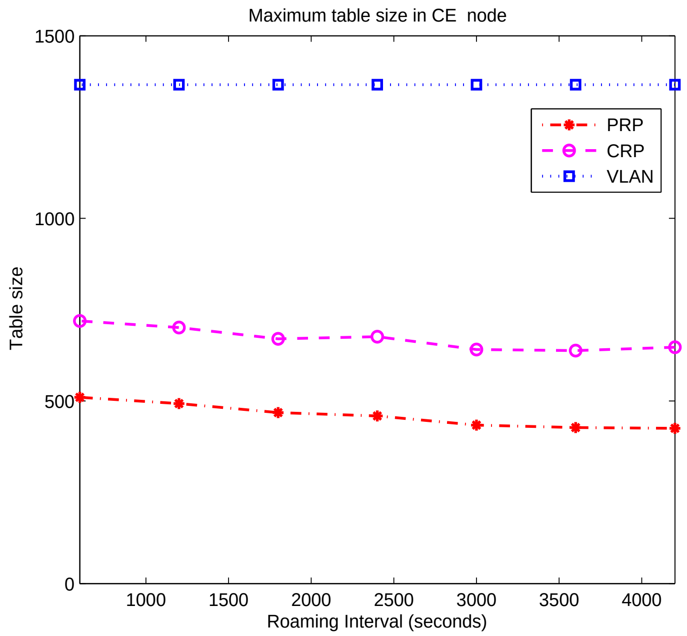


Figure 6.30: Max table size in CE node

ing. The sizes in PRP and CRP decrease as the interval of roaming increases. Compared to the legacy VLAN based scheme, the table sizes in PRP and CRP are still much less even in the situation of frequently roaming.

6.5 Summary

In order to cope with the non-hierarchical addressing and broadcast−based unknown address resolution problems, the efficient PRP and CRP schemes are proposed. The EMiM encapsulation scheme is cooperated with them. By doing MiM encapsulation in the end users, both PE and CE nodes do not need to maintain MAC mapping table. So the forwarding table explosion can be avoided. PRP and CRP are addressed to make Metro Ethernet bandwidth−efficient. In these schemes, a simple registration in PE or CE node makes it efficient to resolve an unknown MAC address. We conducted extensive simulations to evaluate the performance of PRP and CRP, and also compared them with the

VLAN scheme. The simulation results demonstrate our schemes are more robust and scalable. Moreover, they inherit the plug-and-play and self−configuration nature of Ethernet.

Chapter 7

Conclusion and Future Works

In this chapter, we first summarize the works included in this thesis. Then, we discuss the future works.

Video caching which provides cost effective and convenient content distribution services has gained significant attention recently. However, due to its large storage space consumption and time sensitiveness character, the algorithm for video caching is extremely crucial.

Highly developed Ethernet technologies have made the Ethernet an attractive proposition as a MAN transport for service providers. Hence, MEF has defined the services model for Ethernet services. However, Ethernet as mentioned above is originally designed for LAN area which handles limited number of users. When deploying Ethernet in MAN area, the efficiency and scalability have come out. Specially, the following two design issues must be considered: 1) use flat addressing scheme(i.e., non-hierarchical MAC addresses); 2) use broadcast-based mechanism to resolve the location of an unknown address. In order to overcome the scalability problems, some efficient schemes are proposed. We proposed two efficient video caching algorithms, called OCR and CCR, and scalable Metro Ethernet architectures, namely CFT, EMiM, and RP (PRP and CRP).

OCR and CCR aim to maximize the utilization of a cache for video on demand by making groups of certain users based on their arrival time. Users using the cache to record their chunks formed a group naturally after the cache is full. The chunks used by the last user of a group are eliminated. As a consequence, the cache utilization could be improved. In the CCR scheme, the caches cooperate with each other when chunk miss happens. The request could be directed and handled by a nearby cache without resorting to the original server at this circumstance. The simulation results demonstrate that the new schemes outperforms LRU scheme. With these schemes, both the request access latency and server bandwidth consumption could be improved.

CFT is an efficient cache scheme for Metro Ethernet. CFT learns the IP and MAC mapping in a frame and eliminates the subsequent broadcast frames whenever a request is answered by a cached entry. The broadcast looking for the MAC address of another end user could be stopped and answered with the help of the information stored in the new forwarding table. So, the CE and PE nodes that employ the new forwarding scheme not only could determine the next hop but also could handle the ARP request coming from the segments behind it. We also discussed with cache in EMiM. The LRU and MLRU cache replacement schemes are compared. The proposed scheme can save the broadcast messages for address resolution and reduces the forwarding table size in PE nodes. Moreover, CFT is easy to be accomplished and fully backward compatible.

To cope with the unscalable problem caused by the flat addressing, we proposed an EMiM encapsulation scheme for Metro Ethernet. By doing MiM encapsulation in the end users, both the PE and the CE nodes do not need to maintain MAC mapping table. Unlike the traditional MiM encapsulation scheme, our scheme not only reduces the forwarding table size in the CNs, but also in the PE nodes. The proposed scheme can work with any existing Metro Ethernet protocols by only modified the ARP entry format. Hence the proposed scheme can be easily implemented.

The proposed RP are addressed to make metro Ethernet bandwidth-efficient. In these schemes, a simple registration directory in PE or CE node makes them efficient to resolve an unknown MAC address. The limitations of the Metro Ethernet, such as the flat addressing and broadcast resolution to resolve unknown addresses, could be improved. The simulation results demonstrate our scheme is robust and scalable. Moreover, it inherits the plug−and−play setup and self−configuration nature of Ethernet.

7.1 Future works

In the future, for all the schemes, the detailed forwarding protocols could be considered. Many works contribute to find the suitable forwarding method in MANs. The performances of link state protocol and that of STP protocol should be compared. Moreover, how to reduce the overhead in EMiM will be investigated.

Future research area is the multicast in Metro Ethernet. Many researchers are considering to use link state protocol in Ethernet. Compared with STP protocol, link state protocol can benefit in lower bandwidth utilization and better latency. However, how to combine Ethernet multicast with link state protocol is rarely considered. Traditionally, multicast traffic in layer 2 is almost treated as broadcast and all the ports receive the frames, which is wasteful in both bandwidth and processing time.

For CFT, the roaming problems could be considered. As location flexibility is an important issue in the future Metro area, it is meaningful to do simulation with roaming and study the performance of the proposed schemes as the increase of the roaming rate. How to keep the forwarding table up to date and avoid directing frame to previous subscriber could be considered.

Bibliography

[1] Gilbert Held, “ Carrier Ethernet Providing the Need for speed”, Auerbach Publications 2008.

[2] Abdul Kasim, “ Delivering carrier Ethernet : extending Ethernet beyond the LAN”, 2008.

[3] Paul Bedell, “ Gigabit ethernet for metro area networks”, 2003.

[4] Sam Halabi, “ Metro Ethernet ”, 2003.

[5] Daniel Minoli, Peter Johnson, Emma Minoli, “ Ethernet-based metro area networks: planning and designing the provider network”, 2002.

[6] “IEEE Standard for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks”, *IEEE*, 1998.

[7] IEEE, “IEEE Standard 802.1ak for Local and Metropolitan Area Networks Virtual Bridged Local Area Networks - Amendment 07: Multiple Registration Protocol,” June 2007.

[8] M. Ali, G. Chiruvolu and A. Ge, “Traffic Engineering in Metro Ethernet”, *IEEE Networks, v19(2)*, pp10-17, 2005.

[9] G. Chiruvolu, “Issues and Approaches on Extending Ethernet Beyond LANs”, *IEEE Communication Magazine*, v42(3), pp 80-86, 2004.

[10] IEEE 802.1Q, “Virtual LANs”.

[11] IEEE, “Standard for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks,” *IEEE Std 802.1q,* 1998.

[12] Metro Ethernet Forum, “Metro Ethernet Networks - A Technical Overview,” 2004.

[13] MEF, “Ethernet Services Model, Phase I,” Nov. 2003.

[14] MEF, “Ethernet Services Definitions Phase I, Draft v5.5,” Mar. 2004.

[15] D. Cavendish. “Operation, Administration, and Maintenance of Ethernet Services in Wide Area Networks,” IEEE Communications Magazine. March 2004.

[16] S. Clavenna. “Standardizing Ethernet Services.”, Jan, 9th, 2004.

[17] DSL Forum, “Migration to Ethernet Based DSL Aggregation,” *Working Text WT-101,* Oct. 2004.

[18] E. Rosen, A. Viswanathan, and R. Callon, “Multiprotocol Label Switching Architecture,” *IETF RFC 3031*, January 2001.

[19] DSL Forum, “Migration to Ethernet Based DSL Aggregation for Architecture and Transport Working Group,” *DSL Forum WT-101, rev.3,* Oct. 2004.

[20] Telcordia, “GR253-Core: Synchronous Optical Network (SONET) Transport Systems,” 2005.

[21] ITU-T, “Recommendation G.707: Network node interface for the synchronous digital hierarchy (SDH),” 2003.

[22] ITU-T, “Recommendation G.708: Synchronous digital hierarchy (SDH) network to network interface (NNI),” 2003.

[23] Harry G. Perros, “Connection-oriented networks : SONET/SDH, ATM,MPLS, and optical networks,” 2004.

[24] Steven Shepard, “SONET/SDH Demystified,” 2001.

[25] Harry G. Perros, “An introduction to ATM networks,” 2002.

[26] DSL Forum, “Broadband Remote Access Server (BRAS) Requirements Document,” *DSL Forum TR-092*, Aug. 2004.

[27] DSL Forum, “Architecture Requirements for the Delivery of Advanced Broadband Services,” *DSL Forum 2003-427,* Nov. 2003.

[28] P. Francois, C. Filsfils, J. Evans, and O. Bonaventure, “Achieving sub-second IGP convergence in large IP networks,” *ACM SIGCOMM Computer Communication Review,* vol. 35, Issue 3, pp. 35C44, July 2005.

[29] IEEE, “IEEE.” Available at http://www.ieee.org/.

[30] G. Holland. “Carrier Class Metro Networking: The High Availability Features of Riverstones RS Metro Routers”, Riverstone Networks Technology whitepaper # 135.

[31] IETF, “L2 Virtual Private Networks (l2vpn) Working Group,” http://www.ietf.org/html.charters/l2vpn-charter.html, Sept. 2004.

[32] L. Martini, L. Tappan, D. Vlachos, C. Liljenstolpe, G. Heron, and K. Kompella, “Transport of L2 Frames over MPLS,” Internet Draft (Expired), June 2004.

[33] L. Martini, E. Rosen, N. El-Aawar, “Encapsulation Methods for Transport of Layer 2 Frames over MPLS Networks,” June 2007.

[34] K. Kompella and Y. Rehkter, “Virtual Private LAN Service,” IETF Internet Draft, May 2004.

[35] A. Kirstädter, C. Gruber, J. Riedl, and T. Bauschert, “Carrier-grade Ethernet for packet core networks,” vol. 6354, Oct. 2006.

[36] Daniel Minoli, Peter Johnson, Emma Minoli, “Ethernet-based metro area networks: planning and designing the provider network”, 2002.

[37] IEEE, “802.1Qay- Provider Backbone Bridge Traffic Engineering,” IEEE PAR, March 2007.

[38] T-Pack, “PBT: Carrier Grade Ethernet Transport,” White Paper, 2006.

[39] IEEE Std 802.3x-1997 and IEEE Std 802.3y-1997.

[40] J. Postel and J. Reynolds, “A Standard for the Transmission of IP Datagrams over IEEE 802 Networks”, *RFC 1042*, 1988.

[41] MEF, “The Metro Ethernet Network: Comparison to Legacy SONET/SDH MANs for Metro Data Service Providers”, Metro Ethernet Forum Whitepaper July 2003.

[42] MEF, “Carrier Ethernet the Technology of Choice for Access Networks”, March 2006.

[43] D. C. magazine, “ISP backbones,” September 1997.

[44] “BellSouth Metro Ethernet”, http://www.bellsouthlargebusiness.com.

[45] S. Halabi, *Metro Ethernet*, Cisco Press, 2003.

[46] M. Casado, M. J. Freedman and S. Shenker, “Ethane: Taking Control of the Enterprise”, *ACM SIGCOM*, 2007.

[47] A. Myers, T.E. Ng, and H. Zhang, “Rethinking the Service Model: Scaling Ethernet to a Million Nodes”, *Third Workshop on Hot Topics in Networks (HotNets-III)*, 2004.

[48] R. Pallos, J. Farkas, I. Moldovan, and C. Likovszki, “Performance Evaluation of the Rapid Spanning Tree Protocol in Access and Metro Networks,” *IEEE AccessNets 2007, Ottawa, Canada.,* August 2007.

[49] Riverstone Networks, “Scalability of Ethernet Services Networks”, http://www.riverstonenet.com/solutions/ethernet_scalability.shtml.

[50] M. Batayneh, D. A. Schupke, M. Hoffmann, A. Kirstaedter, and B. Mukherjee, “On reliable and cost-efficient design of carrier-grade ethernet in a multi-line rate network under transmission range constrains,” *Post Deadline paper, OFC07,* March 2007.

[51] S. Williams, M. Abrams, C. R. Standridge, G. Abdulla, and E. A. Fox, ”Removal policies in network caches for world-wide web documents,” *SIGCOMM ’96* , 1996.

[52] J. E. Pitkow and M. M. Recker, ”A simple yet robust caching algorithm based on dynamic access patterns , ” *2nd WWW Conference*, 1994.

[53] R. Wooster and M. Abrams, ”Proxy caching that estimates page load delay,” presented at the 6th WWW Conference, Santa Clara, CA, Apr. 1997.

[54] L. Breslau, P. Cao, L. Fan,G. Phillips, and S. Shenker, ”Web caching and Zipf-like distributions: evidence and implications,” *INFOCOM’99* , 1999.

[55] P. Cao and S. Irani, ”Cost-aware WWW proxy caching algorithms,” *1997 USENIX Symp. Internet Technology and Systems*, 1997.

[56] A. P. Foong, Y.Hu, and D.M.Heisey, ”Logistic regression in an adaptive web caching,” *IEEE Internet Comput.*, 1999.

[57] M. Kharbutli and Y. Solihin, ”Counter-based cache replacement algorithms,” *in Proceedings of the 23rd International Conference on Computer Design*, 2005.

[58] M. K. Qureshi, M. A. Suleman, and Y. N. Patt, ”Line distillation: Increasing cache capacity by ?ltering unused words in cache lines,” *in Proceedings of the 13th International Symposium of High-Performance Computer Architecture*, 2007.

[59] J. Alghazo, A. Akaaboune, and N. Botros, ”Sf-lru cache replacement algorithm,” *in Records of the International Workshop on Memory Technology, Design and Testing*, 2004.

[60] D. Bao, X.Li, ”A Cache Scheme Based on LRU-Like Algorithm,” *Proceedings of the 2010 IEEE International Conference on Information and Automation*, 2010.

[61] J. Shodong, A. Bestavros, and A. Iyengar, ”Accelerating internet streaming media delivery using network-aware partial caching,” *in Proc. 22nd Int. Conf. Distributed Computing Systems*, 2002.

[62] S. Sen, J. Rexford, and D. F. Towsley, ”Proxy prefix caching for multimedia streams,” *in Proc. IEEE INFOCOM*, 1999.

[63] S. Chan and F. Tobagi, ”Caching Schemes for Distributed Video Services,” 1999.

[64] BALAFOUTIS, E., PANAGAKIS, A., LAOUTARIS, N., AND STAVRAKAKIS, I., ”The impact of replacement granularity on video caching”. *In IFIP Networking 2002*, 2002.

[65] Podlipnig, S. and B?sz?rmenyi, L., ”Replacement strategies for quality based video caching”, *In Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2002.

[66] Wei-hsiu Ma and David H. C. Du, ”Design a Progressive Video Caching Policy for Video Proxy Servers,” *IEEE TRANSACTIONS ON MULTIMEDIA*, VOL. 6, NO. 4, pp. 599-610, 2004.

[67] S. Acharya and B. Smith, ”Middleman: A Video Caching Proxy Server”. *in Proceedings of the 10th International Workshop on Network and Operating System Support for Digital Audio and Video*, 2000.

[68] Ulas C. Kozat, Oztan Harmanc, Sandeep Kanumuri, Mehmet Umut Demircin and M. Reha Civanlar, ”Peer Assisted Video Streaming With Supply-Demand-Based Cache Optimization,” IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 11, NO. 3, p. 494-508, 2009.

[69] M. Hofmann, T. E. Ng, K. Guo, S. Paul, and H. Zhang, Caching techniques for streaming multimedia over the internet, Bell Labs, Holmdel, NJ, Tech. Rep. BL011345-990409-04TM, 1999.

[70] E. Balafoutis and I. Stavrakakis, Proxy caching and video segmentation based on request frequencies and access costs, *IEEE International Conference on Telecommunications (ICT03)*, 2003.

[71] J. Z. Wang and P. S. Yu, Fragmental proxy caching for streaming multimedia objects, *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 147C156, Jan. 2007.

[72] 29] L. Shen, W. Tu, and E. Steinbach, A flexible starting point based partial caching algorithm for video on demand, *IEEE International Conference on Multimedia and Expo (ICME07)*, Beijing, China, July 2007.

[73] W. Tu, E. Steinbach, M. Muhammad, and X. Li, Proxy caching for video on demand using flexible starting point selection, *in Proc. IEEE Transactions on Multimedia*, VOL. 11, NO. 4, p. 716-729, 2009.

[74] D.C. Plummer, “An Ethernet Address Resolution Protocol or Converting Network Addresses to 48.bit Ethernet Address for Transmission on Ethernet Hardware”, *RFC 826*, 1982.

[75] R. Droms, “Dynamic host configuration protocol”, *RFC 2131*, 1997.

[76] ”Problems with broadcasts”, http://www.ists.dartmouth.edu/classroom/crs/arpbroadcast.php.

[77] IEEE Std. 802.1Q, “Virtual Bridged Local Area Networks”.

[78] IEEE 802.1ah, “Provider Backbone Bridges”.

[79] IEEE, “802.1ad - Provider Bridges,” *tech. rep., IEEE,* May 2006.

[80] Nortel Networks, “Service Delivery Technologies for Metro Ethernet Networks” Nortel Networks Whitepaper Sept. 19 2003 http://www.nortel.com/solutions/optical/collateral/nn-105600-0919-03.pdf.

[81] I. Hadzic, “Hierarchical MAC address space in public Ethernet networks,” *Globecom,* 2001.

[82] IEEE, “802.1D Standard for local and metropolitan area networks - Media Access Control (MAC) Bridges,” June 2004.

[83] K. Elmeleegy, A. L. Cox, and T. S. E. Ng, “On Count-to-Infinity Induced Forwarding Loops in Ethernet Networks,” *IEEE Infocom 2006,* 2006.

[84] IEEE, “Multiple Spanning Trees,” IEEE Std 802.1s.

[85] IEEE, “Multiple Spanning Trees,” *IEEE Std 802.1s.*

[86] Ibanez G., Garcia A., Azcorra A., “Alternative multiple spanning tree protocol (AM-STP) for optical Ethernet backbones” *The 29th Annual IEEE International Conference on Local Computer Networks*, 2004.

[87] S. Sharma, K. Gopalan, S. Nanda, and T. Chiueh, “Viking: A Multispanning Tree Ethernet Architecture for Metropolitan Area and Cluster Networks,” *Infocom,* 2004.

[88] K. Lui, W. Lee, and K. Nahrstedt, "STAR: A Transparent Spanning Tree Bridge Protocol with Alternate Routing," *ACM SIGCOMM 2002,* July 2002.

[89] S. Acharya, B. Gupta, P. Risbood, A. Srivastava. "PESO: Low Overhead Protection for Ethernet over SONET Transport", IEEE INFOCOM, 2004.

[90] T. Gimpelson. "Atrica makes Ethernet resilient", Network World Fusion http://www.nwfusion.com/edge/news/2002/0122atrica.html Jan, 02, 2002.

[91] S. Varadarajan, T. Chiueh "Automatic Fault Detection and Recovery in Real Time Switched Ethernet Networks", IEEE INFOCOM 1999.

[92] G. Ibanez, A. Garcia-Martinez, A. Azcorra, I. Soto, "ABridges: Scalable, self-configuring Ethernet campus networks", *computer Networks*, 2008.

[93] Huynh Minh, Mohapatra Prasant, Goose Stuart, "Cross-over spanning trees Enhancing metro ethernet resilience and load balancing", *BROADNETS*, 2007.

[94] Minh Huynh, Prasant Mohapatra, Stuart Goose, "Spanning tree elevation protocol: Enhancing metro Ethernet performance and QoS", *Computer Communications*, 2009.

[95] Ibanez G., Garcia A., Azcorra A., "Alternative multiple spanning tree protocol (AM-STP) for optical Ethernet backbones" *The 29th Annual IEEE International Conference on Local Computer Networks*, 2004.

[96] M.Tafti, G. Mirjalily, and S. Rajaee, "Topology Design of Metro Ethernet Networks Based on Load Balance Criterion", *Internatioal Symposium on Telecommunications*, 2008.

[97] IEEE, "Virtual Bridged Local Area Networks - Amendment 9: Shortest Path Bridging," *P802.1aq/D0,3 Draft Amendment to IEEE Std 802.1q- 2005,* May 2006.

[98] R. Perlman, S. Gai, and D. G. Dutt, “Rbridges: Base Protocol Specification,” *IETF Draft (work in progress),* March 2007.

[99] R. Perlman, “Rbridges: Transparent routing”. *IEEE Infocom*, 2004.

[100] R. Garcia, J. Duato, and F. Silla, “LSOM: A Link State Protocol Over Mac Addresses for Metropolitan Backbones Using Optical Ethernet Switches”, *The Second IEEE International Symposium on Network Computing and Applications (NCA)*, 2003.

[101] IEEE 802.1D-2004, “Spanning Tree Protocol revision of IEEE std 802.1D”.

[102] IEEE 802.1S, “Multiple Spanning Tree”.

[103] IEEE 802.1W, “Rapid Reconfiguration of Spanning Tree”.

[104] IETF TRILL Working group, “TRILL 2010”, http://datatracker.ietf.org/wg/trill/.

[105] D. Oran (Editor), “OSI IS-IS Intra-domain Routing Protocol,” *Request for Comments 1142, Internet Engineering Task Force,* Feb. 1990.

[106] Huynh Minh, Mohapatra Prasant, “A Scalable Hybrid Approach to Switching in Metro Ethernet Networks Local Computer Networks”, *The 32nd IEEE Conference on Local Computer Networks*, 2007.

[107] S. Ray, R. A. Guerin and R. Sofia, “A Distributed Hash Table Based Address Resolution Scheme for Large-Scale Ethernet Networks”, *IEEE ICC*, 2007.

[108] Changhoon Kim, Rexford, J., “Revisiting Ethernet: Plug-and-play made scalable and efficient”, *The 15th IEEE Workshop on Local and Metropolitan Area Networks* , 2007.

[109] KIM, C., CAESAR, M., GERBER, A., AND REXFORD, J. 2009. "Revisiting route caching: The world should be flat," *In Proceedings of the Symposium on Passive and Active Measurement.*

[110] Changhoon Kim, Matthew Caesar, and Jennifer Rexford, "Floodless in SEATTLE: A Scalable Ethernet Architecture for Large Enterprises", *ACM SIGCOMM*, 2008.

[111] Kim, C., Caesar, M., and Rexford, J., "SEATTLE: A Scalable Ethernet Architecture for Large Enterprises," *ACM Transactions on Computer Systems (TOCS),* Vol. 29, February 2011.

[112] T. L. Rodeheffer, C. A. Thekkath and D. C. Anderson, "SmartBridge: a Scalable Bridge Architecture", *ACM SIGCOMM*, 2000.

[113] E. W. Dijkstra and C. S. Scholten, "Termination Detection for Diffusing Computations," *Information Processing Letters,* vol. 11(1), pp. 1C4, August 1980.

[114] P. Wang, C. Chan, and P. Lin, "Translation for Enabling Scalable Virtual Private LAN Service", *21st International Conference on Advanced Information Networking and Applications Workshops*, 2007.

[115] Ravi Kumar Buregoni, "A Unified Distributed Directory based Service Delivery Architecture for Metro Ethernet Networks", *11th International Conference on Advanced Communication Technology*, 2009.

[116] K. Elmeleegy and A. L. Cox, "EtherProxy: Scaling Ethernet By Suppressing Broadcast Traffic", *The 28th Conference on Computer Communications. IEEE*, 2009.

[117] GREENBERG, A., HAMILTON, J., JAIN, N., KANDULA, S., KIM, C., LAHIRI, P., MALTZ, D., PATEL, P., AND SENGUPTA, S. 2009. "VL2: A scalable and flexible data center network," *In Proceedings of ACM SIGCOMM.*

[118] MYSORE, R. N., PAMBORIS, A., FARRINGTON, N., HUANG, N., MIRI, P., RADHAKRISHNAN, S., AND SUBRAM, V. 2009. “PortLand: A scalable fault-tolerant layer 2 data center network,” *In Proceedings of ACM SIGCOMM*.

[119] SCOTT, M. AND CROWCROFT, J. 2008. “MOOSE: Addressing the Scalability of Ethernet,” *In Proceedings of EuroSys (Poster session)*.

[120] Dongmei Wang, Lynch D., Jian Li, Klincewicz J., Guangzhi Li, Doverspike R. and Segal M. “ Design of metro Ethernet networks,” *2010 17th IEEE Workshop on Local and Metropolitan Area Networks (LANMAN),* May 2010.

[121] M. Kaplan, M. Haenlein, “ Users of the world, unite! The challenges and opportunities of Social Media,” *Business Horizons*, 53, S. 59-68, 2010.

[122] D. Bao and X. Li, “ A cache scheme based on LRU-like algorithm,” *2010 IEEE International Conference on Information and Automation*, 2010.

[123] U. Chejara, H. Chai, and H. Cho, “ Performance Comparison of Different Cache-Replacement Policies for Video Distribution in CDN,” *Proc. High Speed Net. Multimedia Commun.*, 2004.

[124] J. Liu, J. Xu, and X. Chu, “ Fine-Grained Scalable Video Caching for Heterogeneous Clients,” *IEEE Trans. Multimedia.*,2006.

[125] J. Liu and J. Xu, “ Proxy caching for media streaming over the Internet,” *IEEE Commun. Mag.*, 2004.

[126] K. Katsaros, G. Xylomenos, and G. C. Polyzos, “ MultiCache: an incrementally deployable overlay architecture for information-centric networking,” *INFOCOM Work-in-Progress (WiP)*, 2010.