# An efficient and Scalable Metro-Ethernet Architecture

Xiaocui Sun and Zhijun Wang
*Department of Computing, The Hong Kong Polytechnic University, Hong Kong*
*Email: csxcsun@comp.polyu.edu.hk*

***Abstract***

*Ethernet has gained popularity for deploying in Metropolitan Area Networks (MANs) due to its ease of management and highly cost effective. While the flat addressing scheme (i.e., non-hierarchical MAC addresses) and the broadcast based address resolution scheme simplify many aspects of configuration, they also bring poor scalability. This paper aims at developing an efficient and scalable Metro-Ethernet architecture. To achieve this goal, we propose a Distributed Registration based Address Resolution Protocol (DRARP) and an End user enabled Mac-in-Mac (EMiM) encapsulation scheme. DRARP solves an unknown address through unicast, and hence eliminates the broadcast messages. EMiM does Mac-in-Mac (MiM) encapsulation by end users instead of the Provider Edge (PE) nodes, thus significantly reducing the PE node's forwarding table size. The proposed architecture sustains the Ethernet's plug-and-play feature and provides high scalability. The simulation results show that the proposed schemes can save more than 60% communication messages for address resolution and up to 80% forwarding table size in PE nodes.*

***Keywords:*** *Metro-Ethernet Mac-in-Mac Scalability*

## 1. Introduction

Ethernet has increasingly become an important technology to replace layer 3 technologies for deploying Metropolitan Area Networks (MANs) due to its following nice features. Firstly, it requires minimal management and maintenance cost and is ubiquitously deployed. Network administrators are already familiar and comfortable with it; Secondly, it provides the lowest per port cost, which increases sub-linearly as the port capacity increases; Thirdly, almost all the end systems support Ethernet interfaces; Finally, more and more companies run geographically distant campuses, offices, and data servers. The ability of creating virtual LAN environments that link distant campuses, offices and servers together brings significant value to those companies for effective information exchanges and data sharing.

The simplicity, flexibility and cost effective make Ethernet extremely attractive to be deployed in MANs. However, Ethernet originally designed as a Local Area Network (LAN) technology that usually handles a small number of users has some scalability limitations [3] [4] [6]. On one hand, Ethernet uses a broadcast-based address resolution scheme. Some protocols such as ARP [7] and DHCP [8] use 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 [9], high frequency broadcast messages waste a lot of bandwidth for address resolution. Moreover, every end user needs to take resource to handle every broadcast message [10]. On the other hand, it has poor scalability due to the use of a flat addressing scheme (i.e., non-hierarchical MAC addresses). 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 need excessive frame flooding, depending on the actual timeout values for the table entries. To deploy Ethernet technology to MAN, it has to solve all of the above issues.

In this paper, we propose an efficient and scalable Metro-Ethernet architecture. The proposed architecture includes a Distributed Registration based Address Resolution Protocol (DRARP) and an End user enabled Mac-in-Mac (EMiM) encapsulation scheme. In DRARP, multiple ARP registers (acted by Provider Edge (PE) nodes or Customer Edge (CE) nodes) 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 node 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. 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 do address swapping and hence do not need to maintain the entries mapping end user's MAC address to its PE node's MAC address, thus significantly reducing their forwarding table size. The proposed architecture is evaluated by simulations. The results show the proposed schemes can save more than 60% messages for address resolution and reduces up to 80% forwarding table size in PE nodes.

The rest of the paper is organized as follows: Section II gives the related works. Section III presents the details of the proposed architecture including DRARP and EMiM. The performance comparison is shown in Section IV. Finally, Section V concludes the paper and describes some future works.

## 2. Related work

One of the most important challenges by deploying Ethernet in MANs is scalability. There are numerous works focused on the development of scalable solutions for Ethernet. The VLAN technology [21] is a useful scheme to make Metro-Ethernet scalable. It uses a VLAN tag [22][4] to partition a single large Ethernet to multiple VLANs. The broadcasting messages for address resolution are limited to a single VLAN instead of the whole network, and hence the redundancy traffic is reduced. The VLAN tag has only 12 bits and hence can only support up to 4096 active VLANs at any time in the network. To support more active VLANs, Q-in-Q or VLAN stacking encapsulation [23] is proposed. In this scheme, a PE node inserts an additional Q-tag in the Ethernet frame to support more active VLANs. However, this scheme still cannot avoid learning table explosion. The MAC-in-MAC (MiM) encapsulation [4][13] can be used to reduce the forwarding table size in the CNs, but the PE nodes still need to keep a potential large number of forwarding entries for MiM encapsulation.

Some research works focus on spanning tree protocols to enhance scalability. In [20] and [35], an alternative multiple spanning tree is developed to minimize hop distance among switches. Cost [29] and STEP [30] enhanced the resiliency of the MAN and had the potential of utilizing multiple spanning trees. [31] constructed an efficient algorithm to find the best spanning tree with the consideration of both shortest path selection and load balancing on links and switches. The same idea was employed to discover more suitable Metro-Ethernet structure on the basis of load balance in [32]. In [14] and [15], the link state routing protocols were proposed to replace the spanning tree based routing schemes [16][17][18]. In [19], 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. However, all these schemes still have to use the broadcast based ARP.

Another way to improve the scalability is to use directory based ARP [6] 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. 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 [24] [25] [26] 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. But routing frame with unknown destination addresses to designated user may take the frame travel more unnecessary hops, and make the traffic control more difficult. SmartBridge [12] allows finding the shortest forwarding path by exchanging topology information among bridges. Hence a full knowledge of the network topology should be obtained. In [27], a MAC address translation 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. However, it is possible that two MAC addresses are translated to the same structured address. Both [33] and [5] 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 Relay PE maintaining the ARP entry, broadcasting messages can be reduced [33]. Similarly, Etherproxy in [5] caches the ARP entries it learned and suppresses broadcast messages by looking up the cached entries. It retains the plug and play nature of Ethernet and is backward compatible. However, broadcast still happens under the Etherproxy, which scales with the size of the network.

## 3. A scalable and efficient Metro-Ethernet architecture

Figure 1 shows a general Metro-Ethernet including a provider network and multiple LAN segments. A provider network is composed of multiple Provider Edge (PE) nodes and Core Nodes (CNs) (switch or bridge). A LAN segment includes a Customer Edge (CE) node and multiple end users. The proposed Metro-Ethernet architecture includes a Distributed Registration Based Address Resolution Scheme (DRARP) and an End User Enabled Mac-in-Mac (EMiM) scheme. DRARP eliminates the broadcast service for address resolution, and EMiM reduces the PE nodes' forwarding table size. In the following, the details of DRARP are presented.

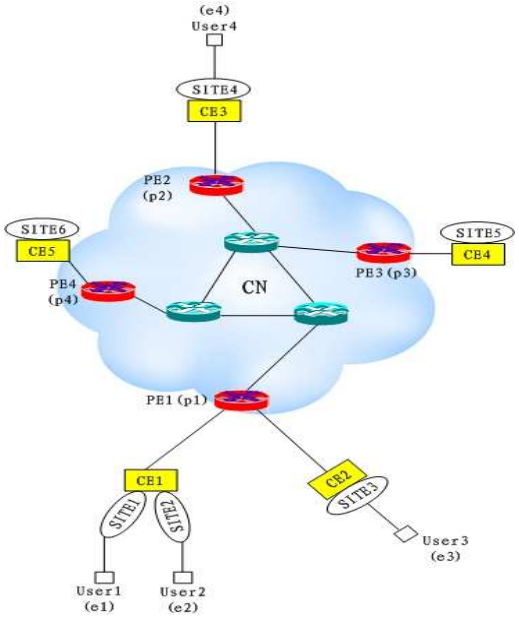


**Figure 1. A Metro-Ethernet**

### 3.1. Distributed Registration Based Address Resolution (DRARP) Scheme

In the proposed DRARP scheme, 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.

**3.1.1. Home and foreign ARP register:** Each PE node is in responsible for all of the ARP entries belonging to a set of IP prefixes. The PE node is called as the home ARP register of these IP addresses. When an end user moves out of its home register, its current PE node is considered as its foreign ARP register. A PE node needs to store all the ARP entries which have the PE node as their home register or foreign register. A PE node also has a table storing the mapping of the IP prefix to its home register for the IP prefixes in the whole Metro-Ethernet. When a user joins in a LAN segment under its home register, it needs to register its ARP entry immediately. When an end user moves to a LAN under another PE node (as the foreign register), it needs to register its ARP entry to its foreign register immediately. The foreign register records the entry and forwards the entry to its home register. Each user needs to periodically send its ARP entry to its home/foreign register. The home register updates the ARP table whenever it is changed. When a user moves from one foreign register to another, a new ARP entry has to be sent to the home register which in turn informs the previous foreign register to remove the ARP entry. Hence each ARP entry needs to be stored at most twice in the network. The registration process is illustrated in Figure 2.

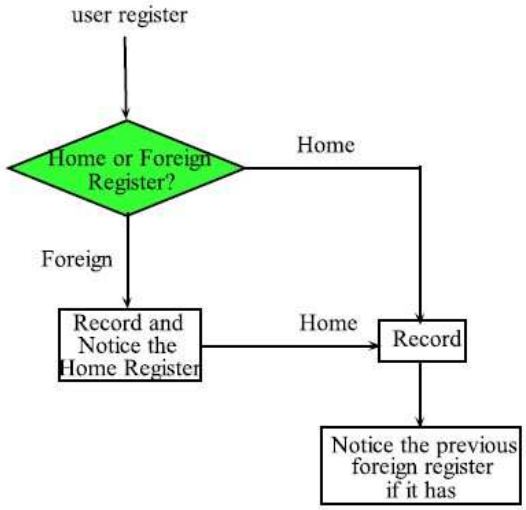


**Figure 2. ARP Registration Process**

A link state routing protocol is used for frame forwarding. Then each node only needs to record all the EMAC addresses behind the same PE node and all the PMAC addresses if the EMiM encapsulation scheme is used.
In summary, the entries in the ARP table in a PE node include:

- All IP addresses using the PE node as their home ARP register;
- All IP prefixes in the network;
- All IP addresses using the PE node as their foreign register.

### 3.2. End user enabled MiM (EMiM) encapsulation scheme

The DRARP scheme eliminates the broadcast messages for address resolution. Now we design an End user enabled MiM (EMiM) encapsulation scheme [34] to reduce the forwarding table size. The existing MiM encapsulation scheme is done by the PE node so that a PE node needs to maintain the entries of mapping end user's MAC address to PE's MAC address. This may still make the mapping table overflow. In EMiM, the MiM encapsulation is done by the end user instead of the PE node. To allow an end user doing MiM encapsulation, the PE node's MAC address is associated with an ARP entry, i.e., an ARP entry including IP, End user's MAC (EMAC), and its PE's MAC (PMAC). The associated PMAC address allows an end user to do MiM encapsulation, and 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. Note that the PMAC in the ARP entry is the end user's current PE node address.

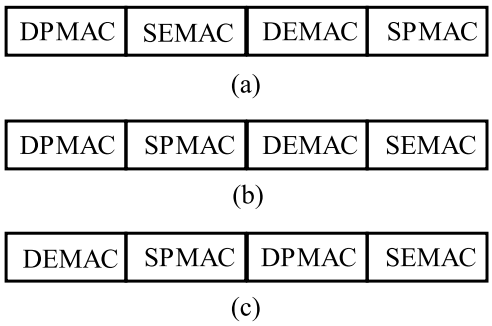


**Figure 3. ARP response and data frame format in EMiM sent by: (a) the sender to the LAN segment; (b) the sender's PE node to the CN node; (c) the receiver's PE node to its LAN segment.**

The entry format of the ARP table is <IP, EMAC, PMAC*, recordtime, age*>, here IP is the end user's IP address; *recordtime* is the time when this ARP is created or refreshed; and *age* indicates if the ARP entry is valid or not. The EMiM encapsulation scheme is used in data frames. The data frame format is as following: if the destination is behind the same PE node, the first pair of addresses are set to DEMAC (Destination EMAC) and SEMAC (Source EMAC), and the second pair of addresses are set to both SPMAC (in this case SPMAC is the same as DPMAC). All the nodes just use the first pair of addresses for frame forwarding and no address swapping is needed. If the destination is in a LAN segment behind another PE node. The two PE nodes need to swap the pair of MAC addresses. The frame formats are shown in Figure 3. Figure 3(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 3(b) gives the frame format sent by the sender's PE node to the CNs, and the source address is swapped to the sender's PE node's address; Figure 3(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 swapped to be the source address.

**3.2.1. Address resolution process and data frame format:** The address resolution and frame forwarding processes 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 sents its <IP, EMAC, NULL> to its 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 that a user moves from one LAN to another belonging to a different PE node, it needs to immediately register to its new local PE node as its foreign ARP register. The foreign register has to forward its ARP entry to its home register.

When a user needs to resolve a MAC address for a destination user, 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 records the destination's IP address, EMAC and PMAC in its ARP entry. Then it is ready to start a data session.

**3.2.2. Illustration of ARP entry learning and data frame forwarding:** We use the Metro-Ethernet given in Figures 1 as an example network. Assume the MAC addresses for users 1, 2, 3 and 4 are *e1, e2, e3* and *e4*; and for PE nodes 1, 2, 3 and 4 are *p1, p2, p3* and *p4*, respectively. The IP prefixes handled by PE nodes 1, 2, 3, 4 are 61.1.0.0/16, 62.1.0.0/16, 63.1.0.0/16, 64.1.0.0/16, respectively.

**ARP registration and address resolving process**

We will discuss two cases: (1) an user is connected to its home register; and (2) an user is connected to a foreign register.

*A user connected to its home register:* Assume user 4 joins in the LAN segment for the first time, it requests its IP address through DHCP if it has no static IP address. After the DHCP server receives the request frame, it assigns IP address 62.1.1.1 to user 4. After user 4 receives the IP address, it sends a registration frame to PE node 2 for registration. The ARP entry < 62.1.1.1*, e4, p2*> is created in PE node 2. When user 1 needs to send frames to user 4, it sends an ARP request message to its local PE node (i.e., PE node 1). PE node 1 finds that the IP address 62.1.1.1 is an IP address in charged by PE node 2 and hence forwards the request to PE node 2. After PE node 2 gets the ARP request frame, it replies the corresponding ARP entry back to PE node 1 which in turn forwards to user 1. Now data frames are ready to send by user 1 to user 4. The process is illustrated in Fig 4. The process is similar when two end users reside under the same PE node.

*A user connected to a foreign register:* Now let us consider the case that user 4 is an immigrant from PE node 3. Assuming PE node 3 is the home register of user 4 and PE node 2 is its foreign register. The IP address of user 4 is 63.1.0.1. After User 4 moves to a LAN behind PE node 2, it needs to send an ARP registration frame to PE node 2 immediately. After checking its IP address, PE node 2 knows that user 4 is not in its home register, and hence PE node 2 records the ARP entry and sends the ARP entry to its home register PE node 3. This can ensure that the entry in the home register always has the most recent ARP entry. When user 1 needs to resolve the MAC address of user 4, it sends an ARP request frame to PE node 1. PE node 1 sends the request to PE node 3 after checking the IP address. PE node 3 replies an ARP response frame back. Now user 1 is ready to send data frames to user 4. When user 4 moves out of PE node 2, PE node 2 will delete its ARP entry after it is noticed. Figure 5 shows the process in this case.

**Data communication process**

Data communication between users behind the same PE node is only using the first pair of addresses. For example, frame format for communication destined to user 1 from user 3 is *e1-e3-p1-p1*. No address swapping is needed.

Data Communication between users behind different PE nodes need address swapping by both the sender's and receiver's PE nodes. Now let us look at the frame format in the communication between user 1 and user 4. The addresses in the frame sent by user 1 are *p2-e1-e4-p1*. *p2* is used

as the destination address, and hence the forwarding nodes behind PE node 1 do not need to record the forwarding entry of *e4*. PE node 1 swaps the addresses to *p2-p1-e4-e1*. When the frame arrives at PE node 2, it swaps the source addresses, and the format is *e4-p1-p2-e1*. The proposed DRARP replaces the ARP broadcast with unicast. The schemes do not need any configurations and hence maintains Ethernet's plug-and-play setup and self-configuration capability.

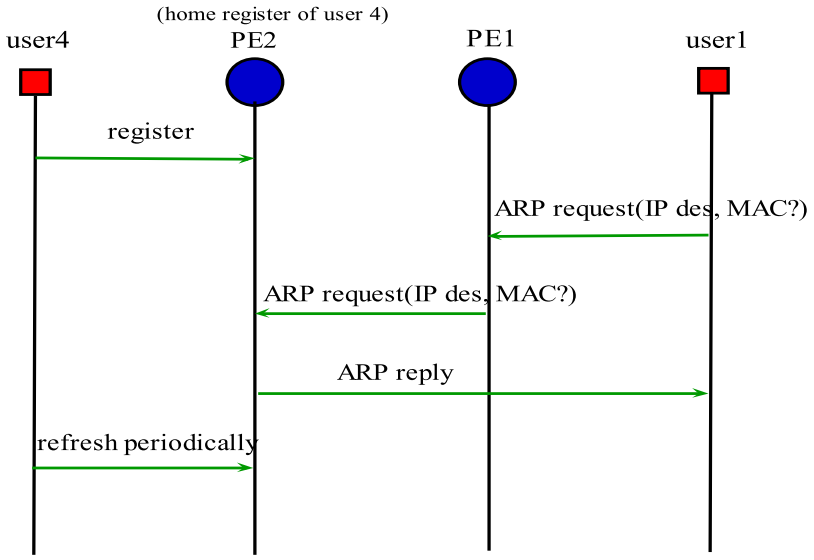


**Figure 4. ARP registration and resolving process case 1: end user locates under its home register**

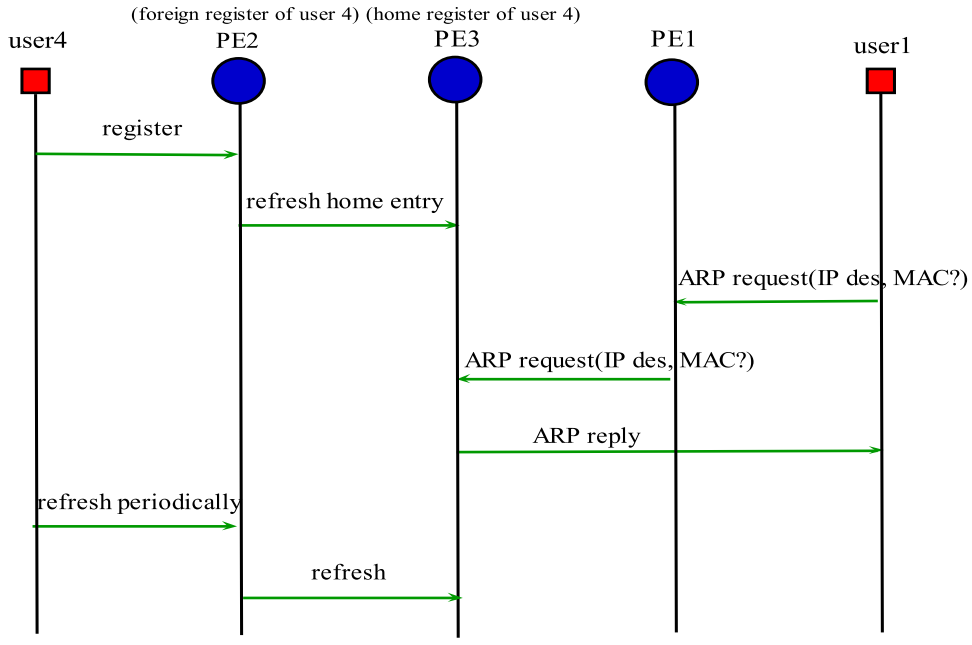

**Figure 5. ARP registration and resolving process case 2: end user locates under a foreign register**

### 3.3. Performance analysis

The proposed architecture (PEDRARP and EMiM) 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 PEDRARP is $t_r$.

We use $S_{vlan}^{max}$ and $S_{new}^{max}$ new 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 $M_{new} = 17{,}000$. That is, 99% messages can be saved.

### 3.4. Customer Edge based Distributed Registration Address Resolution Scheme (CEDRARP)

In order to further lower the registration table size, we can move the ARP register to CE nodes. The basic principles of Customer Edge based Distributed Registration Address Resolution Scheme (CEDRARP) is similar to that in PEDRARP. In CEDRARP, the entries in the ARP table of a PE node include:

(1) All IP prefixes in the network.

(2) The mapping of IP prefixes to their CE node that is behind the PE node.

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

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

(2) The IP addresses of end users use the CE node as their foreign register.

The registration and address resolution process is similar to that in PEDRARP. The EMiM encapsulation scheme is still used to transmit data session frames. In order to cooperate with CEDRARP, the entry format of the ARP table is modified as *<IP address, EMAC, CMAC, PMAC, recordtime, age>*. Here *CMAC* is the MAC address of the CE node. Three pairs of addresses are used for frame forwarding. The 2 CE nodes and 2 PE nodes need to do address swap.

### 3.5. 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 happens in the legacy VLAN based scheme. In DRARP, 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, tow cases are set. The parameters are shown in Table 1. Each PE node directly connects to at least 2 and at most 5 PE nodes. At least 2 and at most 6 CE nodes are behind a PE node. A CE node can have up to 6 sites. There are at least 8 and at most 256 end users behind a CE node. A VLAN has at least 3 and at most 13 sites. A user has probability to communicate with other end users in the same VLAN. The data session interval (to send frame) follows the Poisson distribution. The average data session interval 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 ARP entry is timed out every 120 seconds. In the simulation, we run 1600 seconds and the results of the last 600 seconds are collected. At the beginning, all the forwarding tables are set to be empty for legacy VLAN technology. For DRARP 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 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 DRARP, the PE and CE 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**

**Table 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 | 120 $s$ | 600 $s$ to 4200 $s$ |

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

Figure 6 and Figure 7 show the average number of messages a user handled per second and the average number of messages a PE node handled per second. From the Figures, 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 rate. As data session rate increases, the number of message grows rapidly in traditional systems. This is due to using broadcast to resolve unknown MAC addresses. Whereas, DRARP is much more progressive as the unknown MAC addresses can be resolved by sending a unicast request message to its PE node.

Figure 8 and Figure 9 illustrate the average table size and the maximum table size of PE node. Figure 10 and Figure 11 show the average table size and the maximum table size of CE node. In DRARP, the table of PE/CE node includes the sum of the entries in the registration ARP table and forwarding table. Both table sizes increase linearly with increasing traffic load. However, the table size of the legacy VLAN based one grows much more rapidly compared to that new scheme. Although the registration entries have to be stored, the table size is much smaller in both PEDRARP and CEDRARP. The legacy VLAN based scheme performs even worse as the increase of data session rate. In this situation the routing table of the legacy VLAN based scheme grows rapidly whereas the table size of DRARP scheme is not so sensitive to the increasing of the work load. This is because the size of registration ARP table is not affected by the growth of the traffic load.

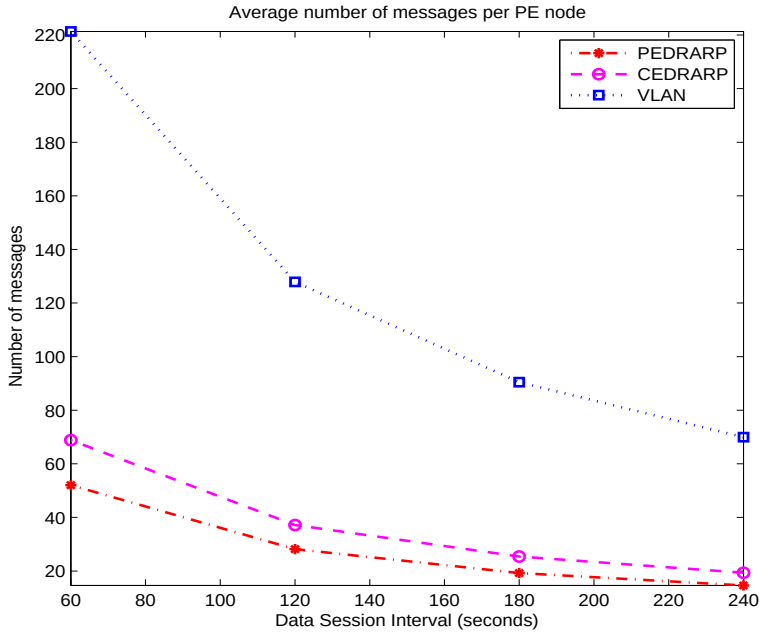


**Figure 6. Average number of messages per PE**

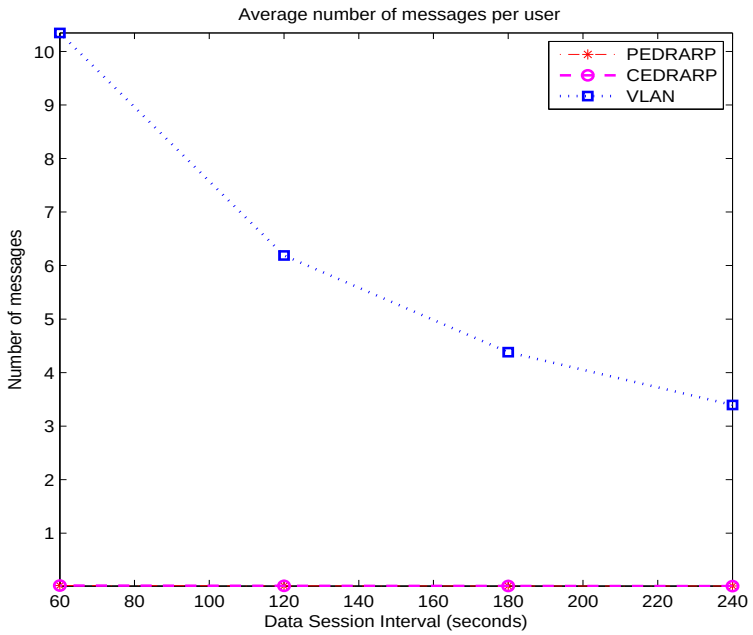


**Figure 7. Average number of messages per user**

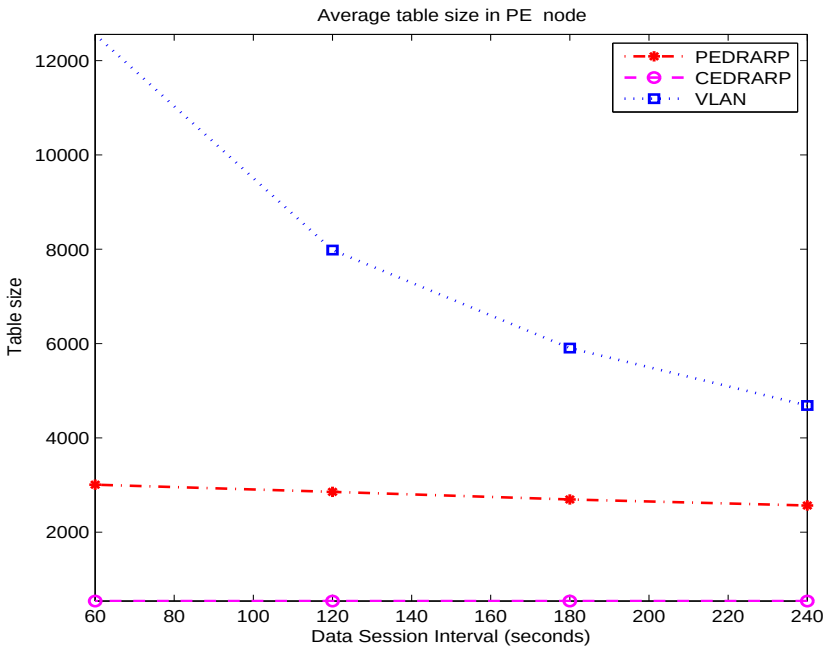


**Figure 8. Average table size in PE node**

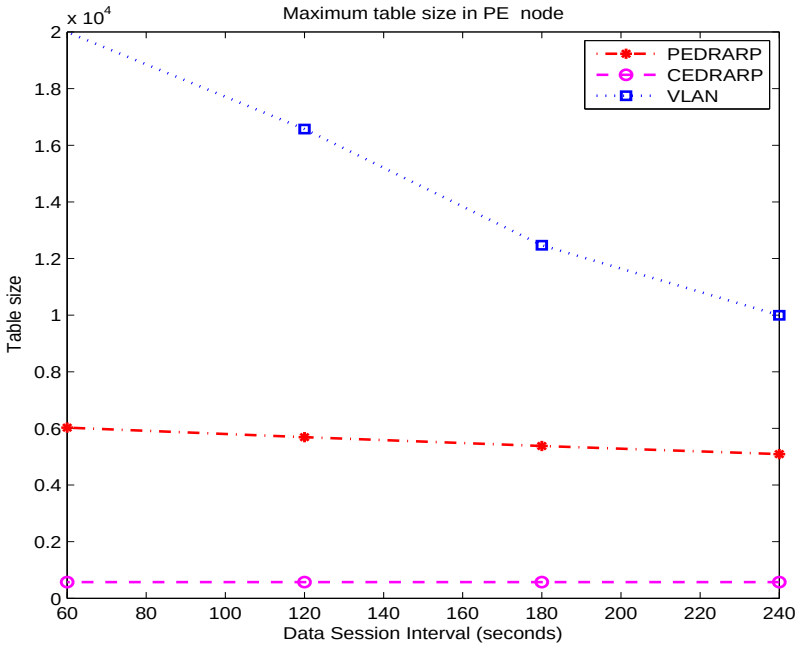


**Figure 9. Max table size in PE node**

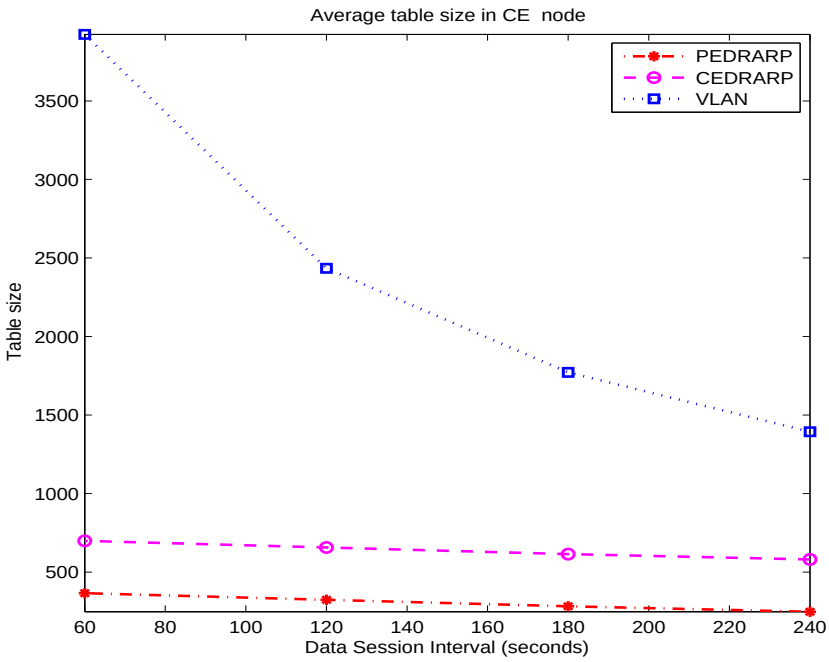


**Figure 10. Average table size in CE node**

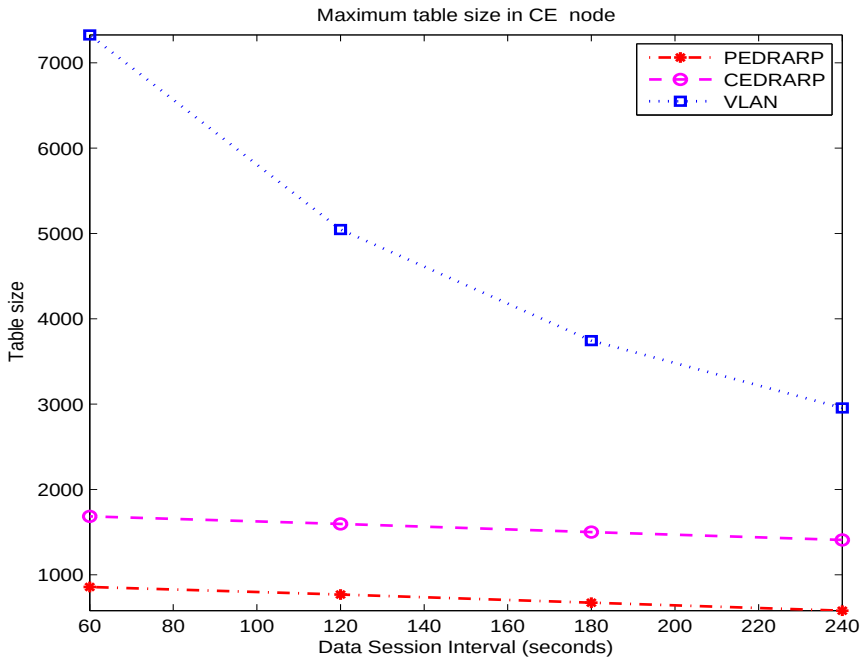


**Figure 11. Max table size in CE node**

**Case 2: Impact of roaming**

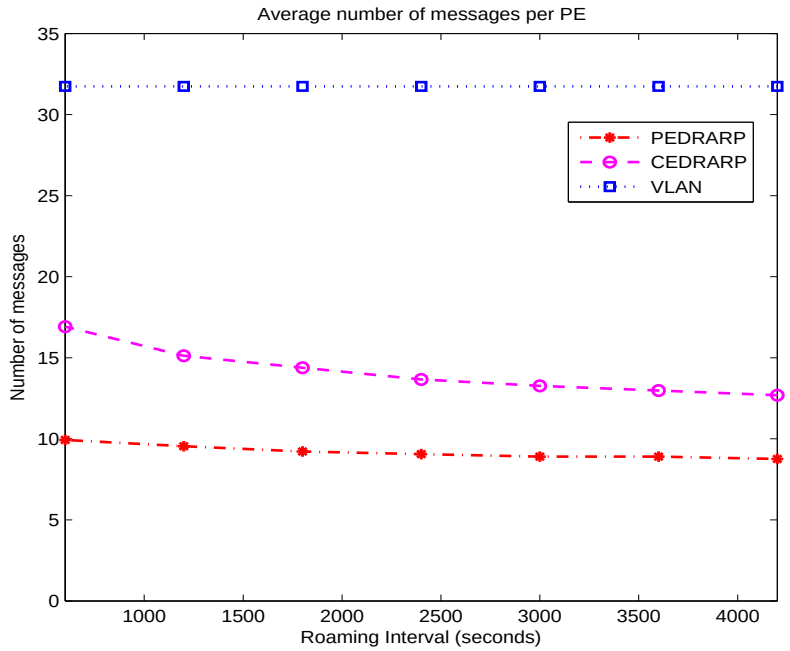


**Figure 12. Average number of messages per PE**

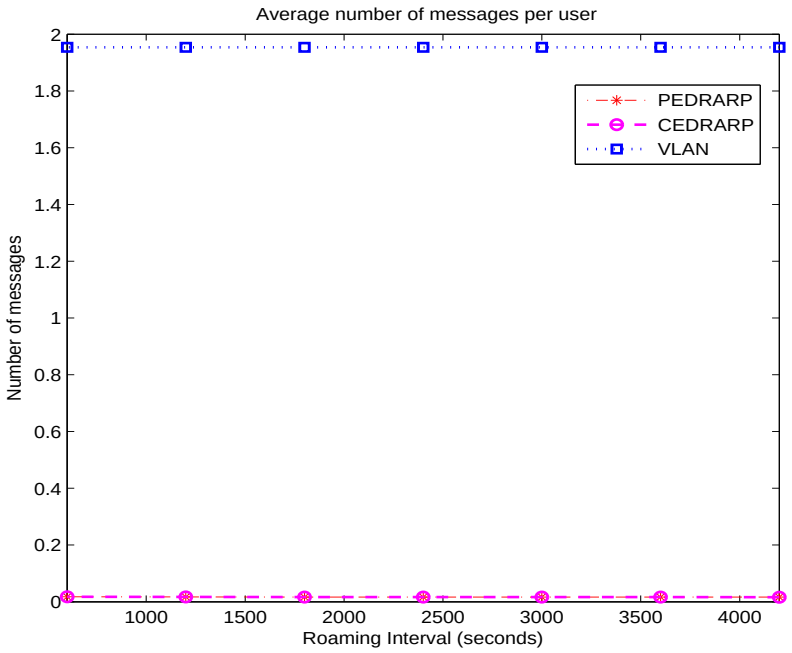


**Figure 13. Average number of messages per user**

This case investigates the user roaming impact on the performance of DRARP. 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 by roaming.

Figure 12 and Figure 13 illustrate the average number of messages handled by a PE node and an end user per second. For PEDRARP and CEDRARP, the number of messages handled by a PE node is almost the same, which are slightly heavier compared to that without roaming. The numbers 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 DRARP schemes reduces as roaming interval increases.

The maximum and average table sizes in PE nodes are shown in Figures 14 and 15. As seen in the figures, the PE table sizes in CEDRARP and the legacy VLAN based scheme are stable with the variance of the roaming interval, while those in PEDRARP have a tendency of decrease. For PEDRARP, the PE table size is larger than that without roaming. However, it is still smaller than that of the legacy VLAN based scheme.

Figure 16 and Figure 17 plot the average and maximum table sizes in CE node. The table sizes are greater in both PEDRARP and CEDRARP with roaming than that without roaming. The sizes in PEDRARP and CEDRARP decrease as the interval of roaming increases. Compared to the legacy VLAN based scheme, the table sizes in PEDRARP and CEDRARP are still much less even in the situation of frequently roaming.

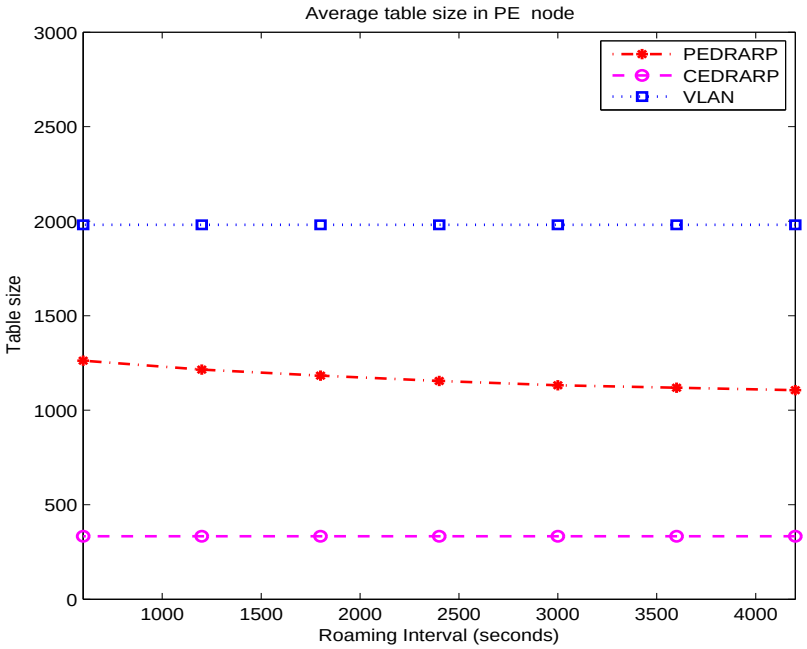


**Figure 14. Average table size in PE node**

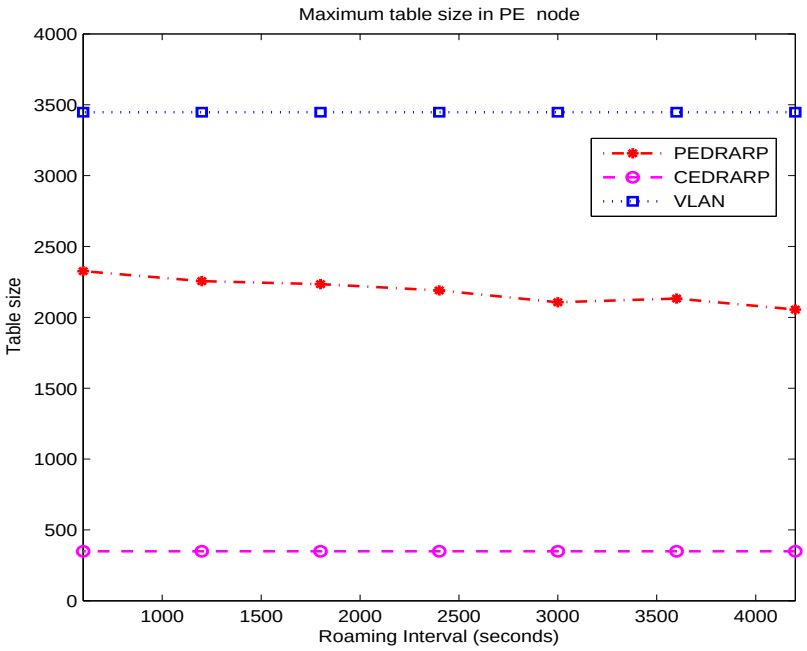


**Figure 15. Max table size in PE node**

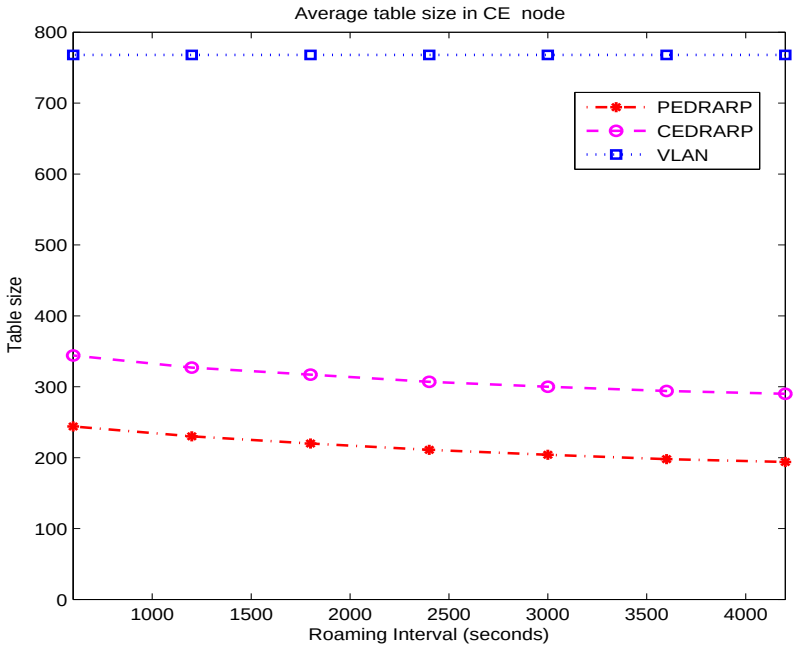


**Figure 16. Average table size in CE node**

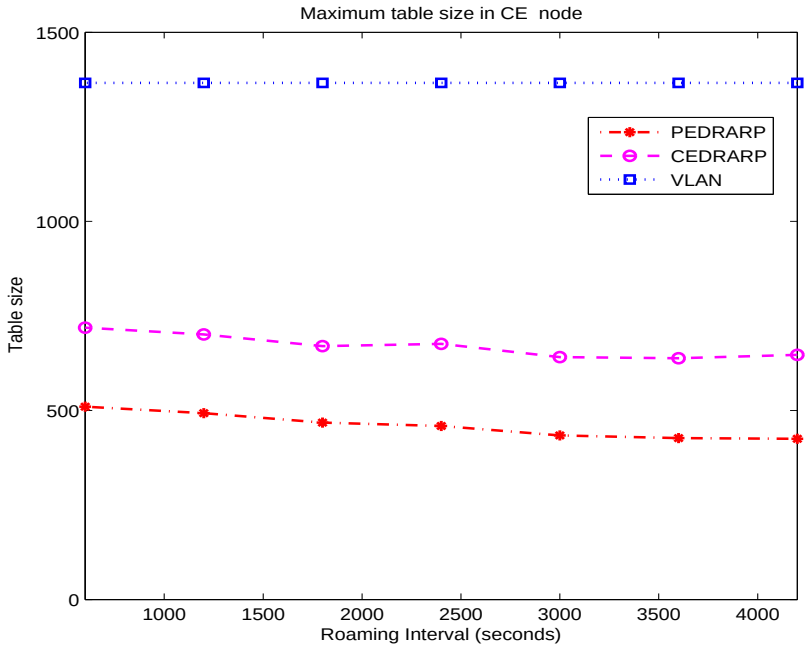


**Figure 17. Max table size in CE node**

## 4 Conclusions

High development of the Ethernet technology has made it an attractive proposition for MANs. However, Ethernet has poor scalability due to using a flat addressing scheme (i.e., non-

hierarchical MAC addresses) and broadcasting based address resolution scheme.

This paper proposes an efficient and scalable Metro-Ethernet architecture. The proposed architecture includes a Distributed Registration based Address Resolution Protocol (DRARP) and an End user enabled Mac-in-Mac (EMiM) encapsulation scheme. DRARP replaces the broadcast service with unicast service for address resolution. EMiM 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 do address swapping and hence do not need to maintain the entries mapping end user's MAC address to its PE node's MAC address, thus significantly reducing their forwarding table sizes. The simulation results show the proposed schemes can save overhead messages for address resolution and significantly reduces the forwarding table size in PE nodes.

In the future, the routing protocol, traffic control and Quality of Service (QoS) will be studied.

## References

[1] IEEE Std 802.3x-1997 and IEEE Std 802.3y-1997.
[2] J. Postel and J. Reynolds, "A Standard for the Transmission of IP Datagrams over IEEE 802 Networks", *RFC 1042*, 1988.
[3] M. Casado, M. J. Freedman and S. Shenker, "Ethane: Taking Control of the Enterprise", *ACM SIGCOM*, 2007.
[4] G. Chiruvolu, "Issues and Approaches on Extending Ethernet Beyond LANs", *IEEE Communication Magazine*, v42(3), pp 80-86, 2004.
[5] K. Elmeleegy and A. L. Cox, "EtherProxy: Scaling Ethernet By Suppressing Broadcast Traffic", *The 28th Conference on Computer Communications. IEEE*, 2009.
[6] 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.
[7] 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.
[8] R. Droms, "Dynamic host configuration protocol", *RFC 2131*, 1997.
[9] S. Halabi, *Metro Ethernet*, Cisco Press, 2003.
[10] "Problems with broadcasts", http://www.ists.dartmouth.edu/classroom/ crs/arp broadcast:php.
[11] IEEE Std. 802.1Q, "Virtual Bridged Local Area Networks".
[12] T. L. Rodeheffer, C. A. Thekkath and D. C. Anderson, "SmartBridge: a Scalable Bridge Architecture", *ACM SIGCOMM*, 2000.
[13] IEEE 802.1ah, "Provider Backbone Bridges".
[14] R. Perlman, "Rbridges: Transparent routing". *IEEE Infocom*, 2004.
[15] 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.
[16] IEEE 802.1D-2004, "Spanning Tree Protocol revision of IEEE std 802.1D".
[17] IEEE 802.1S, "Multiple Spanning Tree".
[18] IEEE 802.1W, "Rapid Reconfiguration of Spanning Tree".
[19] 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.
[20] Ibanez G., Garcia A., Azcorra A., "Alternative multiple spanning tree protocol (AMSTP) for optical Ethernet backbones" *The 29th Annual IEEE International Conference on Local Computer Networks*, 2004.
[21] "IEEE Standard for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks", *IEEE*, 1998.
[22] M. Ali, G. Chiruvolu and A. Ge, "Traffic Engineering in Metro-Ethernet", *IEEE Networks, v19(2)*, pp10-17, 2005.
[23] IEEE 802.1Q, "Virtual LANs".
[24] S. Ray, R. A. Guerin and R. Sofa, "A Distributed Hash Table Based Address Resolution Scheme for Large-Scale Ethernet Networks", *IEEE ICC*, 2007.
[25] Changhoon Kim, Rexford, J., "Revisiting Ethernet: Plug-and-play made scalable and efficient", *The 15th IEEE Workshop on Local and Metropolitan Area Networks* , 2007.
[26] Changhoon Kim, Matthew Caesar, and Jennifer Rexford, "Floodless in SEATTLE: A Scalable Ethernet Architecture for Large Enterprises", *ACM SIGCOMM*, 2008.
[27] 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.
[28] "BellSouth Metro-Ethernet", http://www.bellsouthlargebusiness.com.
[29] Huynh Minh, Mohapatra Prasant, Goose Stuart, "Cross-over spanning trees Enhancing Metro-Ethernet resilience and load balancing", *BROADNETS*, 2007.

[30] Minh Huynh, Prasant Mohapatra, Stuart Goose, "Spanning tree elevation protocol: Enhancing Metro-Ethernet performance and QoS", *Computer Communications*, 2009.
[31] Mirjalily G., Karimi M.H., Adibnia F., Rajai S, "An Approach to Select the Best Spanning Tree in Metro-Ethernet Networks", *Computer and Information Technology*, 2008.
[32] M.Tafti, G. Mirjalily, and S. Rajaee, "Topology Design of Metro-Ethernet Networks Based on Load Balance Criterion", *Internatioal Symposium on Telecommunications*, 2008.
[33] Ravi Kumar Buregoni, "A Unified Distributed Directory based Service Delivery Architecture for Metro-Ethernet Networks", *11th International Conference on Advanced Communication Technology*, 2009.
[34] X. Sun, Z. Wang, H. Che, and F. Zhao, "An End User Enabled MAC-in-MAC Encapsulation Scheme for Metro-Ethernet", *Proceedings of the 2008 IEEE International Symposium on Parallel and Distributed Processing with Applications*, 2008.
[35] Ibanez G., Garcia A., Azcorra A., I. Soto, "ABridges: Scalable, self-configuring Ethernet campus networks", *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 2008.