A Scalable and Resilient Layer-2 Network with Ethernet Compatibility

Chen Qian and Simon S. Lam
UTCS TR-13-19 October 19, 2013

***Abstract*—We present the architecture and protocols of ROME, a layer-2 network designed to be backwards compatible with Ethernet and scalable to tens of thousands of switches and millions of end hosts. Such large-scale networks are needed for emerging applications including data center networks, wide area networks, and metro Ethernet. ROME is based upon a recently developed greedy routing protocol, greedy distance vector (GDV). Protocol design innovations in ROME include a stateless multicast protocol, a Delaunay DHT, as well as routing and host discovery protocols for a hierarchical network. ROME protocols do not use broadcast and provide both control-plane and data-plane scalability. Extensive experimental results from a packet-level event-driven simulator, in which ROME protocols are implemented in detail, show that ROME protocols are efficient and scalable to metropolitan size. Furthermore, ROME protocols are highly resilient to network dynamics. The routing latency of ROME is only slightly higher than shortest-path latency. To demonstrate scalability, we provide simulation performance results for ROME networks with up to 25,000 switches and 1.25 million hosts.**

I. Introduction

Layer-2 networks, *each* scalable to tens of thousands of switches/routers and connecting millions of end hosts, are needed for important future and current applications and services including: data center networks [13], metro Ethernet [1], [4], [14], [16], wide area networks [5], [18], [15], as well as enterprise and provider networks. As an example, Google's globally-distributed database scales up to millions of machines across hundreds of data centers [9].

Ethernet offers plug-and-play functionality and a flat MAC address space. Ethernet MAC addresses, being permanent and location independent, support host mobility and facilitate management functions, such as trouble shooting and access control. For these reasons, Ethernet is easy to manage. However Ethernet is not scalable to a large network because it uses a spanning tree routing protocol that is highly inefficient and not resilient to failures. Also, after a cache miss, it relies on network-wide flooding for host discovery and packet delivery.

Today's metropolitan and wide area Ethernet services provided by network operators are based upon a network of IP (layer-3) and MPLS routers which interconnect relatively small Ethernet LANs [14]. Adding the IP layer to perform end-to-end routing in these networks nullifies Ethernet's desirable properties. IP routing protocols (such as, RIP, OSPF, and IS-IS) provide shortest paths and are much more efficient than Ethernet's spanning tree protocol. However these routing protocols still introduce *scalability problems in both control and data planes* as follows: In networks that use IP routing protocols in the control plane, routers need to either perform link-state broadcasts or exchange $O(N)$ distance vectors, both of which are not scalable. More importantly, large IP networks require massive efforts by human operators to configure and manage, especially for enterprise and data center networks where host mobility and VM migrations are ubiquitous. Networks that use shortest-path routing on flat addresses in layer 2, on the other hand, require a large amount of *data-plane state* (forwarding table entries) to reach every destination in the network. Furthermore, when multicast and VLAN are used, each switch has to store a lot more state information. Such data plane scalability is challenging because high-speed memory is both expensive and power hungry [42].

Besides scalability, resiliency is also an important requirement of large layer-2 networks. According to a recent study by Cisco [3], availability and resilience are the most important network performance metrics for distributed data processing, such as Hadoop, in large data centers. Without effective failure recovery techniques, job completion will be significantly delayed.

Therefore, it is desirable to have a *scalable and resilient layer-2 network that is backwards compatible with Ethernet*, i.e., its switches interact with hosts by Ethernet frames using conventional Ethernet format and semantics. Ethernet compatibility provides plug-and-play functionality and ease of network management. Hosts can still use IP addresses as identifiers but the network does not use IP addresses for routing.

In this paper, we present the architecture and protocols of a scalable and resilient layer-2 network, named ROME.[1] ROME is fully decentralized and self-organizing without any central controller or special nodes. All switches execute the same distributed algorithms in the control plane. *ROME uses greedy routing instead of spanning-tree or shortest-path routing to achieve scalability and resiliency.* ROME provides control-plane scalability by eliminating network broadcast and limiting control message propagation within a local area. ROME provides data-plane scalability because each switch stores small routing and multicast states.

This work was sponsored by National Science Foundation grants CNS-0830939 and CNS-1214239. An abbreviated version of this paper [33] appeared in Proceedings of IEEE ICNP, Austin, TX, November 2012.

Chen Qian is with the Department of Computer Science, University of Kentucky, Lexington, KY 40506 (e-mail: qian@cs.uky.edu). Simon S. Lam is with the Department of Computer Science, The University of Texas at Austin, Austin, TX 78712 (e-mail: lam@cs.utexas.edu).

[1] acronym for Routing On Metropolitan-scale Ethernet

ROME protocols utilize some recent advances in greedy routing, namely, GDV on VPoD [32] and MDT [23], [24]. Unlike greedy routing in wireless sensor and ad hoc networks, switches in ROME do not need any location information. For routing in ROME, a virtual space is first specified, such as, a rectangular area in 2D.[2] Each switch uses the VPoD protocol to compute a position in the virtual space such that the Euclidean distance between two switches in the space is proportional to the routing cost between them. This property enables ROME to provide routing latency only slightly higher than shortest-path latency. Switches construct and maintain a multi-hop Delaunay triangulation (MDT) which guarantees that GDV routing finds the switch closest to a given destination location [32], [24].

A. Contributions and paper outline

Protocol design innovations in this paper include the following: (i) a stateless multicast protocol to support VLAN and other multicast applications; (ii) protocols for host and service discovery using a new method, called Delaunay DHT (D^2HT); (iii) new routing and host discovery protocols for a hierarchical network.

We compare ROME with a recently proposed scalable layer-2 network, SEATTLE [20]. ROME and SEATTLE were evaluated and compared using a packet-level event-driven simulator in which ROME protocols (including GDV, MDT, and VPoD) and SEATTLE protocols are implemented in detail. Every protocol message is routed and processed by switches hop by hop from source to destination. Experimental results show that ROME performed better than SEATTLE by an order of magnitude with respect to each of the following performance metrics: switch storage, control message overhead during initialization and in steady state, and routing failure rate during network dynamics.

The routing latency of ROME is only slightly higher than the shortest-path latency. ROME protocols are highly resilient to network dynamics and switches quickly recover after a period of churn. To demonstrate scalability, we provide simulation performance results for ROME networks with up to 25,000 switches and 1.25 million hosts.

The balance of this paper is organized as follows. In Section II, we discuss related work including protocol services from our prior work used by ROME. In Section III, we present *location hashing* in a virtual space and stateless multicast. In Section IV, we present Delaunay DHT and its application to host discovery, i.e., address and location resolution. In Section V, we present ROME's architecture and routing protocols for hierarchical networks. In Section VI, we present performance evaluation and comparison of ROME and SEATTLE. We conclude in Section VII.

II. Related Work

A. Scalable Ethernet

Towards the goal of scalability, Myers et al. [30] proposed replacing Ethernet broadcast for host discovery by a layer-2 distributed directory service. In 2007, replacing Ethernet broadcast by a distributed hash table (DHT) was proposed independently by Kim and Rexford [21] and Ray et al. [37]. In 2008, Kim et al. presented SEATTLE [20] which uses link-state routing, a one-hop DHT (based on link-state routing) for host discovery, and multicast trees for broadcasting to VLANs. Scalability of SEATTLE is limited by link-state broadcast as well as a large amount of data plane state needed to reach every switch in the network [42]. In 2010, AIR [39] was proposed to replace link-state routing in SEATTLE. However, its latency was found to be larger than the latency of SEATTLE by 1.5 orders of magnitude. In 2011, VIRO [17] was proposed to replace link-state routing. To construct a rooted virtual binary tree for routing, a centralized algorithm was used for large networks (e.g., enterprise and campus networks).

To increase the throughput and scalability of Ethernet for data center networks, SPAIN [29] and PAST [41] proposed the use of many spanning trees for routing. In SPAIN, an offline network controller first pre-computes a set of paths that exploit redundancy in a given network topology. The controller then merges these paths into a set of trees and maps each tree onto a separate VLAN. SPAIN requires modification to end hosts. PAST does not requires end-host modification; instead, a spanning tree is installed in network switches for every host. The important issue of data plane scalability was not addressed in both papers.

In four of the five papers with simulation results to show network performance [20], [39], [17], [41], scalability was demonstrated for networks of several hundred switches. In SPAIN [29], simulation experiments were performed for special data center network topologies (e.g., FatTree) of up to 2,880 switches. In this paper, we demonstrate scalability of ROME from experiments that ran on a packet-level event-driven simulator for up to 25,000 switches and 1.25 million hosts.

B. Greedy routing and virtual coordinates

Many greedy geographic routing protocols have been designed for wireless sensor and ad hoc networks. Two of the earliest protocols, GFG [7] and GPSR [19], use face routing to move packets out of local minima. They require the network topology to be a planar graph in 2D to avoid routing failures. Kim et al. [22] proposed CLDP which, given any connectivity graph, produces a subgraph in which face routing would not cause routing failures. Leong et al. proposed GDSTR [26] for greedy routing without the planar graph assumption by maintaining a hull tree. Lam and Qian proposed MDT [23], [24] for any connectivity graph of nodes with arbitrary coordinates in a d-dimensional Euclidean space ($d \geq 2$). From simulation experiments in which GFG/GPSR, CLDP, GDSTR, and MDT-greedy ran on the same networks, it is shown that MDT-greedy provides the lowest routing stretch and the highest routing success rate (1.0) [24].

Many virtual coordinate schemes have been proposed for wireless networks when node location information is unavailable (e.g., [34], [11], and [8]). In each scheme, the main objective is to improve greedy routing success rate. VPoD [32] is the only virtual coordinate protocol designed to predict and minimize the routing cost between nodes.

[2] 2D, 3D, or a higher dimension can be used [24].

C. Services provided by MDT, VPoD, and GDV

ROME uses greedy routing to provide scalability and resiliency. The protocol used by ROME switches is GDV routing which uses services provided by VPoD and MDT protocols [32], [24]. We next provide a brief overview of these three protocols.

A Delaunay triangulation (DT) is a graph that can be computed from a set of node locations in a Euclidean space [12]. In a DT, two nodes sharing an edge are said to be DT neighbors. For 2D, Bose and Morin [6] proved that greedy forwarding in a DT guarantees to find the destination node. For 2D, 3D, and higher dimensional Euclidean spaces, Lee and Lam [25] generalized their result and proved that greedy forwarding in a DT guarantees to find the node closest to a destination *location*. Since two neighbors in a DT graph may not be directly-connected, nodes maintain forwarding tables for communication between DT neighbors multiple hops apart (hence the name, multi-hop DT [24]).

At network initialization, each ROME switch assigns itself a random location in the virtual space and discovers its directly-connected neighbors. Each pair of directly-connected switches exchange their unique identifiers (e.g., MAC addresses) and self-assigned locations. Then, the switches have enough information to construct and maintain a multi-hop Delaunay triangulation using MDT protocols [24].

ROME switches then repeatedly exchange messages with their neighbors, including multi-hop DT neighbors, and change their positions. Using the VPoD protocol [32], each switch moves its location in the virtual space by comparing, for each neighbor, the Euclidean distance with the routing cost between them. (Routing cost can be in any additive metric.) A switch stops running VPoD when the amount of location change has converged to less than a threshold value. When all switches finish, the distance between two switches in the virtual space approximates the routing cost between them. Then switches use their updated locations to construct a new multi-hop DT to be used by GDV routing [32].

GDV routing. Let y denote a neighbor of switch u. For a packet with destination location t, the estimated routing cost from u to t via y is $R_y = c(u,y) + \tilde{D}(y,t)$, where $c(u,y)$ is the routing cost from u to y and $\tilde{D}(y,t)$ is the distance from y to t computed by u from the locations of y and t in the virtual space. Switch u selects the neighbor v such that R_v minimizes $\{R_y, y \in$ set of directly-connected and DT neighbors of $u\}$. If $R_v < \tilde{D}(u,t)$, u sends the packet to v; else, the packet is marked. Every packet when first created is unmarked. A switch, such as u in the above example, forwards a marked packet by MDT-greedy using node locations in the virtual space without considering routing costs from the switch to its neighbors. Therefore, GDV guarantees to route every packet to the switch that is closest to the packet's destination location [24], [32].

It has been shown that the VPoD protocol is very effective such that GDV's routing cost is not much higher than that of shortest-path routing. Lastly, MDT and VPOD protocols do not use broadcast. In particular, MDT has a very efficient and effective search method for each switch to find its multi-hop DT neighbors. As a result, a multi-hop DT has been shown to be highly resilient to rapid topology changes [24], [32].

III. Routing in ROME

A. Virtual space for switches

Consider a network of switches with an arbitrary topology (any connected graph). Each switch selects one of its MAC addresses to be its identifier. End hosts are connected to switches which provide frame delivery between hosts. Ethernet frames for delivery are encapsulated in ROME packets. Switches interact with hosts by Ethernet frames using conventional Ethernet format and semantics. ROME protocols run only in switches. Link-level delivery is assumed to be reliable.

A Euclidean space (2D, 3D, or a higher dimension) is chosen as the virtual space. The number of dimensions and the minimum and maximum coordinate values of each dimension are known to all switches. Each switch determines for itself a location in the space represented by a set of coordinates.

Location hashing. To start ROME protocols, each switch boots up and assigns itself an initial location randomly by hashing its identifier, ID_S, using a globally-known hash function H. The hash value is a binary number which is converted to a set of coordinates. Our protocol implementation uses the hash function MD5 [38], which outputs a 16-byte binary value. 4 bytes are used for each dimension. Thus locations can be in 2D, 3D, or 4D.[3]

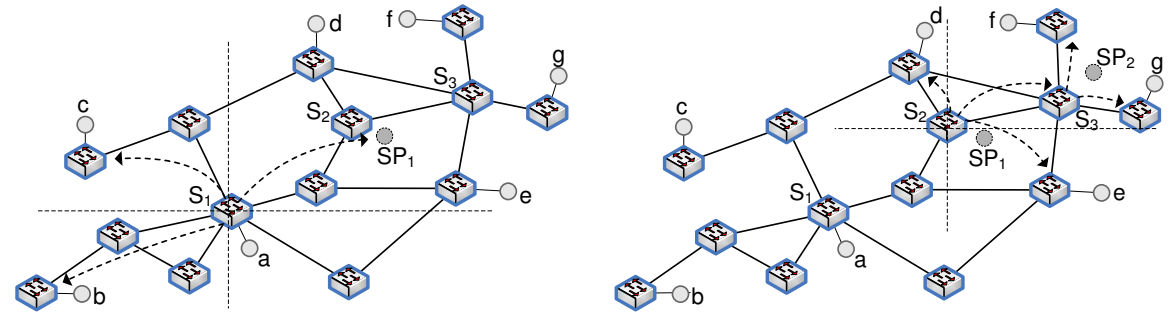
Consider, for example, a network that uses a 2D virtual space. For 2D, the last 8 bytes of $H(ID_S)$ are converted to two 4-byte binary numbers, x and y. Let MAX be the maximum 4-byte binary value, that is, $2^{32}-1$. Also let min_k and max_k be the minimum and maximum coordinate values for the kth dimension. Then the location in 2D obtained from the hash value is $(min_1 + \frac{x}{MAX}(max_1 - min_1),\ min_2 + \frac{y}{MAX}(max_2 - min_2))$, where each coordinate is a real number. The location can be stored in decimal format, using 4 bytes per dimension. Hereafter, for any identifier, ID, we will use $H(ID)$ to represent its location in the virtual space and refer to $H(ID)$ as the identifier's *location hash* or, simply, *location*.

Switches discover their directly-connected neighbors and, using their initial locations, proceed to construct a multi-hop DT [23]. Switches then update their locations using VPoD and construct a new multi-hop DT as described in subsection II-C.

Unicast routing. Unicast packet delivery in ROME is provided by GDV routing in the multi-hop DT maintained by switches. In a correct multi-hop DT, *GDV routing of a packet guarantees to find the switch that is closest to the destination location of the packet* [24], [32] assuming reliable link-level delivery and no packet drop due to congestion.

As in most prior work [20], [42], [39], [17], the issue of multi-path routing and traffic engineering is not addressed herein and will be an interesting problem for future work.

[3]Conceptually, a higher dimensional space gives VPoD more flexibility but requires more storage space and control overhead. Our experimental results show that VPoD's performance in 2D is already very good.



(a) Split at S_1

(b) Splits at S_2 and S_3

Fig. 1. Example of stateless multicast

B. Hosts

Hosts have IP and MAC addresses. Each host is directly connected to a switch called its *access switch*. An access switch knows the IP and MAC addresses of every host connected to it. The routable address of each host is the location of its access switch in the virtual space, also called the *host's location*. Hosts are not aware of ROME protocols and run ARP [31], DHCP [10], and Ethernet protocols in the same way as when they are connected to a conventional Ethernet.

C. Stateless multicast and its applications

To provide the same services as conventional Ethernet, ROME needs to support group-wide broadcast or multicast, for applications, such as, VLAN, teleconferencing, television, replicated storage/update in data centers, etc.

A straightforward way to deliver messages to a group is by using a multicast tree similar to IP multicast [20]. All broadcast packets within a group are delivered through a multicast tree sourced at a dedicated switch, namely a broadcast root, of the group. When a switch detects that one of its hosts is a member of a group, the switch joins the group's multicast tree and stores some multicast state for this group. When there are many groups with many hosts in each group, the amount of multicast state stored in switches can become a scalability problem.

We present a *stateless multicast* protocol for group-wide broadcast in ROME. A group message is delivered using the locations of its receivers without construction of any multicast tree. Switches do not store any state for delivering group messages.

The membership information of stateless multicast is maintained at a *rendezvous point* (RP) for each group. The RP of a group is determined by the location hash $H(ID_G)$, where ID_G is the group's ID. The switch whose location is closest to $H(ID_G)$ serves as the group's RP. The access switch of the sender of a group message sends the message to the RP by unicast. GDV routing guarantees to find the switch closest to $H(ID_G)$.

The RP then forwards the message to other group members (receivers) as follows: The RP partitions the entire virtual space into multiple regions. To each region with one or more receivers, the RP sends a copy of the group message with the region's receivers (their locations) in the message header (actually the ROME packet header). The destination of the group message for each region is a location, called *split position* (SP), which is either (i) the closest receiver location in that region, or (ii) the mid-point of the two closest receiver locations in the region. By GDV routing, the group message will be routed to a switch closest to the SP. This switch will in turn partition its region into multiple sub-regions and send a copy of the group message to the SP of each sub-region. Thus a multicast tree rooted at the RP grows recursively until it reaches all receivers. The tree structure is not stored anywhere. At each step of the tree growth, a switch computes SP's for the next step based on receiver locations in the group message it is to forward.

We present an example of stateless multicast in Figure 1(a). The group consists of 7 hosts a,b,c,d,e,f,g, connected to different switches with locations in a 2D virtual space as shown. Switch S_1 serves as the RP. Host a sends a message to the group by first sending it to S_1. Upon receiving the message, S_1 realizes that it is the RP. S_1 partitions the entire virtual space into four quadrants and sends a copy of the message by unicast to each of the 3 quadrants with at least one receiver. The message to the northeast quadrant with four receivers (d,e,f, and g) is sent to a split position, SP_1, which is the midpoint between the locations of d and e, the two receivers closest to S_1. The message will then be routed by GDV to S_2, the switch closest to SP_1.

Subsequently, S_2 partitions the space into four quadrants and sends a copy of the message to each of the three quadrants with one or more receivers (see Figure 1(b)). For the northeast quadrant that has two receivers, the message is sent to the split position, SP_2, which is the midpoint between the locations of f and g. The message to SP_2 will be routed by GDV to S_3, the switch closest to SP_2, which will unicast copies of the message to f and g.

At any time during the multicast, when a switch realizes that a receiver is a directly-connected host, it can transmit the message directly to the host and removes the host from the set of receivers in the message to be forwarded.

In ROME, for each group, its group membership information is stored in only one switch, the group's RP. For this group, no multicast state is stored in any other switch. This is a major step towards scalability. The tradeoff for this gain is an increase in communication overhead from storing a set of receivers in the ROME header of each group

message. Experimental results in subsection VI-F show that this communication overhead is small. This is because when the group message is forwarded by the RP and other switches, the receiver set is partitioned into smaller and smaller subsets.

The implementation of stateless multicast, as described, is not limited to the use of a 2D space. Also, partitioning of a 2D space at the RP, or at a switch closest to a SP, is not limited to four quadrants. The virtual space can be partitioned into any number of regions evenly or unevenly. A study of other virtual spaces and partitioning methods for implementing stateless multicast will be future work.

Stateless multicast for VLAN. Members of a VLAN are in a logical broadcast domain; their locations may be widely distributed in a large-scale Ethernet. ROME's stateless multicast protocol is used to support VLAN broadcast. When a switch detects that one of its hosts belongs to a VLAN, it sends a Join message to location $H(ID_V)$, where ID_V is the VLAN ID. By GDV, The Join message is routed to the switch closest to $H(ID_V)$, which is the RP of the VLAN. The RP then adds the host to the VLAN membership. The protocol for a host to leave a VLAN is similar. VLAN protocols in ROME are much more efficient than the current VLAN Trunking Protocol used in conventional Ethernet [2]. The number of global VLANs is restricted to 4094 in conventional Ethernet [14]. There is no such restriction in ROME because stateless multicast does not require switches to store VLAN information to perform forwarding.

IV. Host and Service Discovery in ROME

Suppose a host knows the IP address of a destination host from some upper-layer service. To route a packet from its source host to its destination host, switches need to know the MAC address of the destination host as well as its location, i.e., location of its access switch. Such address and location resolution are together referred to as *host discovery*.

A. Delaunay distributed hash table

The benefits of using a DHT for host discovery include the following: (i) uniformly distributing the storage cost of host information over all network switches, and (ii) enabling information retrieval by unicast rather than flooding. The one-hop DHT in SEATTLE [20] uses *consistent hashing* of identifiers into a circular location space and requires that every switch knows *all* other switches. Such global knowledge is made possible by link-state *broadcast*, which limits scalability.

In ROME, the Delaunay DHT (or D^{2}HT) uses *location hashing* of identifiers into a Euclidean space (2D, 3D, or a higher dimension) as described in subsection III-A. D^{2}HT uses greedy routing (GDV) in a multi-hop DT where every switch only needs to know its directly-connected neighbors and its neighbors in the DT graph. Furthermore, each switch uses a very efficient search method to find its multi-hop DT neighbors without broadcast [24].

In D^{2}HT, information about host i is stored as a key-value tuple, $t_i =< k_i, v_i >$, where the key k_i may be the IP (or MAC) address of i, and v_i is host information, such as its MAC address, location, etc. The access switch of host i is the *publisher* of i's tuples. A switch that stores $< k_i, v_i >$ is called a *resolver* of key k_i. The tuples are stored as *soft state*.

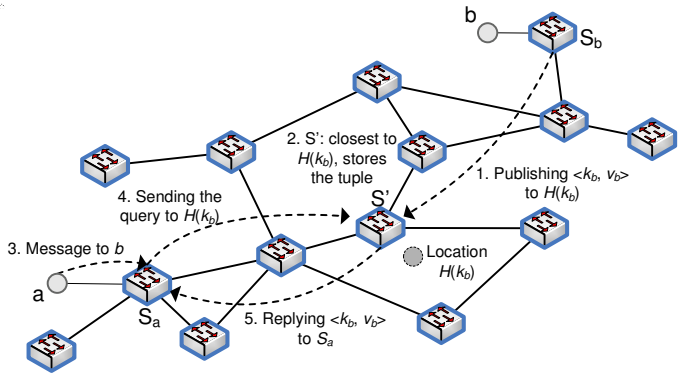


Fig. 2. S_b publishes a tuple of b. S_a performs a lookup of b

To publish a tuple, $t_i =< k_i, v_i >$, the publisher computes its location $H(k_i)$ and sends a publish message of t_i to $H(k_i)$. Location hashes are randomly distributed over the entire virtual space. It is possible but unlikely that a switch exists at the exact location $H(k_i)$. The publish message is routed by GDV to the switch whose location is closest to $H(k_i)$, which then becomes a resolver of k_i. When some other switch needs host i's information, it sends a lookup request message to location $H(k_i)$. The lookup message is routed by GDV to the resolver of k_i, which sends the tuple $< k_i, v_i >$ to the requester. A publish-lookup example is illustrated in Figure 2.

Comparison with GHT. At a high level of abstraction, D^{2}HT bears some similarity to Geographic Hash Table (GHT) [36]. However, D^{2}HT was designed for a network of switches with no physical location information. On the other hand, GHT was designed for a network of sensors in the physical world with the assumption that sensors know their geographic locations through use of GPS or some other localization technique. Also, for greedy routing, GHT uses GPSR which provides delivery of a packet to its destination under the highly restrictive assumption that the network connectivity graph can be planarized [19]. Thus protocols of D^{2}HT and GHT are very different and the network environments of their intended applications are also different.

Comparison with CAN. Both D^{2}HT and Content Addressable Network (CAN) [35] are DHTs and both use a d-dimensional virtual space. However, this is the extent of their similarity. D^{2}HT and CAN are very different in design. In CAN, the entire virtual space is dynamically partitioned into zones each of which is owned by a node. Nodes in a CAN self-organize into an overlay network that depends on the underlying IP network for packet delivery. D^{2}HT, on the other hand, is designed for a layer-2 network without IP routing. D^{2}HT does not have the concepts of zone and zone ownership. Instead, switches find their locations in a virtual space using location hashing described in Section III-A. Each switch in D^{2}HT only knows its directly-connected neighbors and DT neighbors and their virtual locations.

B. Host discovery using D^{2}HT

In ROME, the routable address of host i is i's location c_i, which is the location of its access switch. There are two key-value tuples for each host, for its IP-to-MAC and MAC-to-location mappings.

In a tuple for host i, the key k_i may be its IP or MAC address. If k_i is the MAC address, value v_i includes location c_i and the unique ID, S_i, of i's access switch. If k_i is the IP address, the value v_i includes the MAC address, MAC_i, as well as c_i and S_i. Note that the host location is included in both tuples for each host.

After a host i is plugged into its access switch S_i with location c_i, the switch learns the host's IP and MAC addresses, IP_i and MAC_i, respectively. S_i then constructs two tuples: $< MAC_i, c_i, S_i >$ and $< IP_i, MAC_i, c_i, S_i >$, and stores them in local memory. S_i then sends publish messages of the two tuples to $H(IP_i)$ and $H(MAC_i)$.

Note that each switch stores two kinds of tuples. For a tuple with key k_i stored by switch S, if S is i's access switch, the tuple is a *local tuple* of S. Otherwise, the tuple is published by another switch and is an *external tuple* of S. Switches store key-value tuples as *soft state*.

Each switch interacts with directly-connected hosts using frames with conventional Ethernet format and semantics. When a host j sends its access switch S_j an ARP query frame with destination IP address IP_i and the broadcast MAC address, S_j sends a lookup request to location $H(IP_i)$, which is routed by GDV to a resolver of IP_i. The resolver sends back to S_j the tuple $< IP_i, MAC_i, c_i, S_i >$. After receiving the tuple, the access switch S_j caches the tuple and transmits a conventional ARP reply frame to host j. When j sends an Ethernet frame with destination MAC_i, the access switch S_j retrieves location c_i from its local memory and sends the Ethernet frame to c_i. If S_j cannot find the location of MAC_i in its local memory because, for instance, the cached tuple has been overwritten, it sends a lookup request which is routed by GDV to $H(MAC_i)$ to get the MAC-to-location mapping of host i.

All publish and lookup messages are unicast messages. Host discovery in ROME is accomplished *on demand* and is *flooding-free*.

C. Reducing lookup latency

We designed and evaluated several techniques to speed up key-value lookup for host discovery, namely: (i) using multiple independent hash functions to publish each key-value tuple at multiple locations, (ii) hashing to a smaller region in the virtual space, (iii) caching key-value tuples for popular hosts as well as other shortcuts for faster responses. These latency reduction techniques are described in more detail in the Appendix.

D. Maintaining consistent key-value tuples

A key-value tuple $< k_i, v_i >$ stored as an external tuple in a switch is *consistent* iff (i) the switch is closest to the location $H(k_i)$ among all switches in the virtual space, and (ii) c_i is the correct location of i's access switch. At any time, some key-value tuples may become inconsistent as a result of host or network dynamics.

Host dynamics. A host may change its IP or MAC address, or both. A host may change its access switch, such as, when a mobile node moves to a new physical location or a virtual machine migrates to a new system.

Network dynamics. These include the addition of new switches or links to the network as well as deletion/failure of existing switches and links. MDT and VPoD protocols have been shown to be highly resilient to network dynamics (churn) [23], [32]. Switch states of the multi-hop DT as well as switch locations in the virtual space recover quickly to correct values after churn. The following discussion is limited to how host and network dynamics are handled by switches in the role of publisher and in the role of resolver in D^{2}HT.

As a publisher, each switch ensures that local tuples of its hosts are correct when there are host dynamics. For example, if a host has changed its IP or MAC address, the host's tuples are updated accordingly. If a new host is plugged into the switch, it creates tuples for the new host. New as well as updated tuples are published to the network. In addition to these reactions to host dynamics, switches also periodically refresh tuples they previously published. For every local tuple $< k_i, v_i >$, S sends a refresh message every T_r second to its location $H(k_i)$. The purpose of a refresh message is twofold: (i) If the switch closest to location $H(k_i)$ is the current resolver, timer of the soft-state tuple in the resolver is refreshed. (ii) If the switch closest to $H(k_i)$ is different from the current resolver, the refresh message notifies the switch to become a resolver.

As a resolver, each switch sets a timer for every external tuple stored in local memory. The timer is reset by a request or refresh message for the tuple. If a timer has not been reset for T_e time, timeout occurs and the tuple will be deleted by the resolver. T_e is set to a value several times that of T_r.

For faster recovery from network dynamics, we designed and implemented a technique, called *external tuple handoff*. When a switch detects topology or location changes in the multi-hop DT, it checks the location $H(k_i)$ of every external tuple $< k_i, v_i >$. If the switch finds a physical or DT neighbor closer to $H(k_i)$ than itself, it sends a *handoff message* including the tuple to the closer neighbor. The handoff message will be forwarded by GDV until it reaches the switch closest to $H(k_i)$, which then becomes the tuple's new resolver.

E. DHCP server discovery using D^{2}HT

In a conventional Ethernet, a new host broadcasts a Dynamic Host Configuration Protocol (DHCP) discover message to find a DHCP server. Each DHCP server that has received the discover message allocates an IP address and broadcasts a DHCP offer message to the host. The host broadcasts a DHCP request to accept an offer. The selected server broadcasts a DHCP ACK message. Other DHCP servers, if any, withdraw their offers.

In ROME, the access switch of each DHCP server publishes the server's information to a location using a key known by all switches, such as, "DHCPSERVER1". When the access switch of a host receives a DHCP discover message, the message is routed by GDV to the location of a DHCP server, without

use of flooding. There is no duplicate DHCP offer. To be compatible with a conventional Ethernet, the access switch replies to the host with a DHCP offer and later transmits a DHCP ACK in response to the host's DHCP request.

V. ROME for a Hierarchical Network

A metropolitan or wide area Ethernet spanning across a large geographic area typically has a hierarchical structure comprising many *access networks* interconnected by a *core network* [16]. Each access network has one or more *border switches*. The border switches of all access networks form the core network. Consider a hierarchical network consisting of 500 access networks each of which has 2000 switches. The total number of switches is 1 million. At 100 hosts per switch, the total number of hosts is 100 millions. We believe that a 2-level hierarchy is adequate for metropolitan scale in the foreseeable future.

A. Routing in a hierarchical network

For hierarchical routing in ROME, separate virtual spaces are specified for the core network and each of the access networks, called *regions*. Every switch knows the virtual space of its region (i.e., dimensionality as well as maximum and minimum coordinate values of each dimension). Every border switch knows two virtual spaces, the virtual space of its region and the virtual space of the core network, called *backbone*.

The switches in a region first discover their directly-connected neighbors. They then use MDT and VPoD protocols to determine their locations in the region's virtual space (*regional locations*) and construct a multi-hop DT for the access network. Similarly, the border switches use MDT and VPoD protocols to determine their locations in the virtual space of the backbone (*backbone locations*) and construct a multi-hop DT for the core network. Each border switch sends its information (unique ID, regional and backbone locations) to all switches in its region.

The Delaunay DHT requires the following extension for hierarchical routing: Each key-value tuple $< k_i, v_i >$ of host i stored at a resolver includes additional information, B_i, which specifies the IDs and backbone locations of the border switches in host i's region.

When a host sends an Ethernet frame to another host, its access switch obtains, from its cache or using host discovery, the destination host's key-value tuple, which includes border switch information of the destination region. This information allows the access switch to determine whether to route the frame to its destination using *intra-region* routing or *inter-region* routing.

Intra-region routing. The sender's access switch indicates in the ROME packet header that this is an intra-region packet. The routable address is the regional location of the access switch of the receiver. The packet will be routed by GDV to the access switch of the receiver as previously described. In the example of Figure 3, an intra-region packet is routed by GDV from access switch S_1 to destination host's access switch S_2 in the same regional virtual space.

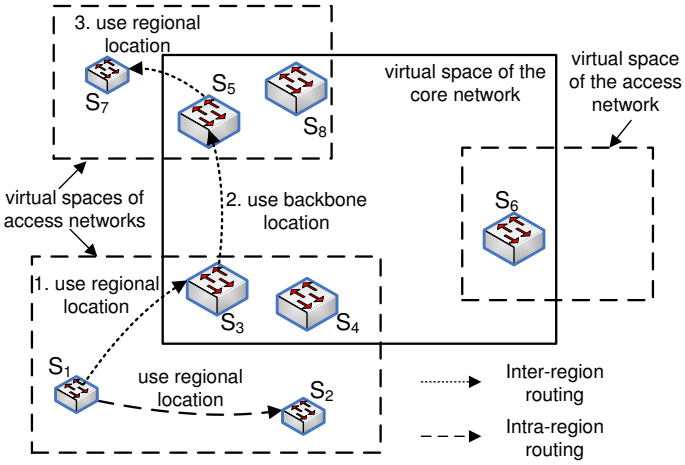


Fig. 3. Routing in a hierarchical network

Inter-region routing. For a destination host in a different region, an access switch learns, from the host's key-value tuple, information about the host's border switches and their backbone locations. This information is included in the ROME header encapsulating every Ethernet frame destined for that host. We describe inter-region routing of a ROME packet as illustrated in Figure 3. The origin switch S_1 computes its distances in the regional virtual space to the region's border switches, S_3 and S_4. S_1 chooses S_3 which is closer to S_1 than S_4. The packet is routed by GDV to S_3 in the regional virtual space.

S_3 learns from the ROME packet header, S_5 and S_8, border switches in the destination's region. S_3 computes their distances to destination S_7 in the destination region's virtual space. S_3 chooses S_5 because it is closer to the destination location. The packet is then routed by GDV in the backbone virtual space to S_5. Lastly, the packet is routed, in the destination region's virtual space, by GDV from S_5 to S_7, which extracts the Ethernet frame from the ROME packet and transmits the frame to the destination host.

Note that at the border switch S_3, it has a *choice of minimizing the distance traveled by the ROME packet in the backbone virtual space or in the destination region's virtual space*. In our current ROME implementation, the distance in the destination region's virtual space is minimized. This is based upon our current assumption that the number of switches in an access network is larger than the number of switches in the core network. This choice at a border switch is programmable and can be easily reversed. Lastly, it is not advisable to use the sum of distances in two different virtual spaces (specified independently) to determine routing because they are not comparable. This restriction may be relaxed but it is beyond the scope of this paper.

B. Host discovery in a hierarchical network

As illustrated in Figure 4, the key-value tuple $< k_i, v_i >$ of host i is published to two resolvers in the entire network, namely: a *regional resolver* and a *global resolver*. The regional resolver is the switch closest to location $H(k_i)$ in the same region as host i; it is labeled by S_{r1} in the figure. The publish

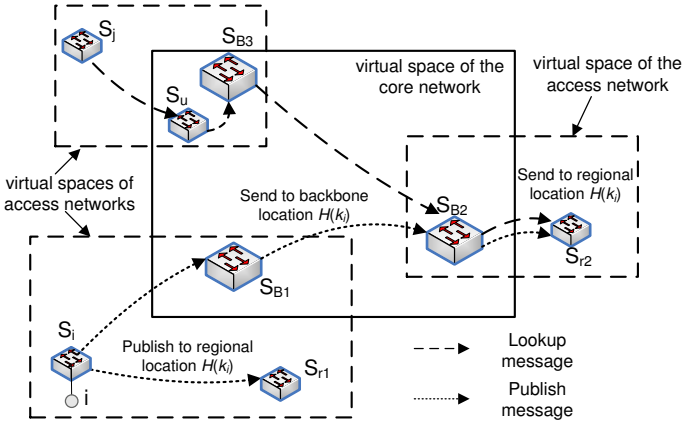


Fig. 4. Tuple publishing and lookup in a hierarchical Ethernet

and lookup protocols are the same as the ones presented in subsection IV-B. To find a tuple with key k_i, a switch sends a lookup message to position $H(k_i)$ in its own region. A regional resolver provides fast responses to queries needed for intra-region communications.

Publish to a global resolver. Switches outside of host i's region cannot find its regional resolver. Therefore, the key-value tuple $< k_i, v_i >$ of host i is also stored in a global resolver to respond to host discovery for inter-region communications. The global resolver can be found by any switch in the entire network. As shown in Figure 4, to publish a tuple $< k_i, v_i >$ to its global resolver, the publish message is first routed by GDV to the regional location of one of the border switches in the region, labeled by S_{B1} in the figure. S_{B1} computes location $H(k_i)$ in the backbone virtual space and includes it with the publish message which is routed by GDV to the border switch closest to backbone location $H(k_i)$ in the core network, labeled by S_{B2} in the figure.

Switch S_{B2} serves as the global resolver of host i if it has enough memory space. Switch S_{B2} can optionally send the tuple to a switch in its region such that all switches in the region share the storage cost of the global resolver function (called *two-level location hashing*). In two-level location hashing, the publish message of tuple $< k_i, v_i >$ sent by S_{B2} is routed by GDV to a switch closest to the regional location $H(k_i)$ (labeled by S_{r2} in the figure) inside S_{B2}'s access network. S_{r2} then becomes a global resolver of host i.

Lookup in a hierarchical network. To discover the key-value tuple $< k_i, v_i >$ of host i, a switch S_j first sends a lookup message to location $H(k_i)$ in its region. As illustrated in Figure 4 (upper left), the lookup message arrives at a switch S_u closest to $H(k_i)$. If S_j and host i were in the same region, S_u would be the regional resolver of i and it would reply to S_j with the key-value tuple of host i. Given that S_j and host i are in different regions, it is very unlikely that S_u happens to be a global resolver of host i (however the probability is nonzero). If S_u cannot find host i's tuple in its local memory, it forwards the lookup message to one of the border switches in its region, S_{B3} in Figure 4. Then S_{B3} computes location $H(k_i)$ in the backbone virtual space and includes it with the lookup message, which is routed by GDV to the border switch S_{B2} closest to $H(k_i)$.

In the scenario illustrated in Figure 4, S_{B2} is not host i's global resolver and it forwards the lookup message to switch S_{r2} closest to the regional location $H(k_i)$, which is the global resolver of host i.

Hash functions. In the above examples, the core and access networks use different virtual spaces but they all use the same hash function H. We note that different hash functions can be used in different networks. It is sufficient that all switches in the same network (access or core) agree on the same hash function, just like they must agree on the same virtual space.

VI. Performance Evaluation

A. Methodology

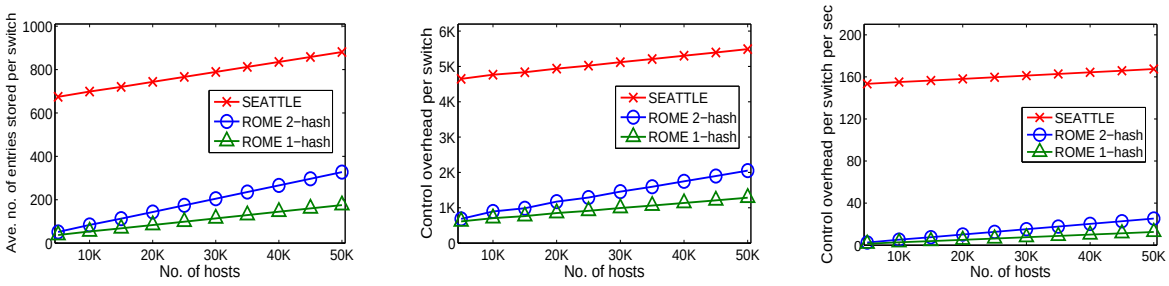
The ROME architecture and protocols were designed with the objectives of *scalability*, *efficiency*, and *reliability*. ROME was evaluated using a packet-level event-driven simulator in which ROME protocols as well as the protocols, GDV, VPoD, and MDT [23], [32] used by ROME are implemented in detail. Every protocol message is routed and processed by switches hop by hop from source to destination. Since our focus is on routing protocol design, queueing delays at switches were not simulated. Packet delays from one switch to another on an Ethernet link are sampled from a uniform distribution in the interval $[50\ \mu s, 150\ \mu s]$ with an average value of 100 μs. This abstraction speeds up simulation runs and allows performance evaluation and comparison of routing protocols unaffected by congestion issues. The same abstraction was used in the packet-level simulator of SEATTLE [20].

For comparison with ROME, we implemented SEATTLE protocols in detail in our simulator. We conducted extensive simulations to evaluate ROME and SEATTLE in large networks and dynamic networks with reproducible topologies. For the link-state protocol used by SEATTLE, we use OSPF [28] in our simulator. The default OSPF link state broadcast frequency is once every 30 seconds. Therefore, in ROME, each switch runs the MDT maintenance protocol once every 30 seconds.

In ROME, a host's key-value tuple may be published using one location hash or two location hashes. In the case of publishing two location hashes for each tuple, the area of the second hash region is 1/4 of the entire virtual space.

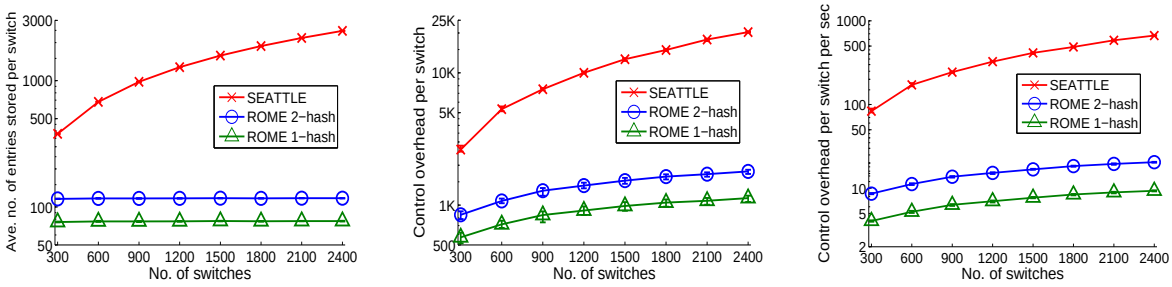
Performance criteria. *Storage cost* is measured by the *average number of entries stored per switch*. These entries include forwarding table entries and host information entries (key-value tuples).

Control overhead is communication cost measured by the *average number of control message transmissions*, for three cases: (i) network initialization, (ii) network in steady state, and (iii) network under churn. Control overhead of ROME for initialization includes those used by switches to determine virtual locations using VPoD, construct a multi-hop DT using MDT protocols, and populate the D^2HT with host information for all hosts. Control overhead of SEATTLE for initialization includes those used by switches for link-state broadcast and to populate the one-hop DHT with host information for all hosts. During steady state (also during churn), switches in SEATTLE and ROME use control messages (i) to detect inconsistencies in forwarding tables and key-value tuples stored locally and



(a) Storage cost (b) Control overhead for initialization (c) Control overhead in steady state

Fig. 5. Performance comparison by varying the number of hosts



(a) Storage cost (b) Control overhead for initialization (c) Control overhead in steady state

Fig. 6. Performance comparison by varying the number of switches

externally, as well as (ii) to repair inconsistencies in forwarding tables and key-value tuples.

We measure two kinds of *latencies to deliver ROME packets*: (i) latency of the first packet to an unknown host, which includes the latency for host discovery, and (ii) latency of a packet to a discovered host.

To evaluate ROME's (also SEATTLE's) resilience under churn, we show the *routing failure rates* of first packets to unknown hosts and packets to discovered hosts. Successful routing of the first packet to an unknown host requires successful host discovery as well as successful packet delivery by switches from source to destination.

Network topologies used. The first set of experiments used the AS-6461 topology with 654 routers from Rocketfuel data [40] where each router is modeled as a switch. To evaluate the performance of ROME as the number of switches increases, synthetic topologies generated by BRITE with the Waxman model [27] at the router level were used. Every data point plotted in Figures 6, 7, and 9 is the average of 20 runs from different topologies generated by BRITE. *Upper and lower bars in the figure show maximum and minimum values of each data point* (these bars are omitted in Figure 7(c) for clarity). Most of the differences between maximum and minimum values in these figures are very small (many not noticeable) with the exception of latency values in Figures 7(a) and (b).

B. Varying the number of hosts

For a network with n switches and m hosts, a conventional Ethernet requires $O(nm)$ storage per switch while SEATTLE requires $O(m)$ storage per switch. We found that ROME also requires $O(m)$ storage per switch with a much smaller absolute value than that of SEATTLE. We performed simulation experiments for a fixed topology (AS-6461) with 654 switches. The number of hosts at each switch varies. The total number of hosts of the entire network varies from 5,000 to 50,000. We found that the storage costs of ROME and SEATTLE for forwarding tables are constant, while their storage costs for host information increase linearly as the number of hosts increases. In Figure 5(a), the difference between the storage costs of ROME and SEATTLE is the difference in their forwarding table storage costs per switch. The host information storage cost of ROME using two (location) hashes is close to, but not larger than, twice the storage cost of ROME using one hash.

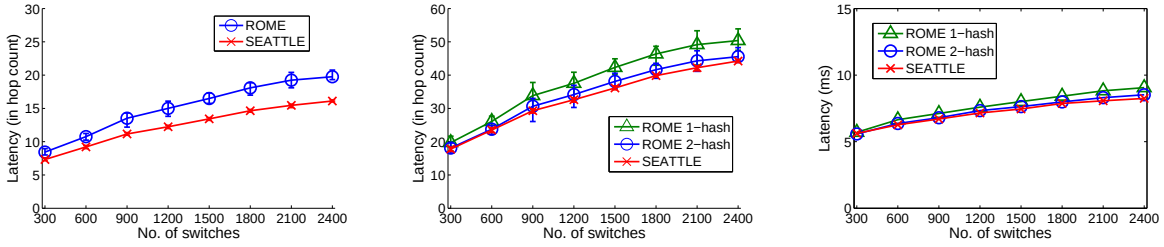
Figures 5(b) and 5(c) show the control overheads of ROME and SEATTLE, for initialization and in steady state. We found that the control overheads for constructing and updating SEATTLE's one-hop DHT and ROME's D^{2}HT both increase linearly with m and they are about the same. However, the figures show that ROME's overall control overhead is much smaller than that of SEATTLE. This is because ROME's forwarding table construction and maintenance are flooding-free and thus much more efficient.

C. Varying the number of switches

In this set of experiments the number n of switches increases from 300 to 2,400 while the average number of hosts per switch is fixed at 20. Thus the total number of hosts of the network also increases linearly from 6,000 to 48,000. The results are shown in Figure 6. Note that each y-axis is in logarithmic scale.

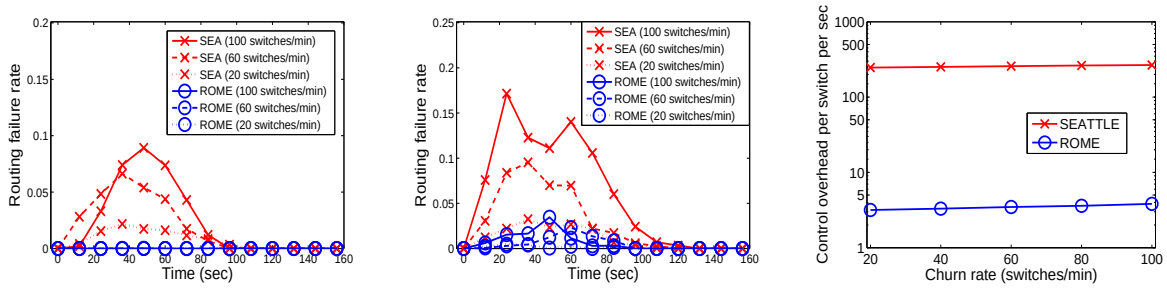
Figure 6(a) shows storage cost versus n. Note that while the storage cost of SEATTLE increases with n, ROME's storage cost is almost flat versus n. At $n = 2400$, ROME's storage cost is less than 1/20 of the storage of SEATTLE.

Figures 6(b) and (c) show that the control overheads of ROME for initialization and in steady state are both substantially lower than those of SEATTLE. These control overheads



(a) Packet to a discovered host (b) 1st packet to an unknown host (c) 1st packet to unknown host in hybrid nets

Fig. 7. Latency vs. number of switches



(a) Routing failure rate to a discovered host (b) Routing failure rate to an unknown host (c) Control overhead

Fig. 8. Performance under network dynamics

of ROME increase slightly with n. This is because the paths from publishers to resolvers in a larger network are longer.

D. Routing latencies

These experiments were performed using the same network topologies (with 20 hosts per switch on average) as in subsection VI-C. Figure 7(a) shows the latency (in average number of hops) of packets to discovered hosts. Note that ROME's latency is not much higher than the shortest-path latency of SEATTLE.

Figure 7(b) shows the latency of first packets to unknown hosts for SEATTLE and for ROME using one and two hashes. This latency includes the round-trip delay between sender and resolver, and the subsequent latency from sender to destination. By using two hashes instead of one, the latency of ROME improves and becomes very close to the latency of SEATTLE. At $n = 300$, the latency of ROME (2-hash) is actually smaller than the latency of SEATTLE.

We also performed experiments to evaluate ROME and SEATTLE latencies in hybrid networks, where 20% of the switches are replaced by wireless switches. The packet delay of a wireless hop is sampled uniformly from $[5\ ms, 15\ ms]$ with an average value of 10 *ms*, much higher than 100 μs for a wired connection. Figure 7(c) shows that SEATTLE still has the lowest latency, but the difference between SEATTLE and ROME is negligible.

E. Resilience to network dynamics

We performed experiments to evaluate the resilience of ROME using two hashes and SEATTLE under network dynamics for networks with 1,000 switches and 20,000 hosts. Before starting each experiment, consistent forwarding tables and DHTs were first constructed. During the period of 0-60 seconds, new switches joined the network and existing switches failed. The rate at which switches join, equal to the rate at which switches fail, is called the churn rate. Figure 8(a) shows the routing failure rates to discovered hosts as a function of time for ROME and SEATTLE. Different curves correspond to churn rates of 20, 60, and 100 switches per minute. At these very high churn rates, the routing failure rate of ROME is close to zero. The routing failure rate of SEATTLE is relatively high but it converged to zero after 100 seconds (40 seconds after churn stopped).

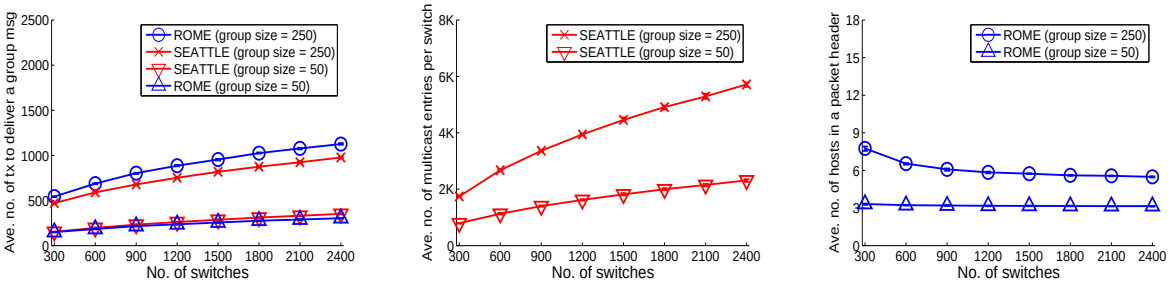
Figure 8(b) shows routing failure rates to unknown hosts versus time. Both SEATTLE and ROME experienced many more routing failures which include host discovery failures. The routing failure rate of ROME at the churn rate of 100 switches/minute is still less than that of SEATTLE at the churn rate of 20 switches/minute.

Figure 8(c) shows the control overhead (per switch per second) during a churn and recovery period versus churn rate during the period. The control overhead of SEATTLE is very high due to link-state broadcast. The control overhead of ROME is about two orders of magnitude smaller than that of SEATTLE.

ROME has much smaller routing failure rates and control overhead because each switch (using the MDT maintenance protocol) can find all its neighbors in the multi-hop DT of switches very efficiently without broadcast.

F. Performance of multicast

Both SEATTLE and ROME provide multicast support for services like VLAN. SEATTLE uses a multicast tree for each group which requires switches in the tree to store some multicast state. ROME uses the stateless multicast protocol described in subsection III-C. We performed experiments using the same network topologies (with 20 hosts per switch on average) as in subsection VI-C. The average multicast group size is 50 or 250 in an experiment. The number of groups is 1/10 of the number of hosts.



(a) Ave. no. of transmissions to deliver a group message (b) Multicast storage cost per switch of SEATTLE (c) Ave. no. of destinations in the header of a ROME group message

Fig. 9. Performance of multicast

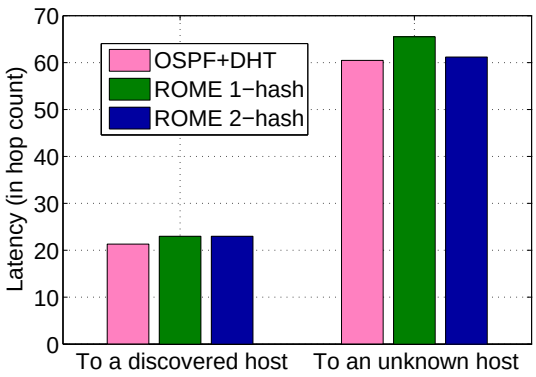


Fig. 10. Latency comparison for a very large hierarchical network (25,000 switches)

Figure 9(a) shows the average number of transmissions used to deliver a group message versus the number n of switches. For multicast using a tree, this is equal to the number of links in the tree. SEATTLE used few transmissions than ROME in experiments for average group size 250. ROME used fewer transmissions in experiments for average group size 50.

Figure 9(b) shows the amount of multicast state (average number of groups) per switch in SEATTLE versus n, the number of switches. (ROME's multicast is stateless.) Each switch in SEATTLE stores multicast state for a large number of groups, i.e., thousands in these experiments. (Group membership information stored at rendezvous points is not included because it is needed by both ROME and SEATTLE.) On the other hand, ROME requires the packet header of each group message to store a subset of hosts in the group. (SEATTLE does not have this overhead.) Figure 9(c) shows the average number of hosts in a ROME packet header. For experiments in which average group size is 50, the number is around 3. For experiments in which average group size is 250, the number is about 6.

G. Performance of a very large hierarchical network

We use a hierarchical network consisting of 25 access networks of 1000 switches each (generated by BRITE at router level). Two switches in each access network serve as border switches in a backbone network of 50 switches with topology generated by Brite at AS level. Kim et al. [20] discussed ideas for a multi-level one-hop DHT. Based upon the discussion, we implemented in our packet-level event-driven simulator an extension to SEATTLE for routing in a hierarchical network, which we refer to as "OSPF+DHT".

We performed experiments for this network of 25,000 switches for 250K to 1.25 million hosts. Figure 10 shows the routing latencies for ROME and OSPF+DHT. ROME's latency to a discovered host is very close to the shortest-path latency of OSPF+DHT, much closer than the latencies in single-region experiments shown in Figure 7(a). ROME's latency to an unknown host is also very close to the shortest-path latency of OSPF+DHT. Figure 11 shows the storage cost per switch, control overheads for initialization and in steady state. The performance of ROME is about an order of magnitude better than the OSPF+DHT approach.

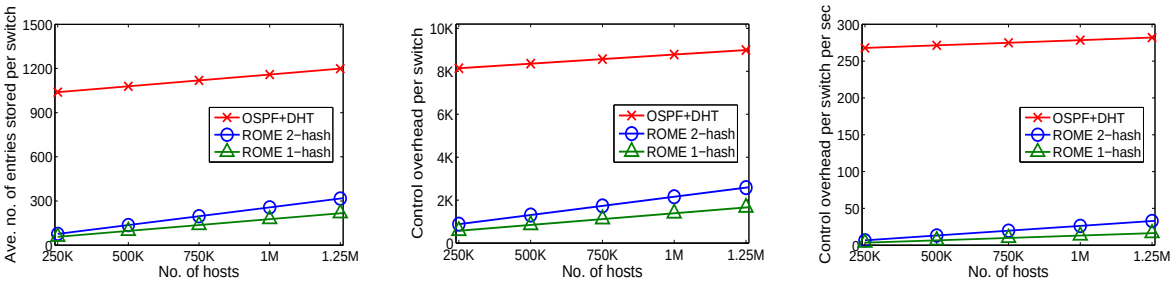
VII. Conclusions

We present the architecture and protocols of ROME, a scalable and resilient layer-2 network that is backwards compatible with Ethernet. Our protocol design innovations include a stateless multicast protocol, a Delaunay DHT (D^2HT), as well as routing and host discovery protocols for a hierarchical network. Experimental results using both real and synthetic network topologies show that ROME protocols are efficient and scalable. ROME protocols are highly resilient to network dynamics and its switches quickly recover after a period of churn. The routing latency of ROME is only slightly higher than the shortest-path latency.

Experimental results show that ROME performs better than SEATTLE by an order of magnitude with respect to each of the following performance metrics: switch storage, control message overhead during initialization and in steady state, and routing failure rate during network dynamics. To demonstrate scalability, we provide simulation performance results for ROME networks with up to 25,000 switches and 1.25 million hosts.

References

[1] Metro ethernet. http://metro-ethernet.org/.
[2] Understanding VLAN Trunk Protocol (VTP). Cisco Technical Support & Documentation, 2007.
[3] Big Data in the Enterprise: Network Design Considerations. Cisco White Paper, 2011.
[4] AT&T. Metro ethernet service. http://www.business.att.com/enterprise/Service/network-services/ethernet/metro-gigabit/.
[5] AT&T. Wide area ethernet. http://www.business.att.com/enterprise/Service/network-services/ethernet/wide-area-vpls/.
[6] P. Bose and P. Morin. Online routing in triangulations. *SIAM journal on computing*, 33(4):937–951, 2004.



(a) Storage cost (b) Control overhead for initialization (c) Control overhead in steady state

Fig. 11. Performance comparison for a very large hierarchical network (25,000 switches)

[7] P. Bose, P. Morin, I. Stojmenovic, and J. Urrutia. Routing with Guaranteed Delivery in Ad Hoc Wireless Networks. In *Proc. of the International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications (DIALM)*, 1999.

[8] A. Caruso, S. Chessa, S. De, and R. Urpi. GPS Free Coordinate Assignment and Routing in Wireless Sensor Networks. In *Proceedings of IEEE INFOCOM*, pages 150–160, 2005.

[9] J. C. Corbett et al. Spanner: Google's Globally-Distributed Database. In *Proceedings of USENIX OSDI*, 2012.

[10] R. Droms. Dynamic Host Configuration Protocol. RFC 2131, 1997.

[11] R. Fonseca, S. Ratnasamy, J. Zhao, C. T. Ee, D. Culler, S. Shenker, and I. Stoica. Beacon-Vector Routing: Scalable Point-to-Point Routing in Wireless Sensor Networks. In *Proc. of NSDI*, 2005.

[12] S. Fortune. Voronoi diagrams and Delaunay triangulations. In J. E. Goodman and J. O'Rourke, editors, *Handbook of Discrete and Computational Geometry*. CRC Press, second edition, 2004.

[13] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta. VL2: a scalable and flexible data center network. In *Proceedings of ACM SIGCOMM*, 2009.

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

[15] C.-Y. Hong et al. Achieving High Utilization with Software-Driven WAN. In *Proceedings of ACM Sigcomm*, 2013.

[16] M. Huynh and P. Mohapatra. Metropolitan Ethernet Network: A Move from LAN to MAN. *Computer Networks*, 51, 2007.

[17] S. Jain, Y. Chen, S. Jain, and Z.-L. Zhang. VIRO: A Scalable, Robust and Name-space Independent Virtual Id ROuting for Future Networks. In *Proc. of IEEE INFOCOM*, 2011.

[18] S. Jain et al. B4: Experience with a Globally-Deployed Software Defined WAN. In *Proceedings of ACM Sigcomm*, 2013.

[19] B. Karp and H. Kung. Greedy Perimeter Stateless Routing for Wireless Networks. In *Proceedings of ACM Mobicom*, 2000.

[20] C. Kim, M. Caesar, and J. Rexford. Floodless in SEATTLE: A Scalable Ethernet Architecture for Large Enterprises. In *Proc. of Sigcomm*, 2008.

[21] C. Kim and J. Rexford. Revisiting Ethernet: Plug-and-play made scalable and efficient. In *Proceedings of IEEE LAN/MAN Workshop*, May 2007.

[22] Y.-J. Kim, R. Govindan, B. Karp, and S. Shenker. Geographic Routing Made Practical. In *Proceedings of USENIX NSDI*, 2005.

[23] S. S. Lam and C. Qian. Geographic Routing in *d*-dimensional Spaces with Guaranteed Delivery and Low Stretch. In *Proceedings of ACM SIGMETRICS*, June 2011.

[24] S. S. Lam and C. Qian. Geographic Routing in *d*-dimensional Spaces with Guaranteed Delivery and Low Stretch. In *IEEE/ACM Transactions on Networking*, volume 21, pages 663–677, 2013.

[25] D.-Y. Lee and S. S. Lam. Protocol design for dynamic Delaunay triangulation. Technical Report TR-06-48, The Univ. of Texas at Austin, Dept. of Computer Science, December 2006; an abbreviated version in *Proceedings IEEE ICDCS*, June 2007.

[26] B. Leong, B. Liskov, and R. Morris. Geographic Routing without Planarization. In *Proceedings of USENIX NSDI*, 2006.

[27] A. Medina, A. Lakhina, I. Matta, , and J. Byers. BRITE: An Approach to Universal Topology Generation. In *Int. Workshop on Modeling, Analysis and Simulation of Computer and Telecom. Systems*, 2001.

[28] J. Moy. OSPF Version 2. RFC 2328, 1998.

[29] J. Mudigonda, P. Yalagandula, M. Al-Fares, and J. C. Mogul. SPAIN: COTS data-center Ethernet for multipathing over arbitrary topologies. In *Proceedings of USENIX NSDI*, 2010.

[30] A. Myers, T. E. Ng, and H. Zhang. Rethinking the service model: Scaling ethernet to a million nodes. In *Proceedings of HotNets*, 2004.

[31] D. Plummer. An Ethernet Address Resolution Protocol. RFC 826, 1982.

[32] C. Qian and S. S. Lam. Greedy Distance Vector Routing. In *Proceedings of IEEE ICDCS*, June 2011.

[33] C. Qian and S. S. Lam. ROME: Routing On Metropolitan-scale Ethernet. In *Proceedings of IEEE ICNP*, 2012.

[34] A. Rao, S. Ratnasamy, C. Papadimitriou, S. Shenker, and I. Stoica. Geographic Routing without Location Information. In *Proceedings of ACM Mobicom*, 2003.

[35] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. A Scalable Content-Addressable Network. In *Proceedings of ACM SIGCOMM*, 2001.

[36] S. Ratnasamy, B. Karp, L. Yin, F. Yu, D. Estrin, R. Govindan, and S. Shenker. GHT: a geographic hash table for data-centric storage. In *Proceedings of ACM WSNA*, 2002.

[37] S. Ray, R. Guerin, and R. Sofia. A distributed hash table based address resolution scheme for large-scale Ethernet networks. In *Proceedings of Int. Conf. on Communications*, June 2007.

[38] R. Rivest. The MD5 Message-Digest Algorithm. RFC 1321, 1992.

[39] D. Sampath, S. Agarwal, and J. Garcia-Luna-Aceves. Ethernet on AIR: Scalable Routing in Very Large Ethernet-based Networks. In *Proceedings of IEEE ICDCS*, 2010.

[40] N. Spring, R. Mahajan, and D. Wetherall. Measuring ISP Topologies with Rocketfuel. In *Proceedings of ACM SIGCOMM*, 2002.

[41] B. Stephens, A. Cox, W. Felter, C. Dixon, and J. Carter. PAST: Scalable Ethernet for Data Centers. In *Proceedings of ACM CoNEXT*, 2012.

[42] M. Yu, A. Fabrikant, and J. Rexford. Buffalo: Bloom filter forwarding architecture for large organizations. In *Proceedings of ACM CoNEXT*, 2009.

VIII. Appendix

A. Latency reduction techniques for host discovery

Multiple location hashes. For each host *i*, the access switch S_i can publish multiple copies of the tuple $< k_i, v_i >$ in the network as follows: S_i applies m independent hash functions on k_i and gets m different location hashes $H_1(k_i)$, ..., $H_m(k_i)$. The switch sends m publish messages to these locations (each message contains the tuple and the id of the hash function). Thus multiple switches, each closest to one of the locations, store the key-value tuple

If a switch wants to request $< k_i, v_i >$, it computes the m location hashes as well as the distances from its own location to the m locations. It then selects the nearest one as the destination of the lookup request. By the property of VPoD, a shorter distance in the virtual space means approximately lower routing cost or latency. Thus using more hash functions trades a higher storage cost for a lower latency in host discovery. We demonstrate this trade-off in experimental results presented in Section VI.

Hashing to a smaller region. To reduce lookup latency, switches can also use two independent hash functions with the second function mapping location hashes to a smaller region in the virtual space. For example, suppose the 2D virtual space of switches is ([0, 100], [0, 100]) and virtual distance

is an accurate estimate of routing latency. If location hashes are distributed over the entire virtual space, the worst-case latency between sending a lookup request and receiving its reply is 282.8. If the hash results are mirrored to a smaller region ([25, 75], [25, 75]), the worst-case lookup latency is 212.1. The average latency is also reduced by using a smaller hash region. This technique can introduce load imbalance among switches, i.e., switches in the smaller region store more tuples than switches outside the region. To avoid load imbalance, the virtual space can be partitioned into two halves; two independent hash functions are used with each function mapping location hashes to one half of the space.

Caching of popular hosts. If there is a set of popular hosts in the network, caching is an effective way to provide fast responses to lookup requests. Each access switch can maintain a cache list that stores the locations of the most popular hosts requested by its hosts.

Shortcuts. A lookup request sent to location $H(k_i)$ does not have to reach the resolver that is closest to $H(k_i)$. Any intermediate switch S can reply to the request for host information and stop forwarding it, under one of three conditions: (1) S is the access switch of k_i; (2) S is a resolver of k_i for a different hash function; (3) k_i's location was previously discovered by S and stored in S's cache.