SAL: Scaling Data Centers Using Smart Address Learning

Alexander Shpiner[1], Isaac Keslassy[1], Carmi Arad[2], Tal Mizrahi[1,2], and Yoram Revah[2]

[1]Technion, {shalex@tx, isaac@ee}.technion.ac.il
[2]Marvell Israel, {carmi, talmi, yoramr}@marvell.com

***Abstract*—Multi-tenant data centers provide a cost-effective many-server infrastructure for hosting large-scale applications. These data centers can run multiple virtual machines (VMs) for each tenant, and potentially place any of these VMs on any of the servers. Therefore, for inter-VM communication, they also need to provide a VM resolution method that can quickly determine the server location of any VM. Unfortunately, existing methods suffer from a scalability bottleneck in the network load of the address resolution messages and/or in the size of the resolution tables.**

In this paper, we propose Smart Address Learning (SAL), a novel approach that expands the scalability of both the network load and the resolution table sizes, making it implementable on faster memory devices. The key property of the approach is to selectively learn the addresses in the resolution tables, by using the fact that the VMs of different tenants do not communicate. We further compare the various resolution methods and analyze the tradeoff between network load and table sizes. We also evaluate our results using real-life trace simulations. Our analysis shows that SAL can reduce both the network load and the resolution table sizes by several orders of magnitude.

I. Introduction

Multi-tenant data centers provide an increasingly popular solution for hosting large-scale service applications [1], [2]. Their appeal comes from their scalability, since they are increasingly cost-effective as they get larger [3]. To ensure scalability, data center providers run multiple virtual machines (VMs) per data center, and can allocate the VMs of a client application to multiple servers, thus also achieving load balancing, fault tolerance and power saving. For efficient implementation of these features, the network has to support unbounded VM placement and migration such that any VM is able to be assigned to any server. In particular, it must provide resolution of the VM location for inter-VM communication: when a new connection is created between two VMs, the initiating VM needs to retrieve the location of the other VM. The services for the physical location resolution of the logical entities have to be supplied by the data center network infrastructure, e.g. by network probing, by the forwarding tables, or by some level of indirection relying on a central database.

Unfortunately, existing location resolution methods often suffer from scalability issues, especially with the resolution network load and the forwarding table size. This is because the network load of the resolution request broadcast messages increases with the number of VMs [4], [5], while it should be kept low in order to leave bandwidth for the application data communication. Moreover, the forwarding-table entries needed for the ever-increasing number of VMs would not fit anymore the on-chip memory that is needed to allow fast access and update times [6], [7]. These issues get especially acute as data centers grow, and may become critical in future multi-tenant data centers, which are expected to include millions of VMs [2], [3].

Several architectures have been proposed to break this scalability bottleneck by using *overlay networks* [4], [8]–[17]. These architectures partition the data center network into segments of broadcast domains, thus solving the problem for intra-segment VM communication given fixed segment sizes. In addition, these architectures use network devices called *edge bridges* to connect between the segments and the network core. The edge bridges provide address resolution for inter-segment VM communication. They can be implemented either in the server hypervisors or at the top-of-rack switches. Unfortunately, they still do not solve the scalability problem of inter-segment address resolution. In fact, [5], [17], [18] state that the overlay network may still suffer from a bottleneck in resolving target stations physical address (MAC or IP) at the overlay edge gateway nodes within the data center. The address resolution broadcast storms may even cause loss of traffic if the cache is too small, and may consume significant bandwidth at large networks.

In this paper, we propose a new address resolution approach called *Smart Address Learning* (SAL). SAL enables scaling the data center while keeping both the resolution table sizes and the network load low. To do so, we use the fact that VMs of different tenants do not communicate directly. Thus, the edge-bridge resolution tables only need to learn addresses of the VMs that belong to the tenants they serve. For instance, if an edge bridge serves a local network with VMs of a tenant i, it only needs to follow the location of the other VMs of tenant i, and can ignore the resolution information of VMs of any tenant $j \neq i$. This selective learning makes the table usage more efficient and increases its hit rate. In addition, SAL decreases the network load, because the VM location updates are only sent to the tables that serve the same tenant, instead of being flooded.

The SAL approach can be easily combined in current data center architectures with any network core routing protocol and it is distributed, scalable and fault-tolerant. It supports any common network core protocol and topology. We introduce two versions of our approach: *pull* and *push*, which differ by the

trigger of address learning.

We further provide an analytical model for the evaluation of the table sizes and the network load under SAL and other resolution methods. In addition, we compare SAL against alternative methods using simulations based on synthetic as well as real-life VM creation, placement and tenancy traces.

To our knowledge, this paper is the first to introduce a model for comparing address resolution methods in data centers, as well as the first to evaluate them using real-life trace simulations.

Our analytical model and simulation results show that SAL can reduce the network load for a given resolution table size by up to four orders of magnitude. It also yields a lower update rate and a higher hit rate in the resolution table, thus potentially enabling implementation of fast on-chip resolution tables even for large multi-tenant data centers.

II. Related Work

Commodity techniques of location resolution in small networks cannot be directly applied to the large-scale data center networks. One such technique, ARP over Ethernet layer-2 infrastructure, limits the scalability of the network due to the high load of the broadcast messages and the large forwarding tables [19], [20]. For instance, [4] states that address resolution traffic constitutes more than 88% of the whole broadcast traffic in the data center networks, and that less than $32,000$ hosts in the same broadcast domain can saturate 100 Mb/s network links with their peak load ARP traffic. Moreover, the broadcast domain is recommended to be limited to several hundreds of nodes. As the network grows, the broadcast messages significantly increase the network load, and the forwarding database tables grow as well, due to a larger number of addresses to learn.

Another commodity technique, the hierarchical IP-based layer-3 addressing, mitigates the advantages of VM migration by limiting it to a specific subnet, because the VM needs to maintain its IP address during its runtime, which can be difficult to do while crossing subnets.

In recent years, several overlay network architectures have been proposed to break this limitations in the data centers [4], [8]–[16], [21]. In these architectures, the VM packets are encapsulated in (or rewritten with) the overlay network headers. The overlay network header is used to route the packet through the network core, which can be implemented using various routing protocols such as commodity Ethernet, hierarchical IP routing, TRILL or MPLS. The encapsulation point, denoted edge bridge, can be for instance the server hypervisor or the top-of-the-rack switch.

Table I summarizes the differences between these approaches, and compares them with our suggested SAL approach. As mentioned, these methods still lack scalability in either network load or table sizes when the number of VMs increases. In addition, the table compares additional desired properties, including distribution, fault-tolerance, and compatibility with the commodity techniques. The overlay methods can be roughly divided into three categories:

Central database — The central database approach is used in VL2 [8] and Portland [12]. The distributed hash table on the aggregation switches, as used in SEATTLE [11], also relates to this category. In this approach each VM location is listed in a unique central consistent database. The edge bridge resolves the location by sending a unicast request message to the consistent directory. Note that the edge bridges also hold a cache table that lists the recently-used resolution entries. The usage of a central address resolution database has several drawbacks. These methods may have scalability problems in large data centers due to frequent resolution updates, unbalanced request rates, fault-tolerance issues, and longer delays for retrieving the information. For instance, [12] states that for maintaining the resolution requests, approximately 70 processing cores are needed, which is beyond the capacity of a single commodity machine. VL2 [8] replicates the database to multiple cached servers. However, this raises consistency and concurrent-replication issues, as well as potential scalability problems when the update rate is high. Moreover, it requires maintaining additional servers for backing up the data. It is also vulnerable to malicious attacks, which lead to service unavailability if the fabric manager fails to perform address resolution [22]. In addition, SEATTLE [11] presents potential fault-tolerance weakness, because the mapping DBs/switches are not backed up, and in a case of DB/switch failure, all the associated mapping information is lost. DHT replication is possible, but generates additional complexity [23].

We next examine two distributed approaches:

Pull — The distributed Pull approach does not rely on a consistent database, but on broadcasting resolution request messages over the network and learning the resolution from the reply. The address resolution is pulled on-demand, meaning, at the time the resolution is required at the edge bridge. This approach is used in EtherProxy [4], Diverter [15], OTV [21], SARP [16] and several other architectures. Unfortunately, this broadcasting may evolve into a vast flooding of the data center network core, and therefore cause a prohibitive network load. Note that here as well, the edge bridges may hold a cache table that lists the recently used resolution entries, and attempt to store entries for the active connections. However, these entries may be inconsistent. Initial VXLAN [9] implementations configured multicast groups per tenant, and broadcasted the address resolution request over the tenant multicast group only. However, the network devices failed to support the large required number of multicast groups. Thus, the VXLAN address resolution was moved to be central-DB-managed in latter implementations [24].

Push — The distributed Push approach relies on sending address resolution updates with each location change. The edge bridges learn the VM addresses at each location update, and manages resolution tables at the edge bridges. Thus, it avoids request broadcasting, but requires larger resolution tables. Keeping the location information consistent and close to the VM allows for a faster start-up time of the new connections and a lower network load. For instance, this approach is used in Netlord NL-ARP address learning approach [13]. Netlord repli-

TABLE I
COMPARISON OF THE RESOLUTION METHODS.

	Location of the consistent information	**Cache inconsistency or miss cost**	**Total number of entries**	**Potential hot spot**
Central DB [8], [12]	Central DB	Request-reply message to DB	Minimal	*Severe*
DHT-based DB [11]	Distributed Hash Table (DHT)	Request message to DHT and redirect	Minimal	*Moderate*
Pull [4], [15], [21]	None	*Resolution request broadcast (high frequency)*	As low as needed	None
Push [13]	Edge bridge (server or TOR switch)	Resolution request broadcast (less likely to happen)	*Maximal*	None
SAL + Pull	None	*Resolution request broadcast (medium frequency)*	As low as needed	None
SAL + Push	Edge bridge (server or TOR switch)	Resolution request broadcast (less likely to happen)	Medium	None

cates the resolution database on every server, and therefore uses a maximal possible number of entries. The edge bridge sends an update message upon every change of the VM status that it is responsible of, similarly to the gratuitous ARP mechanism. Unfortunately, in order to be efficient, this approach requires large tables. Note that if the table capacity is large enough and the update messages always arrive within a negligible time, the push architectures tables are always consistent. Usually, the Push approach is combined with the Pull approach for resolving cases with resolution table inconsistency due to table overflow or resolution packet losses.

In summary, both current centralized and distributed address resolution approaches in the data center have limited scalability when the number of VMs increases.

As mentioned before, our suggested approach is based on selective learning of addresses from the incoming resolution request messages. A similar idea is used in the selective ARP learning [25], where an ARP table is configured to learn a pre-configured specific set of IP addresses. However, the selective ARP learning approach uses only a passive filtering, without dynamic adaption to VM re-placement and to tenancy.

III. NETWORK MODEL AND ASSUMPTIONS

We begin by defining the network model, as illustrated in Figure 1. The model is fairly standard and follows recent literature [4], [8], [12], [16].

We use the terms *application address* (AA) and *location address* (LA) to define both the addresses in the user VM address space and in the physical data center address space, respectively [8]. Note that both those address spaces can be assigned in Ethernet layer 2, IP layer 3, or any other proprietary protocol of the data center provider. For example, in VL2 [8] both AA and LA are IP addresses, while in Portland [12] both AA and LA are MAC addresses. The VM AA is combined from a pair of identifiers: the tenant ID within the data center and the VM ID within the tenant, and thus allows easy association of the VM to a tenant. By the term *resolution* we further refer to the translation of the AA address of a VM into its LA address.

We denote as an *edge bridge* (EB) the encapsulation point where the inter-VM data packets are encapsulated in (or rewritten with) the overlay data center network header. In general, the encapsulation point can be either the ToR switch, the aggregation switch, or the server hypervisor. We assume for simplicity that the communication inside the local network under the EB is L2-protocol based, but other methods would hold as well.

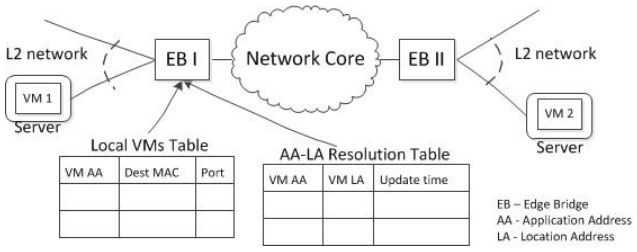


Fig. 1. Network Model. The edge bridge (EB) connects the VMs in its L2 network to the other VMs through the data center network core. The EB implements overlay network encapsulation. It uses two tables for the address resolution. The first is a consistent table that lists all the local VMs under the EB, while the second lists the address resolution of the VMs outside the network under the EB. In the paper, we focus on the scalability of the second table.

Incidentally, the server hypervisor typically implements a virtual switch to connect the hosted virtual machines between themselves and to the network. The virtual machines use a virtual network interface to connect to the virtual switch and believe that they are regular machines connected connected to the network.

One advantage of this approach is that the VMs in a network under the EB are interconnected over an L2 network, and do not necessarily need to send internal messages through the EB. Furthermore, broadcast ARP-request messages that are injected by a VM are stopped at the EB and do not propagate to the core network. For the address resolution requests for VMs outside the L2 network, the EB replies using an ARP-reply message with its own MAC address. This common approach is also used by many other overlay network architectures [4], [8], [12], [16].

Our model supports any common network core protocol and topology. The routing between the EBs can be implemented using standard IP routing with ECMP, MPLS or TRILL tunnels, layer-2 Ethernet with VLANs [26], or any other protocol, as long as each EB can communicate with each other EB.

Each edge bridge stores an *LA-to-AA resolution table* and a local *forwarding table*. The LA-to-AA resolution table is used

to resolve the destination AA for a given LA. The next section introduces SAL, a novel learning scheme for the resolution entries. In addition, the local forwarding table lists the AAs of all the VMs under the edge bridge layer-2 network together with their layer-2 MAC addresses and the output port towards them. We assume that the placement controller of the data center keeps the forwarding table consistent.

The time-out mechanism is popular in the conventional ARP tables, because the tables are stored in a shared memory space and the timeout mechanism avoids their overflow on the account of other system processes. Hence, the ARP tables intend to store the entries for the active entries only. However, in our model, the EBs use dedicated fast memory to store the resolution entries in order to allow fast access times and high bandwidth. So, there is no cost of storing inconsistent entries. On the other hand, repeatedly acquiring information for the resolution entries that were removed by the time-out mechanism cost in additional network load. Therefore, the resolution tables in our model avoid using the time-out mechanism for the entries. The old, last recently used, inconsistent entries are overwritten, when a new resolution information is required to be written to a full memory.

Finally, in the multi-tenant environment, the VMs are divided into groups of tenants. The VMs of a tenant are assumed to communicate only between themselves, and possibly with hosts outside of the data center, but not with VMs of other tenants. This is logical, since they belong to different applications. It also makes sense for security isolation. Therefore, VMs typically only communicate with a small number of other VMs [11], [27], [28]. We will leverage this assumption in the paper to reduce the amount of information that needs to be stored in the resolution tables. For simplicity, we start by focusing on internal VM-to-VM communication in this paper, and neglect the communications to hosts outside of the data center. We later discuss how the inter-tenant communication support can be implemented with our approach.

IV. Smart Address Learning (SAL)

The current approaches for address resolution tables on the overlay network edge simply rely on a list of addresses that can fit within a table size, without using any optimization technique to select the most useful addresses. We would like to present a new approach to improving the usage efficiency of the resolution tables. This approach should allow usage of smaller memories for the table storage, as well as lower the network load of the resolution packets.

The key idea behind our new approach is to selectively learn the addresses from the received address resolution requests that are broadcasted over the network. The main question is *which resolution information needs to be stored in the limited EB tables*.

A first approach, which is quite conservative, is to decide that each EB only learns and updates the resolution entries that it requested, and not those requested by other EBs.

However, to exploit the bi-directional nature of the communication, a second approach is to learn also about outside VMs requesting an inside VM that is located in the EB's subnet, since this inside VM will also probably later communicate with these outside VMs. This method is similar to the scheme used in conventional ARP messages.

In addition, in a third approach, we can leverage the multi-tenant nature of the data center to further restrict the number of learned entries. The inside VMs under a specific EB can potentially communicate only with outside VMs of the same tenants. Therefore, when receiving resolution requests, the EB learns about the VMs of tenants that also have inside VMs – but it does not learn about VMs from other tenants. This is the basis of our SAL algorithm.

In Section VI we compare all these approaches using simulations.

A. SAL Overview

This section presents our suggested Smart Address Learning (SAL) approach. SAL implements a *distributed resolution database*, in which the resolution tables are stored on the edge bridges (EBs).

In SAL, the EB resolution tables only store the addresses of the VMs that belong to the tenants of the VMs hosted in the EB network. For example, consider Figure 2. The servers under EB I host VMs of tenants A and B only. Therefore, the resolution table of EB I only stores the addresses of VMs of the tenants A and B. Likewise, the servers under EB III host VMs of tenant B only, hence the resolution table of EB III would only store VM addresses for tenant B.

More specifically, any EB that broadcasts an address resolution request message will include the AA and LA of its requesting VM. Upon receiving the message, the other EBs will selectively learn this AA-to-LA mapping in their resolution table *if and only if* their network contains another VM of the same tenant as the requesting VM. Therefore, EBs without VMs of this tenant can disregard this message, and as a result their resolution can typically be smaller than without this selective learning. The EBs do not need to store any global resolution information.

In Section V we analyze the scalability properties of SAL and show that SAL is fully scalable if the number of tenants increases proportionally to the number of VMs. In other words, if the number of VMs per tenant is kept fixed as the number of the VMs in the data center increases, both the network load and the resolution tables size are kept constant. However, if the number of VMs increases because each tenant has new VMs, and not because the data center has more customers (tenants), then the advantage of the SAL approach diminishes, since the probability to find at least one VM of a specific tenant on each rack increases.

This section presents how our suggested SAL algorithm updates the EB resolution tables following a VM *location update*, i.e. following a VM creation, destruction or migration. We consider two variants of the update method: *pull* and *push*.

In the *pull* version, the location information is pulled by the EB when this information is required by the encapsulation process, and is not available in its resolution table. On the other

hand, in the *push* variant, the location updates are immediately propagated to other forwarding databases on selected EBs.

Intuitively, the *pull* version is preferable when the location update rate is high relative to the address resolution request rate, and when pushing the updates through broadcasting is costly. We further analyze the tradeoffs involved in the next sections.

B. Pull Update (On-Demand Update)

In the *pull* variant, the location information is pulled to the EB resolution table at the time of resolution request if the information is unavailable in the table. The update is done by broadcasting an address resolution request message to all the other EBs, and receiving a reply from the EB that hosts the requested VM. The request message also contains the AA and LA of the source VM that requests the resolution. In SAL, the smart learning ensures that other EBs that receive this request message only insert this LA address in their tables if they host VM of the same tenant. Each EB knows which tenant VMs it serves using the information from the local VMs forwarding table.

Figure 2 illustrates the *pull* variant. It shows how EB I requests information on VM A.4 by broadcasting a request message, thus *pulling* information from the network. It further emphasizes how only EBs with VMs from tenant will add information on VM A.1, while other EBs such as EB III will not. This is the core selection principle behind the SAL algorithm.

Note that if a VM is migrated during an active connection, its resolution update can be pushed immediately in order to avoid a communication disruption by the migration process. In addition, due to the inconsistent information in the resolution tables, it may happen that an EB receives a data message that is destined to the VM that was previously hosted in its network, but already migrated from it. Then the EB answers the source EB with an error message, and the source EB will re-initiate the full address resolutions process. Incidentally, an optional alternative implementation for the EB is to redirect the packets to the EB of the updated VM location, and then ask it to inform back the source EB of the new location.

For simplicity, we assume that each use and update of an entry in the table refreshes its update timestamp. When the table fills up, the oldest entry is cleared from the table.

C. Push Update (On-Change Update)

In the *push* variant, the updates are pushed to the resolution tables. In our suggested SAL algorithm, in order to reduce network load, the location update broadcast is replaced with messages (several unicast or single multicast) to selected EBs only.

Specifically, upon VM location change, the update is propagated (pushed by either the migration source or the destination EB) immediately only to the EBs that host VMs of the same tenant of the VM. Note that SAL does not require the EB to have a global knowledge on which tenants have clients under each EB, but only the information of tenant VMs that it serves. The destination EBs are known to the sending EB, because it holds the address resolution of all the tenant VMs in its address resolution table. When an update needs to be sent, the EB selects from the forwarding database the location addresses (the destination EBs) of all the VMs of the tenant whose VM is updated. An easy and fast selection can be achieved by assigning application addresses (AAs) that contain the tenant ID in the specific bits, or even better, by logically organizing the table as a tree with a single node per AA, pointing to the different VMs. No additional global information about the VM location is required to be stored.

An alternative implementation is to send the location update message from a data center controller that decides on the placement of the VMs. This controller manages the VM placement and thus has a consistent VM location information.

Figure 3 depicts the *push* process. It shows how EB I *pushes* information on newly-created VM A.5 of tenant A, by selectively sending an update message only to the relevant EBs that contain VMs from the same tenant A. Thus, the network load is typically less than in a full broadcast message.

Special treatment is required in the following two cases. First, when a VM is assigned to an EB network where no other VM of the same tenant exists, the EB needs to retrieve the location information of all other VMs of the tenant. This can be done by broadcasting a request message to all other EBs, or with the assistance of the data center placement controller.

In a second special case, the last VM of a tenant in an EB is removed due to deletion or migration to other EBs. In this case, the EB can remove all the location entries of all other VMs of this tenant in other EBs. This can be done easily by the EB itself, by checking the number of remaining VMs of the tenant in its resolution table after removing a VM.

In order to preserve consistency of the updates, we use a timestamping mechanism based on synchronized clocks in the EBs. An update message holds a timestamp of the update time. Before the table update, the EB validates that the received message timestamp is newer than the last time the entry was updated. Each table entry update refreshes its recently used timestamp. When the table fills up, the oldest entry is cleared from the table.

Inconsistency of the information in the resolution tables may still occur with the push variant. It can happen if the message arrival fails, or if the table is filled up. To overcome this inconsistency, the pull update mechanism is still preserved in the push variant. If the requested entry is missing from the resolution table in the EB, a resolution request is broadcasted to all the other EBs.

D. Inter-tenant Communication Support

As stated in [29], inter-tenant communication is also possible in the data centers. Our design can be extended to support it by reserving entries in the resolution tables for the addresses of such global tenants that communicate with other tenants. A combination of different resolution approaches can be used for the global tenant addresses and the regular tenant addresses. For example, an EB initiates global inter-tenant resolution request by broadcasting the request message as in the *pull* variant,

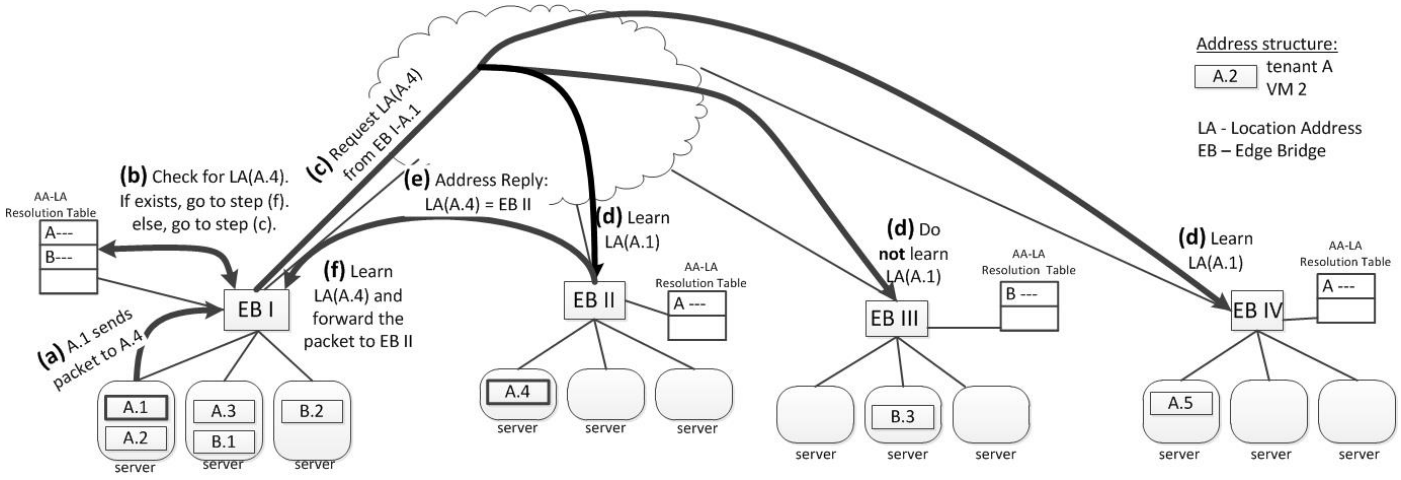


Fig. 2. *Pull* variant of SAL. VM $A.1$ initializes a connection with VM $A.4$ and the resolution process starts. (a) A.1 transmits a data packet destined to A.4. Both VMs belong to tenant A. The packet arrives to EB I, which needs to encapsulate it with the LA of A.4. (b) The LA of A.4 is checked in the resolution table. If it is absent, (c) the EB creates an address resolution message and broadcasts it to other EBs in the network. The address resolution message contains the LA information of the source A.1. (d) Upon reception of this address resolution message, each EB that serves VMs of tenant A learns or updates the LA of A.1. Other EBs do not learn the address in their tables. (e) In addition, EB II that serves the destination A.4, replies by unicast message to EB I with the LA of A.4. (f) Finally, EB I inserts the LA of A.4 in the resolution table and forwards the data packet from A.1 to A.4.

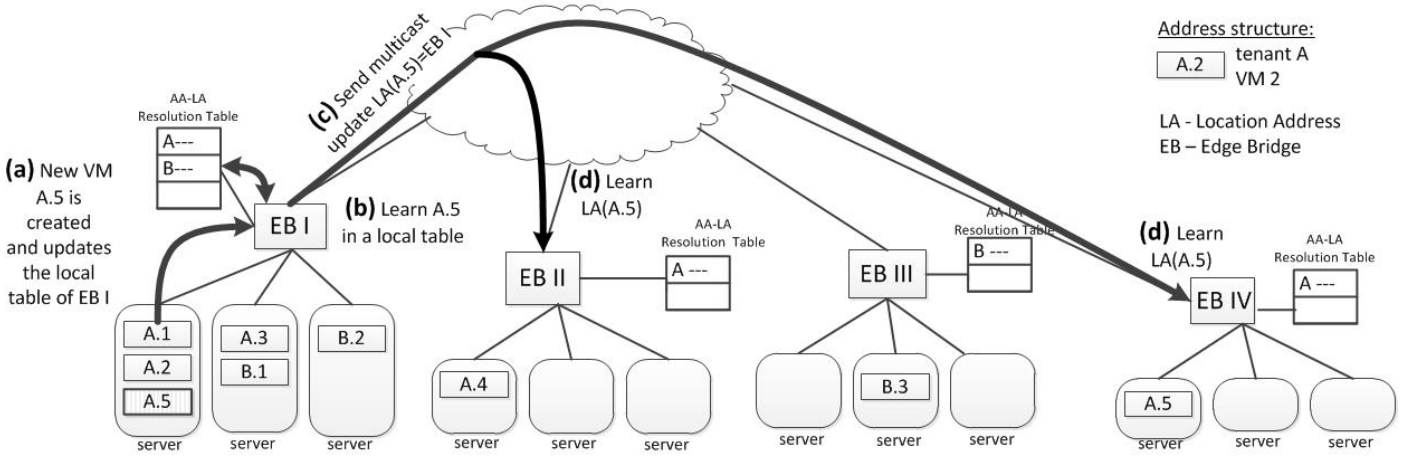


Fig. 3. *Push* variant of SAL. A new VM is created and the resolution update process occurs. (a) The new VM A.5 of tenant A is created in a server under EB I. (b) The VM location is updated in the local EB table. (c) EB I multicasts an address update message with the LA of A.5 to the other EBs that serve VMs of tenant A. EB I determines the EBs to multicast using the entries of tenant A VMs in its resolution table. (d) Finally, the multicast EBs insert the new address in their resolution tables.

even when the intra-tenant resolution uses the *push* variant. The received resolution information from the replied message is stored in the reserved entries for the global tenants addresses.

V. Analytical Model

A. Notations and Assumptions

We would now like to compare the various approaches by formally analyzing their performance. Unfortunately, the performance of each approach is sensitive to many parameters, such as the data center topology, the placement policy, the number of tenants, the distribution of VMs per tenant, the rate of VM creations, migrations and destructions, the burstiness of the application changes, and so on. As a result, to gain some insight, we are reduced to providing a first model in significantly simplified settings. In our analysis we compare the following approaches: Central DB, Push with and without SAL, and Pull with and without SAL. Note that in our model, the Central DB results are also valid for the DHT-Based DB.

Table II illustrates the settings for our model. We make several simplifying assumptions. First, we assume equal-sized tenants, with a fixed number of VMs per tenant, and a fixed table capacity at each EB. As well we assume fixed rates of various VM location update events and VM resolution requests, each following exponentially-distributed inter-event times. We further assume links with infinite capacity and zero propagation time. These assumptions are of course somewhat simplistic, yet they help us better understand the tradeoffs involved in the algorithm design.

Additionally, we consider two simple VM placement strategies: *packed* and *round-robin*, similarly to [14]. These two

TABLE II
ANALYSIS NOTATIONS

N	# of EBs	128
V	# of VMs per EB	640
T	# of tenants	5000
U	# of VMs per tenant ($\equiv VN/T$)	16
C	Active Connections ($\leq VN(U-1)$)	$1.2 \cdot 10^6$
B	EB resolution table capacity	10^5
λ_c	Total VM creation rate (1/sec)	10
λ_m	Total VM migration rate (1/sec)	1
λ_d	Total VM destruction rate (1/sec)	10
λ_u	Total VM location update rate (1/sec) ($\equiv \lambda_c + \lambda_m + \lambda_d$)	21
λ_s	Total resolution request rate (1/sec)	10^5

placement strategies are two extremes that typically cause the best- and worst-performance cases. The *best case* typically corresponds to the *packed placement*, in which VMs of a tenant are locally packed under the lowest number of EBs as possible. This placement is typically chosen to minimize the network load. On the other hand, the *worst case* typically corresponds to the *round-robin placement*, in which VMs of a tenant are spread equally among the servers. This placement strategy may be chosen for its fault-tolerance properties.

For each of the placements we impose an additional condition. We distinguish two cases for each of the placements.
For the packed placement: *Case 1a:* If the number of VMs per tenant is small enough to be placed under a single EB ($U \leq V$ or $N \leq T$), no intra-tenant VM communication is passed through EBs.
Case 1b: Otherwise, ($U > V$ or $N > T$), each tenant occupies several EBs.
For the round-robin placement: *Case 2a:* For the *round-robin* placement, if the number of VMs per tenant is smaller than the number of EBs ($U \leq N$ or $V \leq T$), there are no two VMs of any tenant under the same EB.
Case 2b: Otherwise, if the number of VMs per tenant is larger than the number of EBs ($U > N$ or $V > T$), there are VMs of all T tenants under each EB, and each EB serves $\frac{V}{T}$ VMs of a tenant.

In addition, to quantify our models, as shown in Table II, we assume some typical values (based on [2], [3], [8], [12], [14], [30]–[37], as well as private talks to industry engineers). Also, we assume that the EB is the ToR switch.

B. Resolution Table Length

In this section we evaluate for each of the compared architectures *the resolution table length* (or occupancy), i.e. the required number of resolution entries in the consistent resolution database tables for minimal resolution network load. Specifically, in the *pull* variant, the table length in an EB is the number of resolution table entries needed to support the active connections outgoing from the EB. In the *push* variant, the table length in an EB is the number of VM addresses that are stored in the resolution table. In other words, for a general push architecture it is simply the number of VMs in the data center, while for the push architecture with SAL it is the number of VMs of the tenants that have some VMs under the EB. The local VMs of the EB are not counted. For the central DB, it is simply the number of VMs in the data center.

Table III provides the resolution table length in each of the compared methods, as explained in the next paragraphs. It also provides numerical estimations that are based on the assumptions of Table II.

1) Central DB: The Central DB architecture maintains a central resolution data base that provides resolution for all the VMs, thus the final resolution table length is simply the number of VMs in the data center, which is VN.

2) Push: In the general push architecture each EB table stores the resolution entries for all the VMs in the data center, except the local VMs under EB, thus the resolution table length is $V(N-1)$.

In the push architecture with SAL, the number of resolution entries depends on the placement of VMs, as discussed in Section V-A.

For the *packed* placement, in *Case 1a*, no intra-tenant VM communication is passed through EBs, thus the resolution tables are empty. Otherwise, in *Case 1b*, each resolution table stores entries for one tenant only, of all VMs of a tenant besides the ones that are located under the EB, i.e. a total of $U - V = \frac{V(N-T)}{T}$ entries.

For the *round-robin* placement, in *Case 2a*, there are no two VMs of any tenant under the same EB, thus each resolution table needs to store entries for all other U VMs of a tenant, for each of its V VMs, except for a single local VM, i.e. a total of $V(U-1)$ entries. Otherwise, in *Case 2b*, there are VMs of all T tenants under each EB, and each EB serves $\frac{V}{T}$ VMs of a tenant. Thus, the resolution table, for each of the tenants, stores entries of VMs under other EBs, i.e. total of $T(U - \frac{V}{T}) = V(N-1)$ entries.

3) Pull: In the Pull and SAL-Pull architectures the consistency in the resolution tables is kept only for the entries used in the active connections C. Therefore, each EB resolution table stores consistent entries for the connection between its VMs and VMs under other EBs.

Each tenant has an average of $\frac{C}{T}$ connections between its VMs, out of the $U(U-1)$ possible connections between its U VMs. Therefore, given a pair of VMs, the probability that there is a connection between them is $P_{connect} = \frac{C/T}{U(U-1)} \approx \frac{CT}{V^2N^2}$.

Next, for the evaluation, the previously defined types of placements are considered.

For the *packed* placement, we again observe the two cases. In *Case 1a*, there is no connection between the EBs, thus the resolution tables are empty. Otherwise, in *Case 1b*, the V VMs under EB communicating with other $U-V$ VMs of the tenant outside the EB. Thus, the possible number of VMs to connect to outside of the EB is $V(U-V)$. The probability that an resolution entry for any VM X outside the EB is needed, is the probability that exists any VM from the V VMs under EB that connecting with this VM X. It is equal to $1-(1-P_{connect})^V$. Then, the number of resolution entries in the EB is the product of the number of all resolution entries for the tenant $U-V$

TABLE III
RESOLUTION TABLE LENGTHS (ENTRIES).

Architecture	Packed Placement	Estimation	Round-robin Placement	Estimation
Central DB	VN	$8.2 \cdot 10^4$	VN	$8.2 \cdot 10^4$
Push	$V(N-1)$	$8.2 \cdot 10^4$	$V(N-1)$	$8.2 \cdot 10^4$
SAL-Push	$\frac{V \max\{0,(N-T)\}}{T}$	0	$V(\min\{U,N\}-1)$	$9.8 \cdot 10^3$
Pull, SAL-Pull	$\max\{0, U-V\}(1-(1-\frac{CT}{V^2N^2})^V)$	0	$\frac{C}{N}$, if $U \leq N$; $V(N-1)(1-(1-\frac{CT}{V^2N^2})^{\frac{V}{T}})$, if $U > N$	$9.4 \cdot 10^3$

by the probability that this entry is needed: $(U-V)(1-(1-P_{connect})^V) = (U-V)(1-(1-\frac{CT}{V^2N^2})^V)$.

For the *round-robin* placement, in *Case 2a*, there are $U-1$ potential connections for each for the V VMs under EB. Therefore the number of active connections through the EB is $\frac{CT}{V^2N^2} \cdot V(U-1) \approx \frac{C}{N}$. Otherwise, in *Case 2b*, each of the T tenants has $\frac{V}{T}$ VMs under the EB. For each tenant, the possible number of VMs to connect to outside of the EB is $U - \frac{V}{T}$, and the probability that the entry for VM is needed is $1-(1-P_{connect})^{\frac{V}{T}}$. Thus, the number of connections out of each EB is $T(U-\frac{V}{T})(1-(1-P_{connect})^{\frac{V}{T}}) = T(U-\frac{V}{T})(1-(1-\frac{CT}{V^2N^2})^{\frac{V}{T}}) = V(N-1)(1-(1-\frac{CT}{V^2N^2})^{\frac{V}{T}})$.

4) Summary: Figure 4 plots the resolution table lengths as a function of the number N of EBs, based on Table III. Figures 4(a) and 4(b) show the resolution table length as a function of the number of EBs for the packed and for the round-robin placement, respectively. The values for SAL and for Pull in Figure 4(a) are equal to 0 for $N < T$, therefore they are not seen in the left side of the graphs. Similarly, Figures 4(c) and 4(d) plot the same table lengths, but assuming that the number of tenants T is scaled such that the ratio $\frac{N}{T}$ is kept fixed. The values for SAL and for Pull in Figure 4(c) are equal to 0 for all N, therefore they are not seen in the graphs. Figure 4(d) is interesting since its illustrates the scalability of SAL.

C. Network Load

Next we evaluate the network load of the address resolution management packets as a function of VM location updates and address resolution requests rates. The network load is expressed as the rate of address resolution packets. For ease of an evaluation, a single multicast or broadcast packet to k destinations is counted as k packets.

The network load estimation of the resolution architectures for the packed and round-robin placements is summarized in Table IV.

1) Preliminary Notations: Before we begin with the analysis of the network load, we define several probability notations.

First, we denote the P_{miss} as the probability that the resolution entry is unknown in the table. We assume a uniform probability of each entry to store resolution of any VM. For general Pull or Push method, the P_{miss} is approximated as $P_{\text{miss}} = 1-\min\{1, \frac{B}{VN}\}$, since B is the table length and VN is the total number of VMs to store. Similarly, for SAL-Push approach P_{miss} is approximated as $P_{\text{miss}} = 1-\min\{1, \frac{B}{U \cdot \text{tenants per EB}}\}$, because the table stores entries for VMs of the served tenants only. We consider two types of placement. For the packed placement, the number of tenants per EB is approximated as $\max\{1, \frac{V}{U}\} = \max\{1, \frac{T}{N}\}$ ($P_{\text{miss}} = 1-\min\{1, \frac{B}{U \cdot \max\{1, \frac{V}{U}\}}\}$). For the round-robin placement, it is $\min\{V,T\}$ ($P_{\text{miss}} = 1-\min\{1, \frac{B}{U \cdot \min\{V,T\}}\}$).

We also define P_{wrong} as the probability that the resolution entry in the table is inconsistent. It equals $P_{\text{wrong}} = \frac{\lambda_u}{\lambda_u + \frac{\lambda_s}{N}}$ for Pull. It is equal to 0 for Push, since the entries in Push are consistent.

We also define the probability $P_{\text{in other EB}}$ that the resolution is to another EB and it is calculated as follows. In the packed placement, each tenant occupies $\lceil \frac{U}{V} \rceil = \lceil \frac{N}{T} \rceil$ EBs, so the probability to find the resolution destination under another EB is the complimentary to a probability of finding the destination under current EB, and is equal to $P_{\text{in other EB}} = 1 - \frac{1}{\lceil \frac{U}{V} \rceil}$. In the round-robin placement, each tenant has $\lceil \frac{U}{N} \rceil = \lceil \frac{V}{T} \rceil$ VMs under EB. Therefore, for each tenant there are up to $\lceil \frac{V}{T} \rceil$ VMs in a specific EB out of its all U VMs, so the probability to find the connection destination in another EB is the complimentary to a probability of finding the destination under a specific EB, thus it is equal to $P_{\text{in other EB}} = 1 - \min\{1, \frac{\lceil \frac{V}{T} \rceil - 1}{U}\}$.

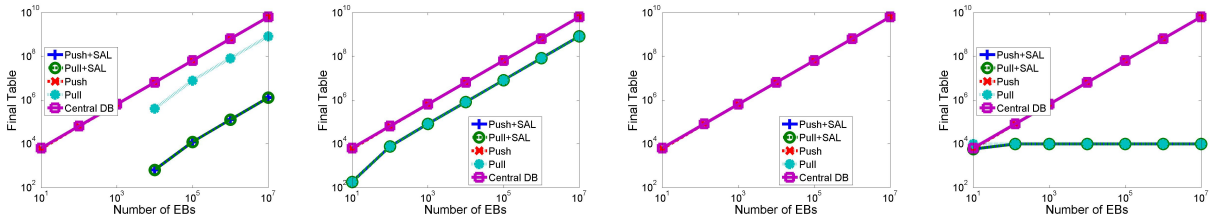
2) Central DB: In the Central DB architecture each new connection retrieves the resolution from central DB. Each new connection with unknown resolution requires two messages: one for the request to the DB and one for the reply back. For the wrong inconsistent resolution entry in the EB table, two additional messages are required: one for sending to a wrong destination and another one for error reply.

Therefore the network load for Central DB is:

$$\begin{aligned} &NL_c \approx \\ &= 2\lambda_s P_{\text{in other EB}}(P_{\text{miss}} + 2(1-P_{\text{miss}})P_{\text{wrong}}). \end{aligned} \tag{1}$$

For the packed placement it is:

$$\begin{aligned} &NL_{c-packed} \approx \\ &2\lambda_s(1-\frac{1}{\lceil \frac{U}{V} \rceil}) \cdot (1-\min\{1, \frac{B}{VN}\} \\ &+ 2(\min\{1, \frac{B}{VN}\})\frac{\lambda_u}{\lambda_u + \frac{\lambda_s}{N}}). \end{aligned} \tag{2}$$



(a) Packed Placement. Fixed number of tenants. (b) Round-robin Placement. Fixed number of tenants. (c) Packed Placement. Number of tenants is scaled with N. (d) Round-robin Placement. Number of tenants is scaled with N.

Fig. 4. Model. Resolution Table Length as a Function of Number N of EBs.

and for the round-robin placement it is:

$$NL_{c-roundrobin} \approxeq 2\lambda_s(1-\min\{1,\frac{\lceil\frac{V}{T}\rceil-1}{U}\})\cdot(1-\min\{1,\frac{B}{VN}\} + 2(\min\{1,\frac{B}{VN}\})\frac{\lambda_u}{\lambda_u+\frac{\lambda_s}{N}}). \quad (3)$$

In DHT-based DB architecture the address resolution is done by the resolver switches, thus it is similar to Central DB with the exception that the resolution requests are to the resolution switches in which the DHT is located. Therefore the network load of the resolution packets is similar to the load in Central DB architecture.

3) Push: In the Push architecture each location update involves broadcasting update messages to all other $N-1$ edge bridges. Also, in the absence of the requested entry from the table, the EB needs to broadcast the resolution request to $N-1$ other EBs and receive one reply. Therefore, the network load in the Push architecture is:

$$NL_{push} \approxeq \lambda_u(N-1)+\lambda_s N P_{\text{in other EB}} P_{\text{miss}}. \quad (4)$$

For the packed placement it equals:

$$NL_{push-packed} \approxeq \approx \lambda_u(N-1)+\lambda_s N(1-\frac{1}{\lceil\frac{U}{V}\rceil})(1-\min\{1,\frac{B}{VN}\}). \quad (5)$$

and for the round-robin placement it equals:

$$NL_{push-roundrobin} \approxeq \lambda_u(N-1)+ + \lambda_s N(1-\min\{1,\frac{\lceil\frac{V}{T}\rceil-1}{U}\})(1-\min\{1,\frac{B}{VN}\}). \quad (6)$$

4) SAL-Push: SAL-Push variant is similar to general Push with the difference that the update messages are sent only to the selected EBs. In the packed placement, the average number of EBs under which the VMs of a single tenant VMs are stored is $\lfloor\frac{U}{V}\rfloor$. In *Case 1a* it is equal to 0, and no update messages are required. In *Case 1b* the network load equals:

$$NL_{S-push-packed} \approxeq \lambda_u(\lfloor\frac{U}{V}\rfloor-1)+\lambda_s N P_{\text{in other EB}} P_{\text{miss}} = = \lambda_u(\lfloor\frac{U}{V}\rfloor-1)+ + \lambda_s N(1-\frac{1}{\lceil\frac{U}{V}\rceil})(1-\min\{1,\frac{B}{U\cdot\max\{1,\frac{V}{U}\}}\}). \quad (7)$$

In the round-robin placement, in *Case 2a*, each location update requires $U-1$ messages, one for each other VM of a tenant. Therefore, the network load equals:

$$NL_{S-push-roundrobin-Case3} \approxeq \lambda_u(U-1)+\lambda_s N P_{\text{in other EB}} P_{\text{miss}} = = \lambda_u(U-1)+ + \lambda_s N(1-\min\{1,\frac{\lceil\frac{V}{T}\rceil-1}{U}\})(1-\min\{1,\frac{B}{U\cdot\min\{V,T\}}\}). \quad (8)$$

Otherwise, in *Case 2b*, there is a VM of each tenant on each EB, and location update requires a message to every other EB. Therefore, the network load is:

$$NL_{S-push-roundrobin-Case4} \approxeq \lambda_u(N-1)+\lambda_s N P_{\text{in other EB}} P_{\text{miss}} = = \lambda_u(N-1)+ + \lambda_s N(1-\min\{1,\frac{\lceil\frac{V}{T}\rceil-1}{U}\})(1-\min\{1,\frac{B}{U\cdot\min\{V,T\}}\}). \quad (9)$$

5) Pull: In Pull architectures the network load consists of the cost of broadcasting resolution request messages to all other EBs. The broadcasting happens when the requested entry is not in the table, or when the entry in the table, but holds the wrong resolution. The last case can happen if the requested VM has moved since its last entry update in the table. The network load for Pull architecture equals:

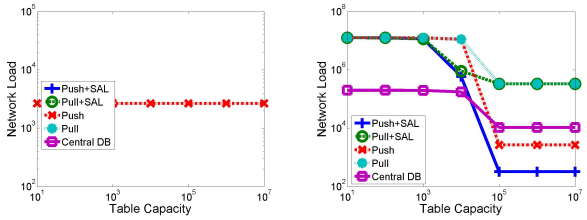
$$NL_{pull} = NL_{S-pull} \approxeq \lambda_s N P_{\text{in other EB}}(P_{\text{miss}}+(1-P_{\text{miss}})P_{\text{wrong}}). \quad (10)$$

Using the notations defined in Section V-C1, with packed placement, the network load for general Pull equals:

$$NL_{Pull-packed} \approxeq \lambda_s N(1-\frac{1}{\lceil\frac{U}{V}\rceil})\cdot \cdot((1-\min\{1,\frac{B}{VN}\})+(\min\{1,\frac{B}{VN}\})\frac{\lambda_u}{\lambda_u+\frac{\lambda_s}{N}}). \quad (11)$$

and with round-robin placement, the network load for general Pull equals:

$$NL_{Pull-roundrobin} \approxeq \lambda_s N(1-\frac{\lceil V/T\rceil-1}{U})\cdot \cdot((1-\min\{1,\frac{B}{VN}\})+\min\{1,\frac{B}{VN}\}\frac{\lambda_u}{\lambda_u+\frac{\lambda_s}{N}}). \quad (12)$$



(a) Packed Placement. (b) Round-robin Placement.

Fig. 5. Model. Network Load as a Function of Table Capacity B.

6) SAL-Pull: In continue to Equation 10 the network load of SAL-Pull with packed placement equals:

$$NL_{S-pull-packed}$$
$$\lambda_s N(1-\frac{1}{\lceil\frac{U}{V}\rceil})\cdot((1-\min\{1,\frac{B}{U\cdot\max\{1,\frac{V}{U}\}}\})+ \quad (13)$$
$$+(\min\{1,\frac{B}{U\cdot\max\{1,\frac{V}{U}\}}\})\frac{\lambda_u}{\lambda_u+\frac{\lambda_s}{N}}).$$

and for the round-robin placement:

$$NL_{S-Pull-roundrobin}$$
$$\lambda_s N(1-\frac{\lceil V/T\rceil-1}{U})((1-\min\{1,\frac{B}{U\cdot\min\{V,T\}}\})+ \quad (14)$$
$$+(\min\{1,\frac{B}{U\cdot\min\{V,T\}}\})\frac{\lambda_u}{\lambda_u+\frac{\lambda_s}{N}}).$$

7) Summary: The expressions are next evaluated in Figures 5 and 6.

Figure 5 shows the network load of the resolution packets as a function of the table capacity B in the packed placement and the round-robin placement. The packed placement result is trivial, as resolution packets are sent in the Push architecture only. The round-robin placement result is explained next. We can see that for small table capacities, the hit rate is low in the EBs under all architecture, therefore the Central DB architecture has lower network load, because the resolution requests are sent in unicast and not flooded as in other architectures. In large table capacities, the Push architectures has lower network load than in other architectures, because the tables are large enough to store the resolution for all the VMs, the information is updated instantly and the multiple request broadcasts are avoided. For Pull, the SAL approach improves the network load under middle table capacities. For small table capacities, the difference in the gain of a slightly better hit rate in the tables is negligible considering the miss cost, and for the large table capacities, the tables are large enough. Both in Pull with and without SAL the table hit rate is not bounded by the table capacity.

Figure 6 shows the network load as function of number of EBs N in the packed placement and round-robin placement.

Figures 6(a) and 6(b) present the network load for the packed and round-robin placements, respectively, keeping other parameters fixed. In the packed placement, the network load for $N \leq T$ is equal to 0 in all the architectures besides Push. For large N, Push with SAL and the Central DB outperform the other approaches. For the round-robin placement, the Central DB approach outperforms the distributed approaches. Note that the Central DB approach has several drawbacks that were discussed in Section II and are not reflected in the figure. Figures 6(c) and 6(d) present the network load as function of number of EBs (N) for the packed and round-robin placements, respectively, scaling also the number of tenants (T), such that the ratio N/T is kept fixed. Figure 6(d) is especially interesting because it illustrates the scalability of the push version of SAL. Moreover, Figures 6(e) and 6(f) present the network load as function of number of EBs (N) for the packed and round-robin placements, respectively, scaling also the number of tenants (T) and the rates λ_c, λ_m, λ_d and λ_s with N. Finally, Figures 6(g) and 6(h) present the network load as function of number of EBs (N) for the packed and round-robin placements, respectively, scaling also the number of tenants (T), the rates λ_c, λ_m, λ_d and λ_s, and the table capacities (B) with N. With the packed placement, only the Push architecture has a positive network load, since all the VMs of each tenant are served by a single EB. With round-robin placement, Push with SAL outperforms the other approaches. Figures 6(i) and 6(j) present the network load as a function of the number of EBs (N) for the packed and round-robin placements, respectively, scaling also the rates λ_c, λ_m, λ_d and λ_s, and the table capacities (B) proportionally to N, but keeping the number of tenants (T) fixed.

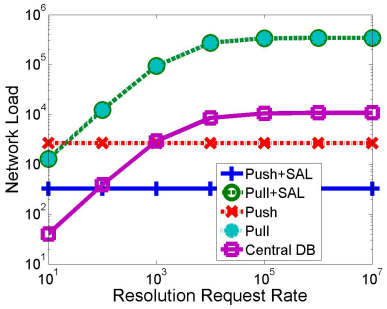


Fig. 7. Model. Network Load as a Function of Resolution Request Rate

D. Impact of Resolution Request Rate

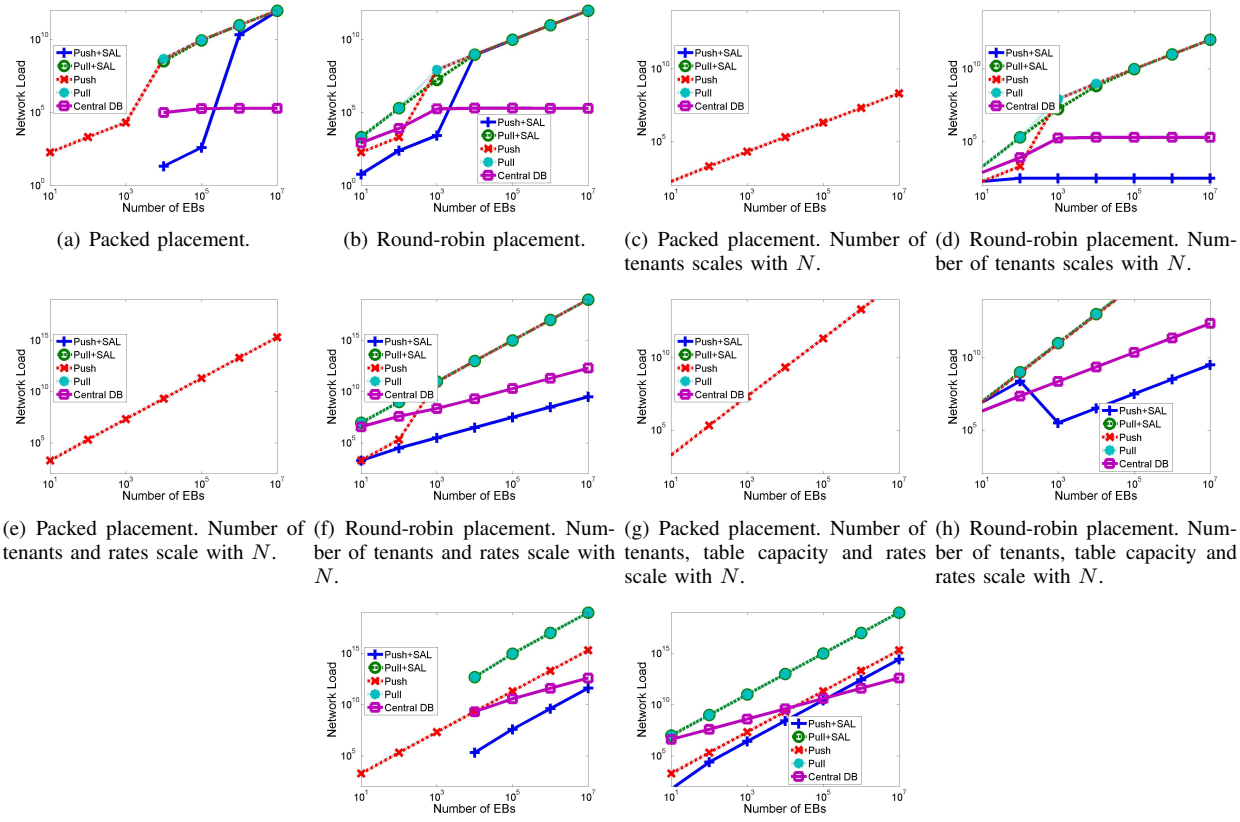
The choice between Push-based and Pull-based models depends on the ratio between the resolution request rate and the VM location update rate. Figure 7 shows the change in the network load as a function of the resolution request rate. All other parameters are set as in Table II. The network load under the Push-based schemes does not increase following the change in the resolution request rate, while it does under the Pull-based schemes. Therefore, under low resolution requests rates, the Pull-based schemes are preferable, but under high rates, the Push-based schemes are preferable.

E. Impact of Number of Tenants

Next, we evaluate how the number of tenants impacts the system performance. Figure 8 shows the change in the network load caused by resolution packets as a function of the number of tenants. For the evaluation we varied the number of tenants

TABLE IV
NETWORK LOAD.

Architecture	Packed Placement	Round-robin Placement
Central DB	$2\lambda_s(1 - \frac{1}{\lceil \frac{U}{V} \rceil})(1 - \min\{1, \frac{B}{VN}\} + 2(\min\{1, \frac{B}{VN}\})\frac{\lambda_u}{\lambda_u + \frac{\lambda_s}{N}})$	$2\lambda_s(1 - \min\{1, \frac{\lceil \frac{V}{T} \rceil - 1}{U}\})(1 - \min\{1, \frac{B}{VN}\} + 2(\min\{1, \frac{B}{VN}\})\frac{\lambda_u}{\lambda_u + \frac{\lambda_s}{N}})$
Push	$\lambda_u(N-1) + \lambda_s N(1 - \frac{1}{\lceil \frac{U}{V} \rceil})(1 - \min\{1, \frac{B}{VN}\})$	$\lambda_u(N-1) + \lambda_s N(1 - \min\{1, \frac{\lceil \frac{V}{T} \rceil - 1}{U}\})(1 - \min\{1, \frac{B}{VN}\})$
SAL-Push	$\lambda_u(\lfloor \frac{U}{V} \rfloor - 1) + \lambda_s(N-1)(1 - \frac{1}{\lceil \frac{U}{V} \rceil})(1 - \min\{1, \frac{B}{U \cdot \max\{U,V\}}\})$	$\lambda_u(\min\{U, N\} - 1) + \lambda_s(N-1)(1 - \min\{1, \frac{\lceil \frac{V}{T} \rceil - 1}{U}\})(1 - \min\{1, \frac{B}{U \cdot \min\{V,T\}}\})$
Pull	$\lambda_s N(1 - \frac{1}{\lceil \frac{U}{V} \rceil})((1 - \min\{1, \frac{B}{VN}\}) + (\min\{1, \frac{B}{VN}\})\frac{\lambda_u}{\lambda_u + \frac{\lambda_s}{N}})$	$\lambda_s N(1 - \frac{\lceil V/T \rceil - 1}{U})((1 - \min\{1, \frac{B}{VN}\}) + \min\{1, \frac{B}{VN}\}\frac{\lambda_u}{\lambda_u + \frac{\lambda_s}{N}})$
SAL-Pull	$\lambda_s N(1 - \frac{1}{\lceil \frac{U}{V} \rceil})((1 - \min\{1, \frac{B}{\max\{U,V\}}\}) + (\min\{1, \frac{B}{U \cdot \max\{U,V\}}\})\frac{\lambda_u}{\lambda_u + \frac{\lambda_s}{N}})$	$\lambda_s N(1 - \frac{\lceil V/T \rceil - 1}{U})((1 - \min\{1, \frac{B}{U \cdot \min\{V,T\}}\}) + (\min\{1, \frac{B}{U \cdot \min\{V,T\}}\})\frac{\lambda_u}{\lambda_u + \frac{\lambda_s}{N}})$

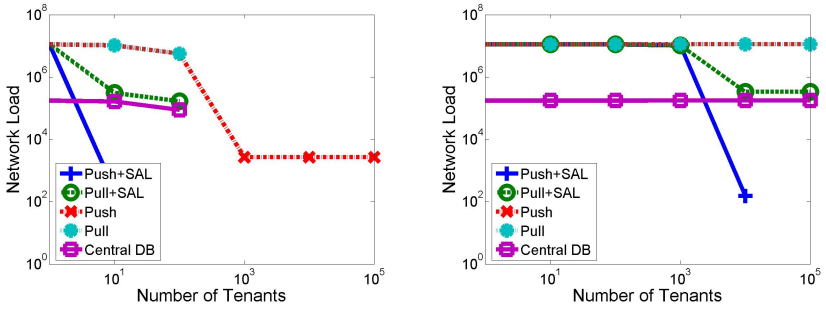


(a) Packed placement. (b) Round-robin placement. (c) Packed placement. Number of tenants scales with N. (d) Round-robin placement. Number of tenants scales with N.

(e) Packed placement. Number of tenants and rates scale with N. (f) Round-robin placement. Number of tenants and rates scale with N. (g) Packed placement. Number of tenants, table capacity and rates scale with N. (h) Round-robin placement. Number of tenants, table capacity and rates scale with N.

(i) Packed placement. Table capacity and rates scale with N. (j) Round-robin placement. Table capacity and rates scale with N.

Fig. 6. Model. Network Load vs Number of EBs N.



(a) Packed placement. (b) Round-robin placement.

Fig. 8. Model. Network Load vs Number of Tenants T.

from $T = 1$ to $T = 10^5$ and set the resolution table size to $B = 10^4$ entries with all other parameters as in Table II. As the number of tenants increase above $T = 10^3$ a significant drop in the network load is observed in our SAL-based algorithms.

VI. Simulations

In this section, we describe a set of simulation results evaluating SAL and comparing it to several existing address resolution methods.

A. Simulator

We implemented an event-driven simulation of the data center network address resolution system. The simulation includes *VM location update events*, i.e. creations, migrations and destructions, as well as *VM address resolution events*, which are initiated by VMs and request for a resolution of other VMs.

In the simulation, tenants are defined as disjoint sets of virtual machines. The VMs are assigned to the hosting edge bridges independently of the tenant they belong to. Furthermore, the source and destination VMs of each resolution request are chosen uniformly within the VMs of each tenant.

We implemented the following address resolution schemes: *Central DB*, *Push* with and without SAL, and *Pull* with and without SAL. The Pull scheme consists of three variants besides SAL. On the figures they are marked by *Pull (complete)*, *Pull (connection)* and *Pull (conservative)*. The difference between these three Pull schemes is the way in which the EB learns resolution information from the incoming broadcast resolution requests that are not destined to the EB. In *Pull (complete)*, the EBs stores information of each incoming resolution request message. In *Pull (connection)*, only the entries that already exist in the resolution table are updated, but no new entry is learned. This method is similar to ARP. Lastly, in *Pull (conservative)*, the EB learns only from resolution requests it initiated and the resolution requests that are destined to it.

In all the schemes, the table lengths are limited by a fixed table capacity. When an entry is added to a full table, the oldest entry is overwritten. In addition, an entry with a wrong information is revealed when it is accessed. The wrong entries and the missing entries are resolved by the broadcast resolution request messages to all the servers — except for the Central DB scheme, where the resolution is done by an access to the central directory.

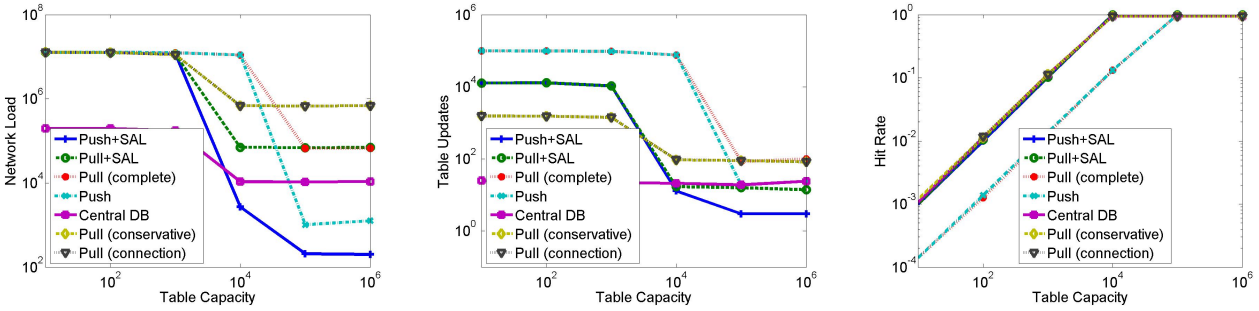
The output of the simulator includes the number of transmitted resolution messages, as well as the occupancy, the number of updates, and the hit percentage of the resolution tables. For simplicity, we neglect the impact of the network topology. Thus, each unicast message between a pair of VMs is counted as a single message, and a multicast or a broadcast message is counted as the number of recipients. For example, a request broadcast by an EB in a data center with N EBs is counted as $N - 1$ messages, since it is sent to $N - 1$ EBs; and the unicast reply is counted as a single message. Pulling the address resolution data base in the Central DB architecture is counted as two messages: one for the request and one for the reply. Revealing a wrong entry costs two additional messages: one for sending a packet to a wrong destination, and the second for receiving a reply message indicating that the destination is wrong.

B. Synthetic Trace Simulation Results

We start by running simulations with a synthetically-generated trace. We use the typical values from Table II, and vary the table capacity B from 10 to 10^6 entries. The placement distribution is uniform, such that at every placement decision, the edge bridge for each VM is chosen uniformly over all VMs. New VMs pick uniformly their tenants. The VM chosen for migration or destruction are also picked uniformly. At the initial state of the simulations, the data center is full with random VMs up to its capacity ($V \cdot N$). The simulations are run until the steady state.

Figures 9(a), 9(b) and 9(c) show the impact of the resolution table capacity on the mean resolution packet network load, the largest mean update rate of a table, and the mean hit rate, respectively, for each of the architectures. Note that for the Central DB, the shown table capacity is for the tables in EBs and not for the central data base.

Specifically, Figure 9(a) confirms our intuition that as table capacity increases, the miss rate decreases and therefore network load decreases, up to a specific large value of table capacity, beyond which there are no further gains. The result also supports our insight from the model that for larger table sizes, the Push architectures perform better than the Pull



(a) Network Load vs. Resolution Table Capacity. (b) Largest Table Update Rate vs. Resolution Table Capacity. (c) Mean Table Hit Rate vs. Resolution Table Capacity.

Fig. 9. Synthetic Event Trace Simulation Results.

architectures, and that the SAL approach for both Push and Pull reduces the network load for some ranges of table capacities, while never increasing the network load. The plot can be divided into three regions of interest: small, medium and large table sizes. For small table sizes, it seems that the preferred resolution method is the Central DB. However, it also relies on a large memory storage with a central data base that holds the resolution of all the VMs. This large memory is not reflected in this plot. It has additional drawbacks that were discussed in Section II. The small-table sizes regions presents the case where the memory is so small that the distributed approaches must rely on frequent broadcasts of the resolution request. Thus, the differences between the push and pull variants are diminished. On the other side, in the large table region, the memories are large enough and can store enough entries. The push variant is preferred over the pull since there is always enough memory to store the consistent entries. The SAL addition to push decreases the network load, due to its selective updates instead of the broadcast. The middle-sized-table region is the most interesting one, because it reflects the network load saving due to the SAL addition both to the push and pull variants.

Figure 9(b) presents the update rate of a single table. By table update, we define each change of the resolution entry in the table, including the address change in an existing entry, and an old entry overwrite for a different VM. For the Central DB architecture, the updates are counted on the central data base, since it suffers a larger update rate than the EB tables. Since the central data base is updated upon each VM location change only, the shown update rate for Central DB is fixed for any table size in the EB.

Figure 9(c) confirms the intuition that the table hit rate increases with the table capacity, and that for the Push and complete Pull architectures, the hit rate is lower than for the other approaches, since in these architectures the resolution tables store information about VMs that are irrelevant.

C. Benchmark Trace

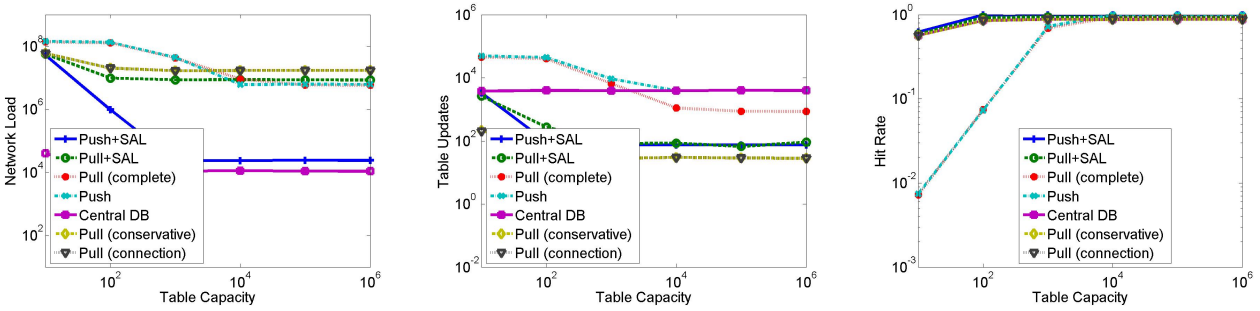
Next, we evaluate the system with a benchmark trace from the IBM Research Compute Cloud (RCCv2), where the customer data was anonymized [38]. The extracted events from the trace are (a) the creation and (b) the destruction times for 38,000 various VMs distributed among 7,000 tenants placed under 3,000 servers, which we assume to function as edge bridges in the data center, as well as (c) their placement, and (d) their tenant assignment. Furthermore, the address resolutions are randomly added with a ratio of 100 resolution events per VM location update event. The RCCv2 system does not include migration, thus the update events only consist of VM creation and destruction events.

Figure 10 is analogous to Figure 9 of the synthetic trace simulations. Most algorithms behave similarly. Moreover, since we now use a slightly higher rate of VM location updates compared to the resolution request rate, the Central DB approach presents a lower asymptotic network load than the Push architectures. This is because a higher location update rate requires unnecessary location update messages in the Push architectures, since an update message may be unnecessary in practice when a VM is moved again before its location resolution is requested by the other EBs. Although the Central DB architecture slightly outperforms the Push with SAL approach, it still requires a higher table update rate, as shown in Figure VI-C. Clearly, higher VM update rates have are detrimental for Push architectures, other parameters being equal. Lastly, Figure 10(c) shows that the SAL approach also improves the resolution table hit rate.

D. Placement Strategy Effect

Next we check the effect of the placement strategy on the resolution packets network load and the resolution table length. We already discussed in Section V-A the two extreme placement strategies: *packed* and *round-robin*. We simulate *hybrid placement strategies* in which, given a parameter p between 0 and 1, each placement decision picks the *packed* placement strategy with probability p, and the *round-robin* placement strategy with probability $1-p$. We run simulations with the hybrid placement strategies by varying p from 0 to 1 in steps of 0.2. The resolution table capacities are chosen as infinity large so as to evaluate the resolution table length in an unconstrained manner. Other parameters are chosen based on the values in Table II.

Figure 11 shows the largest resolution table at the end of the simulation run vs. the cumulative number of resolution packets



(a) Network Load vs. Resolution Table Capacity. (b) Largest Table Update Rate vs. Resolution Table Capacity. (c) Mean Table Hit Rate vs. Resolution Table Capacity.

Fig. 10. Benchmark Trace Simulation Results.

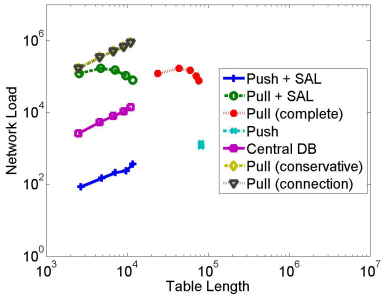


Fig. 11. Placement Effect on Network Load and Resolution Table Length. $p = 0$ is the rightmost point for each architecture. $p = 0.8$ is the leftmost point for each architecture. For $p = 1$(packed placement only) the values are equal to 0, thus are not shown.

sent for the synthetic trace. For each type of architecture the results from various hybrid placement strategies are connected by a line. The rightmost point for each architecture line is for $p = 0$, and the leftmost point is for $p = 0.8$. For the packed placement ($p = 1$), all the VMs of each of the tenants are packed under a single EB, thus no resolution request is exchanged between the EBs and no updates are pushed in SAL. Therefore, the result for $p = 1$ is omitted, since all the values are equal to 0. The only exception is the Push architecture, in which the updates are still pushed between the EBs and the network load and table sizes are larger than 0. Also, for the Push architecture, all the values (including $p = 1$) are concentrated in the graph, since it is less affected by the placement strategy.

It appears that the relative performance of diverse approaches is relatively insensitive to the placement strategy. Therefore, the main insight is that the impact of the placement strategy is *less significant* than we expected before running the simulation.

VII. Conclusions

In the paper we proposed Smart Address Learning (SAL), a novel approach that expands the scalability of current address resolution mechanisms in the data centers, for both the network load and the resolution table sizes, which makes it possible to be implemented on faster memory devices. The key property of the approach is to selectively learn the addresses in the resolution tables, based on the fact that the VMs of different tenants do not communicate.

We presented an analytical model of the network load and resolution table sizes for the presented resolution methods. We further used the model and simulations to evaluate the tradeoff of the network load and the resolution table size. Our analysis showed that both the network load and the resolution table sizes can be reduced by orders of magnitude depending on the system parameters.

More generally, to our knowledge, this paper is the first to introduce a model for comparing address resolution methods in data centers, as well as the first to evaluate them using real-life trace simulations. A more advanced analysis of the optimal address resolution tradeoff in data centers is left for future work.

Acknowledgment

The authors would like to thank Orna Agmon Ben-Yehuda and Aran Bergman for their helpful comments, as well as Mariusz Sabath and David Breitgand, IBM WRC, who kindly shared the data of the IBM Research Compute Cloud (RCCv2) traces [38]. This work was partly supported by the Hasso Plattner Institute Research School, the Intel ICRI-CI Center, the Israel Ministry of Science and Technology, and European Research Council Starting Grant No. 210389.

References

[1] M. Saluan, "Want to Provide Cloud Services? You Need to Understand Multi-Tenancy," http://mspmentor.net/blog/want-provide-cloud-services-you-need-understand-multi-tenancy, 2013.

[2] J. Metzler, A. Metzler, and et al., "The emerging data center LAN," *Webtorials Analyst Division, Cloud Networking Reports 2010 - 2012*.

[3] N. Ilyadis, "The evolution of next-generation data center networks for high capacity computing," in *VLSI Circuits (VLSIC)*, 2012.

[4] K. Elmeleegy and A. Cox, "Etherproxy: Scaling Ethernet by suppressing broadcast traffic," in *IEEE INFOCOM*, 2009.

[5] L. Dunbar, S. Hares, M. Sridharan, N. Venkataramaiah, and B. Schliesser, "Address resolution for large data center problem statement," in *ARMD BOF*, 2011. [Online]. Available: http://tools.ietf.org/html/draft-dunbar-armd-problem-statement-01

[6] D. Meyer, L. Zhang, and K. Fall, "Report from the IAB workshop on routing and addressing," in *IETF, RFC 4984*, 2007.

[7] G. Hankins, "Pushing the limits, a perspective on router architecture challenges," in *North American Network Operators Group, NANOG 53*, 2011.

[8] 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," ser. ACM SIGCOMM, 2009.

[9] M. Mahalingam and et al., "VXLAN: A framework for overlaying virtualized layer 2 networks over layer 3 networks," in *Network Working Group Internet Draft*, 2011.

[10] M. Sridharan and et al., "NVGRE: Network virtualization using generic routing encapsulation," in *Network Working Group Internet Draft*, 2011.

[11] C. Kim, M. Caesar, and J. Rexford, "Floodless in seattle: a scalable ethernet architecture for large enterprises," in *ACM SIGCOMM*, 2008.

[12] R. Niranjan Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya, and A. Vahdat, "Portland: a scalable fault-tolerant layer 2 data center network fabric," in *ACM SIGCOMM*, 2009.

[13] J. Mudigonda, P. Yalagandula, J. Mogul, B. Stiekes, and Y. Pouffary, "NetLord: a scalable multi-tenant network architecture for virtualized datacenters," ser. ACM SIGCOMM, 2011.

[14] T. Benson, A. Akella, A. Shaikh, and S. Sahu, "CloudNaaS: a cloud networking platform for enterprise applications," in *SOCC*, 2011.

[15] A. Edwards, A. Fischer, and A. Lain, "Diverter: a new approach to networking within virtualized infrastructures," in *ACM WREN*, 2009.

[16] Y. Nachum, L. Dunbar, I. Yerushalmi, and T. Mizrahi, "Scaling the address resolution protocol for large data centers (SARP)," in *INTAREA Working Group Internet Draft (work in progress)*, 2013. [Online]. Available: http://tools.ietf.org/html/draft-nachum-sarp-04

[17] T. Narten, M. Karir, and I. Foo, "Address resolution problems in large data center networks," in *Internet Engineering Task Force (IETF)*, 2013. [Online]. Available: http://tools.ietf.org/html/rfc6820

[18] L. Dunbar, W. Kumari, and I. Gashinsky, "Practices for scaling ARP and ND for large data centers," in *Network Working Group Internet Draft (work in progress)*, 2013. [Online]. Available: http://tools.ietf.org/pdf/draft-dunbar-armd-arp-nd-scaling-practices-06.pdf

[19] A. Myers, T. E. Ng, and H. Zhang, "Rethinking the service model: Scaling ethernet to a million nodes," 2004.

[20] B. Stephens, A. L. Cox, S. Rixner, and T. S. E. Ng, "A scalability study of enterprise network architectures," in *ACM/IEEE ANCS*, 2011.

[21] Cisco, "Overlay transport virtualization (OTV)," http://www.cisco.com/c/en/us/solutions/data-center-virtualization/overlay-transport-virtualization-otv/index.html.

[22] F. Bari, R. Boutaba, R. Esteves, M. Podlesny, G. Rabbani, Q. Zhang, F. Zhani, and L. Granville, "Data center network virtualization: A survey," *IEEE Communications Surveys and Tutorials*, 2012.

[23] R. Rodrigues and B. Liskov, "High availability in dhts: Erasure coding vs. replication," in *Proceedings of the 4th International Conference on Peer-to-Peer Systems*. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 226–239. [Online]. Available: http://dx.doi.org/10.1007/11558989_21

[24] G. Kinghorn, "Cisco VXLAN innovations overcoming IP multicast challenges," http://blogs.cisco.com/datacenter/cisco-vxlan-innovations-overcoming-ip-multicast-challenges/, 2013.

[25] R. Chamarajanagar, P. Hunt, S. Kimble, T. Nguyen, and G. Rashiyamany, "Selective passive address resolution learning," in *US Patent Application 20080144634*, 2008.

[26] J. Mudigonda, P. Yalagandula, M. Al-Fares, and J. C. Mogul, "SPAIN: COTS data-center Ethernet for multipathing over arbitrary topologies," ser. USENIX NSDI'10, 2010.

[27] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken, "The nature of data center traffic: measurements and analysis," in *ACM IMC*, 2009.

[28] P. Bodík, I. Menache, M. Chowdhury, P. Mani, D. A. Maltz, and I. Stoica, "Surviving failures in bandwidth-constrained datacenters," in *ACM SIGCOMM*, 2012.

[29] H. Ballani, K. Jang, T. Karagiannis, C. Kim, D. Gunawardena, and G. O'Shea, "Chatty tenants and the cloud network sharing problem," in *Proceedings of the 10th USENIX conference on Networked Systems Design and Implementation*. Berkeley, CA, USA: USENIX Association, 2013, pp. 171–184. [Online]. Available: http://dl.acm.org/citation.cfm?id=2482626.2482644

[30] R. Katz, "Tech titans building boom," *IEEE Spectrum*, vol. 46, no. 2, pp. 40 –54, Feb. 2009.

[31] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: state-of-the-art and research challenges," *Journal of Internet Services and Applications*, 2010.

[32] C. Guo, G. Lu, H. J. Wang, S. Yang, C. Kong, P. Sun, W. Wu, and Y. Zhang, "SecondNet: a data center network virtualization architecture with bandwidth guarantees," in *ACM Co-NEXT*, 2010.

[33] "Amazon web services LLC," https://aws.amazon.com.

[34] "Microsoft Corporation, an overview of Windows Azure," http://www.microsoft.com/downloads/details.aspx?displaylang=en\&FamilyID=96d08ded-bbb9-450b-b180-b9d1f04c3b7f.

[35] B. Stephens, A. Cox, W. Felter, C. Dixon, and J. Carter, "PAST: scalable ethernet for data centers," in *ACM CoNEXT*, 2012.

[36] A. Shieh, S. Kandula, A. Greenberg, and C. Kim, "Seawall: performance isolation for cloud datacenter networks," in *HotCloud*, 2010.

[37] D. Ármannsson, G. Hjálmtýsson, P. D. Smith, and L. Mathy, "Controlling the effects of anomalous arp behaviour on ethernet networks," in *ACM CoNEXT*, 2005.

[38] G. Ammons et al., "RC2: A living lab for cloud computing," *IBM, IBM Research Report RC24947*, 2010.