Cloudcasting: A New Architecture for Cloud Centric Networks

Richard Li, Kiran Makhijani, Lin Han
American Research Center
Huawei Technologies
Santa Clara, California, USA
e-mail: {renwei.li, kiran.makhijani, lin.han}@huawei.com

***Abstract*— Network overlays play a key role in the adoption of cloud oriented networks, which are required to scale and grow elastically and dynamically up/down and in/out, be provisioned with agility and allow for mobility. Cloud oriented networks span over multiple sites and interconnect with Virtual Private Network (VPN) like services across multiple domains. In literature, there have been some proposals to implement network overlays such as, Virtual eXtensible Local Area Networks (VXLAN) as the data plane and Border Gateway Protocol/Ethernet VPN (BGP/EVPN) as the control plane. However, none of them meets all the above requirements. This paper presents the new network architecture, called Cloudcasting, along with its reference model and related protocols, both on the control plane and the data plane, which can demonstrably meet all the requirements. The cloudcasting architecture includes four elements: Cloud Rendezvous Point (CRP), Cloud Switching Point (CSP), Cloud Control (CCC) protocol, and Virtual Extensible Network (VXN) Encapsulation Protocol.**

Keywords-Cloud; Network Overlay; Network Virtualization; Routing, Multi-Tenancy Virtual Data Center; VXLAN; BGP; EVPN.

I. INTRODUCTION

The key characteristics of Cloud-oriented data center architectures are resource virtualization, multi-site distribution, scalability, multi-tenancy and workload mobility. These are typically enabled through network virtualization overlay technologies. Initial network virtualization approaches relate to layer-2 multi-path mechanisms such as, Shortest Path Bridging (SPB) [3] and Transparent Interconnection of Lots of Links (TRILL) [5] to address un-utilized links and to limit broadcast domains. Later, much of the focus was put into the data plane aspects of the network virtualization, for example, VXLAN [1], Network Virtualization using Generic Routing Encapsulation (NVGRE) [2], and Generic Network Virtualization Encapsulation(GENEVE) [9]. These tunneling solutions provide the means to carry layer-2 and/or layer-3 packets of tenant networks over a shared IP network infrastructure to create logical networks. Though, due to their lack of corresponding control plane schemes, they require painstaking orchestration of the system for the virtual network setup and maintenance [10][11]. Even more recently, MP-BGP/EVPN [4] has been proposed as a control plane for virtual network distribution, and has foundations of the VPN style provisioning model. This requires additional changes to an already complex and a heavy protocol that was originally designed for the inter-domain routing. The deployment of MP-BGP/EVPN in data center networks also brings in corresponding configurations, for example, defining autonomous systems (AS), that are not really relevant to the data center infrastructure network.

The existing solutions such as, multi-path, custom-orchestrations and Multiprotocol-BGP (MP-BGP) [6][7] are a class of virtual network architectures that consume protocol data structures of substrate networks, therefore, we refer to them as *Embedded Virtual Networks*. The term *substrate network* henceforth will be used to describe a base, underlying, or an infrastructure network upon which user networks are built as virtual network overlays.

In this paper, a new network virtualization approach is proposed, which does not require changing the substrate protocols. It can connect different types of virtual networks through its own routing scheme. Since, such scheme can be organized over any substrate network topology and routing arrangement; it is referred to as *Extended Virtual Networks*.

Even though *Embedded Virtual Network* (the term is inspired from [17]) solutions mentioned above have irrefutable benefits, they also have several limitations. Of which the most significant and relevant to cloud-scale environments is their dependence on the substrate networks. In addition to being scalable and reliable, a cloud scale network must also be elastic, dynamic, agile, infrastructure-independent, and capable of multi-domain support. There has not been a single technology which works as a converged architecture for network virtualization. In this paper, we propose an *Extended Virtual Network* framework that operates on top of substrate network and offers primitives for cloud auto-discovery, dynamic route distribution as needed. Operationally, this new network architecture allows for agile provisioning and allows for the interconnection of hybrid clouds.

The rest of the paper is organized as follows. Section II describes a reference model for cloudcasting, and major functions of its reference points. Section III explains the signaling communication primitives between the cloudcasting reference points and Section IV uses a multi-tenant virtualized data center as a deployment example. While Section V highlights the advantages and strengths of the solutions, Section VI compare our solution the related work. Lastly, Section VII briefly lays out the directions for our future work.

II. CLOUDCASTING MODEL

A converged virtual routing scheme can be described by two primary factors; an infrastructure-independent virtual network framework, and a unified mechanism to build an overlay of various types of user networks with different

 ISBN: 978-1-61208-455-8

address schemes. On these basis, a new virtual routing scheme called *Cloudcasting,* is proposed with the following characteristics

(1) *Auto discovery*: A signaling scheme that enables us to add, delete, expand and virtualize a tenant's network with minimum configuration.

(2) *Auto distribution:* A signaling scheme that connects multiple virtual networks with each other or asymmetrically as needed.

(3) *Auto Scale:* The ability to provide and serve high scale of tenants in a location-agnostic manner.

A ***cloudcasting network*** is an IP network, which is shared and used by multiple tenant clouds to route traffic within a single virtual network or between different virtual networks. We use the terminology of tenant cloud to emphasize that a tenant or a user network may reside anywhere on the substrate network with a highly dynamic routing table. The IP address space in one tenant cloud may overlap with that in another cloud and these are not exposed to the shared IP infrastructure network.

The cloudcasting reference model, is shown in Figure 1. Each customer has its own network shown as *Tenant Cloud A, B and C*, a shared substrate IP network that was built independently and can encompass multiple administrative domains. This model describes a centralized conversational scheme, in which tenant clouds or Virtual Extensible Networks (VXNs) announce their presence as well as membership interests to a centralized designated authority, called Cloudcasting Rendezvous Point (CRP), via a cloudcasting network virtualization edge element called Cloudcasting Switching Point (CSP).

To communicate among the network elements, a new signaling protocol, called **C**loud**C**asting **C**ontrol (CCC) protocol is defined with three simple primitives facilitating cloud auto-discovery and cloud route distribution. The protocol primitives are defined as below and are further illustrated in Figure 2.

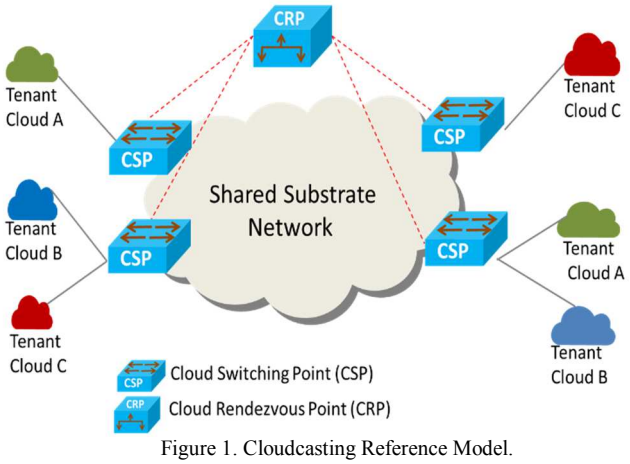

Figure 1. Cloudcasting Reference Model.

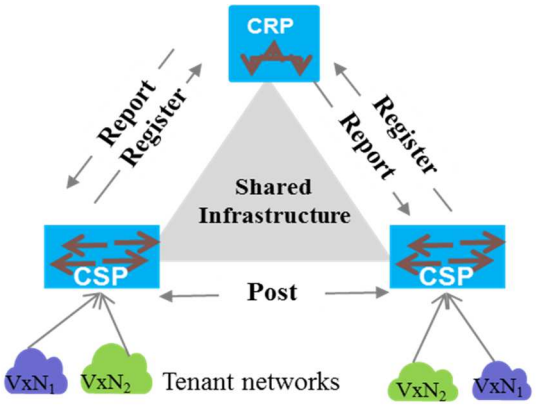

Figure 2. Cloudcasting Framework.

Register message: A virtual network interest and self-identifying announcement primitive from CSP to CRP.

Report message: A response from CRP to all CSPs with similar virtual network interests.

Post message: A CSP to CSP virtual network route distribution primitive.

The details of aforementioned cloudcasting network elements and their properties in cloudcasting framework are discussed as below.

A. *Virtual Extensible Network*

A Virtual Extensible Network is a tenant cloud or a user network. It is represented by a unique identifier with a global significance in cloudcasting network. Using this construct, it is possible to discover all its instances on the substrate IP fabric via CRP. VXN identifiers are registered with CRP from CSPs to announce their presence. There are various possible formats to define the VXN, for instance, an alphanumeric value, number or any other string format. In the preliminary work we have defined it as a named string which is mapped to a 28-bit integer identifier, thus enabling support for up to 256 million clouds.

B. *Cloud Switch Point*

A Cloud Switch Point is a network function that connects virtual networks on one side to the substrate IP network on the other side. It can be understood as an edge of a virtual network that is cloudcasting equivalent of a Virtual Tunnel End Point (VTEP) [1] in VXLAN networks or an Ingress/Egress Tunnel Router (xTR) in the LISP domain [15] and may similarly be co-located with either on a service provider's edge (PE) router, on a top of rack (ToR) switch in a data center, or on both.

A CSP participates in both auto-discovery and auto-route distribution. In order to establish a forwarding path between two endpoints of a virtual network or of two different virtual networks, a CSP first registers with the CRP its address and VXN identifiers it intends to connect to. Then the CRP will report to all CSPs that have the same VXN membership interest. Finally, the CSP will

 ISBN: 978-1-61208-455-8

communicate with those other CSPs and exchange their routing information. On the data forwarding plane, a CSP builds a virtual Forwarding Information Base (vFIB) table on per VXN basis and route/switch traffic to the destination virtual networks accordingly.

C. Cloud Rendezvous Point

A Cloud Rendezvous Point is a single logical entity that stores, maintains and manages information about Cloud Switching Points and their VXN participation. The CRP maintains the latest VXN to CSP membership database and distributes this information to relevant CSPs so that they can form peer connection and exchange virtual network routes automatically. A report message is always generated whenever there is a change in the virtual network membership database. However, CRP is oblivious to any change in vFIB (described above in CSP).

III. CLOUDCASTING COMMUNICATION PRIMITIVES

Now, we describe cloudcasting communication primitives used among CRP and CSPs. Figure 3 illustrates the layering of the virtual routing over any substrate layer and overlay control messages between CSP and CRP.

The encapsulation message format is shown above in Figure 4. A well-known TCP destination port identifies the cloudcasting protocol and CCC header contains the specification for the register, report and post messages.

A. Cloudcasting Register Message

An auto-discovery of virtual networks involves two messages. The first message is the Cloudcasting Register Message that originates from CSPs to announce their presence and interests with the CRP to learn about the other CSPs with the same interest of VXNs. A Register message specification includes the CSP address and VXN identifier list of its interest. An interest is defined as an intent to participate in a specific virtual network. For example, a vxn_{red} on csp_1 expresses 'interest' to join vxn_{red} on csp_2.

As an example, consider virtual networks vxn_{red} and vxn_{green} are attached to csp_1. Then, the register message contains a tuple as follows

$$Register\ \{sender:\ csp_1,\ [vxn_{red},\ vxn_{green}]\}$$

After the CRP receives a cloudcasting register message, it scans its CSP membership database to look for the same VXN identifiers.

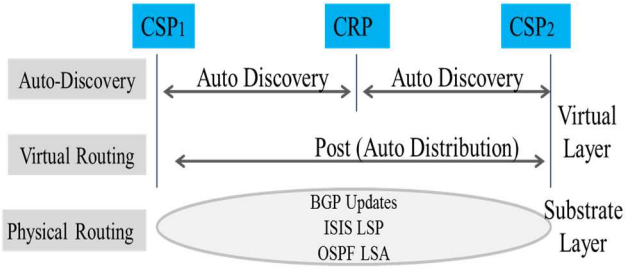

Figure 3. Cloudcasting Protocol Primitives

If it finds one (or more), a cloudcasting report message is generated and sent to all the CSPs with the same interest, otherwise, it simply logs the VXN in its CSP database.

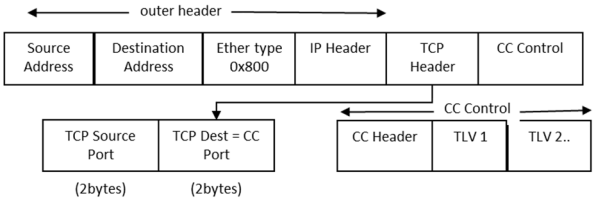

Figure 4. Cloudcasting Control Message Format

B. Cloudcasting Report Message

The CRP generates cloudcasting report messages in response to a cloudcasting register message to inform CSPs of other CSPs' address and their associated VXN identifiers. If the CRP finds other CSP(s) with the same VXN membership (or interested VXNs), then the Report messages are generated for that CSP as well as the other found CSPs. A Report message is sent to each CSP, that contains other CSP addresses for the shared interest VXNs. As an example, consider CRP already has csp_2 with interest vxn_{red}. Upon receiving a cloudcasting register message from csp1 as described earlier, two report messages are generated as below for csp_2 and csp_1, respectively:

$$Report\ (csp_2)\ \{to:\ csp_1,\ [interest:\ vxn_{red}]\}$$

$$Report\ (csp_1)\ \{to:\ csp_2,\ [interest:\ vxn_{red}]\}$$

In this manner, auto-discovery of virtual network locations is accomplished that is based on interest and announcement criteria.

C. Cloudcasting Post Message

The cloudcasting post messages facilitate route distribution as needed. As a cloudcasting report message is received, the CSP will connect with other CSPs to exchange their routing information that includes VXN identifiers, a Generic VXN encapsulation (GVE) tag and the network reachability information within the VXN along with the address family. The list of network reachability information type includes but not restricted to IP prefixes (such as, IPv4, IPv6), VLANs, MAC addresses or any other user defined address scheme.

As an example, when a report as described earlier is received, the following Post will originate from csp_1.

$$Post\ (csp_1,\ csp_2)\ \{vxn_{red},\ gve:\ i,\ [AF:\ IPv4,\ prefix\ list\ldots]\}$$

In the example above, it is shown that csp_1 sends a post update to csp_2 which states that vxn_{red} will use encapsulation tag *'i'*; and that it has certain ipv4 prefixes in its IP network.

The routing (network reachability) information has the flexibility to support various address families (AF) defined by IANA as well as certain extensions not covered under the IANA namespace.

 ISBN: 978-1-61208-455-8

D. Cloudcasting Transport - Generic VXN Encapsulation

In a cloudcasting network, all network devices will work exactly the same as before on the data plane except the Cloud Switch Points (CSP). A CSP will perform encapsulation and decapsulation by following the VXN vFIB table. A VXN vFIB table includes the routing information for a virtual network on a remote CSP where a packet should be destined to. The route information was learned by exchanging Post messages between CSPs.

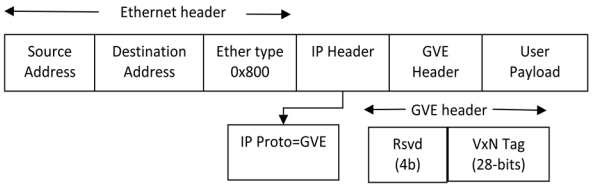

Figure 5. Cloudcasting Data Plane Encapsulation

The format for VXN encapsulation is shown in Figure 5 above in which IP protocol is set to GVE and following IPv4 header is the 32-bit GVE-header. The protocol number for GVE will be assigned by IANA.

IV. USE CASES

Figure 6 shows a cloudcasting-enabled virtualized data center. As discussed earlier in Section I, the CRP is a logically centralized node that is accessible by all the CSPs. A leaf-spine switch architecture is used as a reference to explain cloudcasting deployment. A plausible co-location for CRP could be with the spine node, however, it may be anywhere in the substrate network as long as CSPs can reach it with the infrastructure address space. In Figure 6, several tenant networks are shown as connected to different CSPs and CSP function itself is co-resident with the leaf switches. Each CSP has a virtual FIB table for both encapsulation and decapsulation of traffic along with the tenant network to CSP memberships (dynamically learned through auto-discovery).

The cloudcasting control protocol flow is shown in lighter color lines between CRP and CSPs and among CSPs. At the bottom of the Figure 6. only the logical GVE data path tunnels with dotted lines for tenant 1 on CSP-1, CSP-3 and CSP-4 are shown.

V. EVALUATION AND ANALYSIS

The cloudcasting architecture and primitives have been implemented in our research laboratory. We have successfully used the cloudcasting architecture and control protocol to implement the following use cases:

- Multi-Tenancy Virtual Data Centers
- Multi-Site Interconnection of Data Centers
- Interconnection of Hybrid Clouds
- VPN Accesses to Virtual Data Centers

First and foremost, we emphasize that the cloudcasting architecture represents a paradigm shift. It is a truly converged technology for virtual networks, clouds, and VPNs. No matter what the structure of the underlying substrate network is, any/all types of virtual tenant networks can be constructed in the same way by using cloudcasting.

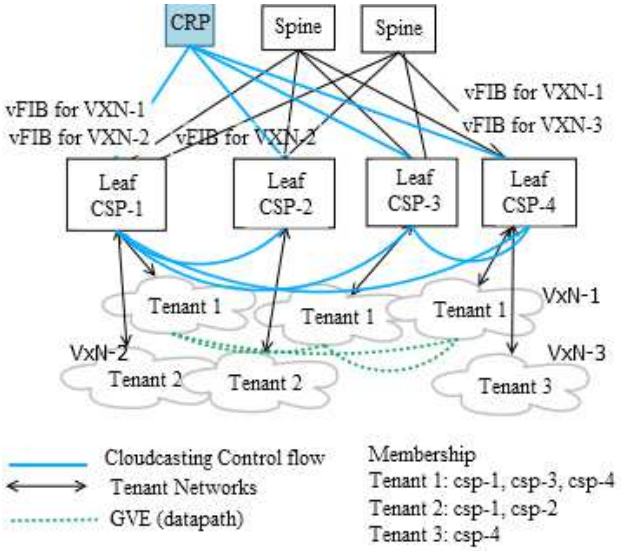

Figure 6. Cloudcasting Enabled Deployment.

The Cloudcasting suitability and applicability can only be verified vis-à-vis characteristics of the cloud-scale environments. Therefore, we have laid importance on the primary characteristics of cloud centric networks that are elasticity, efficiency, agility, and distribution.

The Cloudcasting control plane is *elastic*, because it can grow and shrink independently of (1) the heterogeneous protocols of the substrate network, (2) number of virtual network attachment points, the CSPs, (3) number of domains (autonomous systems), (4) number of routes within a user's virtual network, and (5) mobile nature of the host stations.

The Cloudcasting control plane is *efficient*, because (1) no CSP distributes routes to other CSPs that they are not interested in, (2) thus, no CSP receives and stores routes of virtual networks of non-interest or the ones it is not connected to. In addition, the control plane is fully *distributed* in such a manner that through a single primitive (post-update); change in the tenant networks can be announced immediately, from the spot of change without configuration changes.

The Cloudcasting allows for *agile* networking. Every time when a new CSP is added, it is only required to configure the newly added CSP by using a few lines of commands. Every time when a CSP is deleted, no additional configuration change or for that matter nothing else needs to be done. This is because cloudcasting has a built-in auto-discovery mechanism that has not been seen in the *embedded virtual networks*.

The Cloudcasting data plane *scales* as well. Its default GVE encapsulation protocol allows to support 256 million clouds. In other technology such as, VXLAN, it only up to 16 million clouds are supported.

 ISBN: 978-1-61208-455-8

Due to the limitation of space, we won't discuss and describe other more desirable characteristics.

VI. RELATED WORK

There are several works available that partially solve network virtualization problem; however, they do not provide a complete and consistent solution that sufficiently fulfills all basic requirements discussed earlier in this paper. In what follows, we discuss and compare a few prominent network-overlay approaches.

A. IETF NVO3

The cloudcasting architecture and protocol shares some goals chartered by the IETF working group NVO3 (Network Virtualization Overlays over Layer 3) [16]. The purpose of the NVO3 is to develop a set of protocols and/or protocol extensions that enable network virtualization within a data center environment that assumes an IP-based underlay. Cloudcasting varies from NOV3 in that cloudcasting is not just restricted to the data center, and it doesn't expect a specific structure or protocol conventions in the underlay. The NVO3 architecture may seem to be a reformulation of the BGP architecture, where NVEs (Network Virtualization Edge) and NVA (Network Virtualization Authority) resemble iBGP speakers and Route Reflectors, respectively, and NVO3-VNTP [14] resembles BGP update messages between an iBGP speaker and its Route Reflector. And therefore, NVA (RR) needs to learn and store routes from NVE (iBGP speaker) and then distribute those routes to other NVEs (iBGP speakers). However, this is not the case in Cloudcasting, wherein virtual route information is a function between CSPs, and the CRP is not involved. CRP is used for cloud membership auto-discovery and thus enables agile provisioning. Auto-discovery is missing from NVO3. We should emphasize that CRP has no route database inside. Auto-discovery mechanism in the Cloudcasting has a significant impact on the size of the database in CSP, and is also a common differentiator with other related work as discussed in the following sections.

B. VXLAN and BGP/EVPN

VXLAN is a data plane for network overlay encapsulation and decapsulation, and BGP/EVPN has been proposed as the control plane for VXLAN [4][12][13]. It works by adding new patches to BGP, which was originally designed for inter-domain routing for service providers.

The use of BGP in a data center will require some unnecessary operational actions and design concepts. For example, in order to deploy BGP/EVPN, the network operator must configure something like an AS (autonomous system) in substrate networks, which is not really a data center design concept.

Running BGP in a data center can also lead to serious scalability problems of peering sessions between iBGP speakers (VTEP-BGP). Typically, to address this problem, deployment of Route Reflectors (RR) is suggested which then speaks with every other VTEP-BGP to synchronize their BGP-RIB. As a result, no matter if a VTEP needs routes, all the other VTEPs will always send their routes to the VTEP either directly or indirectly through a Route Reflector, and the VTEP is required to filter out not needed routes through Route Target and other BGP policies. Distributing not needed virtual routes from RR to VTEP-BGP will levy an unnecessary overhead on the substrate network and burn CPU power, processing these BGP messages.

Operating BGP in the data centers not only makes operational cost of data centers as high as that of a service provider's network it also lacks the agility because BGP heavily relies on configurations (it is well known that configuration errors are a major cause of system failures [8]). For example, when a new BGP-VTEP is added/removed the operator has to configure all the BGP peering relationships by stating which BGP neighbors are peering among each other.

Finally, observe that when BGP was first designed, some principles were built-in; for example, iBGP peers should have received and synchronized the same copies of routes. In the case of clouds, many such principles are not applicable anymore.

Compared with BGP/EVPN, our cloudcasting architecture does not suffer from the drawbacks described above. By the means of auto-discovery and route distribution, only specific routes of a virtual network are distributed. Moreover, the role of CRP does not require it to be an intermediate hop between two CSPs to distribute the routes. The detailed comparison and evaluation is in progress and will be published elsewhere.

C. LISP based data center virtualization

Although Locator ID Separation Protocol (LISP) [15] is not an inherent data center virtualization technology, it has a framework to support network overlays. LISP achieves this by distributing encapsulated tenant (customer) routing information and traffic over provider (substrate) network through its control plane based on a mapping system. The LISP architecture includes Ingress/Egress Tunnel Routers (xTRs) and a mapping system (MS/MR) that maintains mappings from LISP Endpoint Identifiers (EIDs) to Routing Locators (RLOCs). LISP requires mapping information to be pulled on-demand and data-driven, xTRs also implement a caching and aging mechanism for local copies of mapping information.

Compared with LISP, Cloudcasting CSPs and LISP xTRs are similar in that they are the virtualization tunnel endpoints performing encapsulation and decapsulation. But the virtual route RIB or mapping databases are different in that (1) LISP's mapping database is a separate protocol element and xTR's local mapping database is built by pulling and kept by caching and aging, while a CSP's virtual network RIB is local and significant only to the CSP itself; (2) An xTR's local database is built on demand after receiving a packet without knowing its mapping information, while CSP's virtual network RIB is signaled through the cloudcasting control protocol; (3) A CSP can auto-discover other CSPs which join the same cloud, while

 ISBN: 978-1-61208-455-8

LISP xTR can only know about another particular xTR after querying the mapping database.

VII. FUTURE WORK

In this paper, we have presented a new routing scheme, called cloudcasting, for virtual networks. Some of the areas being further investigated and formulated include:

- Multicast Support: How does multicast work for different tenant networks on a common and shared substrate network?
- Interface between substrate and virtual network: A careful, thorough research on such interfaces is an area that may be covered through a converged policy model for cloudcasting.

While deliberately left out are discussions on various related topics such as, security, mobility, global scalability and inter-domain deployment models, these topics are being actively worked upon and are the key research areas moving forward.

VIII. CONCLUSION

Cloud-scale networking environments require a technology where virtual networks are first class objects; such that the coarse policies and routing decisions can be defined and applied on the virtual networks. Cloudcasting is a routing system based on converged, unified network virtualization and will evolve better because of lower provisioning costs and enhanced agility through auto discovery when compared with current virtual network schemes where virtual networks will rise only up to the limits of substrate network technologies.

ACKNOWLEDGMENT

The authors are thankful to the members of the Cloudcasting project, particularly those involved in the early design and proof of concept and prototype development for their collaboration and fruitful discussions.

REFERENCES

[1] M. Mahalingam, et al, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, 2014.

[2] M. Sridharan, et al, "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-08 (work in progress), April 13, 2015.

[3] IEEE 802.1aq, “IEEE Standard for Local and metropolitan area networks—Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks—Amendment 20: Shortest Path Bridging”, June 2012, doi: 10.1109/IEEESTD.2012.6231597.

[4] A. Sajassi et al., "A Network Virtualization Overlay Solution using EVPN", draft-ietf-bess-evpn-overlay-01.txt, work in progress, February 24, 2015.

[5] J. Touch, R. Perlman, "Transparent Interconnection of Lots of Links (TRILL): Problem and Applicability Statement", RFC 5556, May 2009.

[6] D. Fedyk, P. Ashwood-Smith, Allan, A. Bragg, and P. Unbehagen, "IS-IS Extensions Supporting IEEE 802.1aq Shortest Path Bridging", RFC 6329, April 2012.

[7] D. Eastlake, T. Senevirathne, A. Ghanwani, D. Dutt, and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 7176, May 2014.

[8] T. Xu, Y. Zhou, “Systems Approaches to Tackling Configuration Errors: A Survey”, Article No.: 70, ACM Computing Surveys (CSUR) Volume 47 Issue 4, July 2015, pp. 70.1-70.41, doi:10.1145/2791577.

[9] J. Gross, T. Sridhar, et al., "Geneve: Generic Network Virtualization Encapsulation", draft-ietf-nvo3-geneve-00, May 2015.

[10] Cisco Nexus 7000 Series NX-OS VXLAN Configuration Guide. Available from: http://www.cisco.com/c/en/us/td/docs/switches/datacenter/sw/nx-os/vxlan/configuration/guide/b_NX-OS_VXLAN_Configuration_Guide.html.

[11] VMware® NSX for vSphere (NSX-V) Network Virtualization Design Guide. Available from: https://www.vmware.com/files/pdf/products/nsx/vmw-nsx-network-virtualization-design-guide.pdf.

[12] E. Rosen, and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, 2006.

[13] Sami et al., draft-boutros-bess-vxlan-evpn-00.txt, “VXLAN DCI Using EVPN”, January 2016.

[14] Z. Gu, Virtual Network Transport Protocol (VNTP), draft-gu-nvo3-vntp-01, October 2015.

[15] D. Farinacci, V. Fuller, D. Meyer, and D. Lewis, "The Locator/ID Separation Protocol (LISP)", RFC 6830, January 2013.

[16] M. Lasserre, et al., "Framework for Data Center (DC) Network Virtualization".

[17] M. Yu, Y. Yi, J. Rexford, and M. Chiang, “Rethinking virtual network embedding Substrate support for path splitting and migration”, ACM SIGCOMM Computer Communication Review, Volume 38 Issue 2, April 2008, pp. 17–29.

 ISBN: 978-1-61208-455-8