Demand-Based Control Planes for Switching Fabrics

Modern switching fabrics use virtual network overlays to support mobility, segmentation, and programmability at very large scale. Overlays are a key enabler of the Software-Defined Networking (SDN) model of control and data plane separation for network abstraction and programmability. Locator Identity Separation Protocol (LISP) as a control plane for overlays is particularly well suited to accommodate the logical centralization of the control plane for programmability. The virtual network overlay uses a control plane to keep the mapping of endpoints to their network location up to date as endpoints move around the network. This control plane must be scalable, nimble, and extensible in handling the endpoint-to-location mappings. It is not, however, required to handle the path calculations that traditional routing protocols must handle. The path calculations required to transport traffic between locations over a multipathed network are executed by the control plane of the underlying network and are therefore abstracted from the overlay control plane.

The functionality required in an overlay control plane is similar to what the DNS infrastructure provides. DNS handles the worldwide scale of name-to-IP mappings. At its foundation, this is simply a mapping database that is consulted by anyone requiring a connection when the connection is required. The results of the consultation are cached by the requester for a finite amount of time. Furthermore, a demand system such as DNS can be enriched with information beyond names and IP addresses, making it easily extensible.

These essential characteristics of a demand-based protocol are the foundation of two key attributes of modern switching fabrics: scalability and extensibility. For the next generation of switching infrastructure, we have chosen to optimize the use of a demand-based protocol in order to improve scale and convergence and to be able to extend the scope of functionality delivered by the overlay. In a majority of the cases, we have also been able to simplify the mechanics of the protocol to support all the traditional services expected from a network and more.

Segmentation

Segmentation has become a baseline requirement in building access networks, in light of the increasing exposure to security threats that spread laterally within a network. The use of segmentation limits the reach of these threats by limiting the radius of connectivity allowed to any given device. For example, heating, ventilation, and air conditioning devices may not reach user devices.

Segmentation at Layer 2 and Layer 3 is built into the demand-based solutions. The segmentation attribute is simply an extension of the host address (IP, MAC, or other). A segmentation attribute is used to color the control plane exchange as well as the data plane transmittal of traffic. The coloring of the traffic defines the segment to which a particular packet or message belongs. In the LISP system, the segmentation attribute is referred to as the instance ID or IID. The fundamental LISP operations remain unchanged when segmentation is enabled, as the instance ID is simply an extension of the endpoint identifier. The instance ID attribute is used to scope control plane messages as well as data plane traffic into specific mapping and forwarding tables (Virtual Routing and Forwarding [VRF] or bridge domains).

The first implementation of such a demand-based switched fabric was the Cisco Application Centric Infrastructure (Cisco ACI™) fabric, in which a conversational learning mechanism was used to consult a centralized databased of endpoint locations. The next generation of switching fabrics is implemented with Internet Engineering Task Force (IETF) Standards Track protocols. For the overlay control plane, the next generation of switching fabrics uses LISP, which is the demand-based protocol with the longest deployment track record and the most advanced formalization of standards at the IETF.

The next few sections describe the key functional areas addressed by the use of demand-based protocols as the overlay control plane in modern switching fabrics. Where appropriate, we highlight the implications of using a traditional routing protocol instead.

Mobility

In a network overlay, the operation of mobility basically involves maintaining an accurate view of the location of the different endpoints. As an endpoint moves, its location is updated. For IP endpoints, this implies that reachability must be procured with host-level granularity, as prefixes are no longer aligned with locations. As endpoints move around the fabric, the underlying routing between locations remains unaffected, while there isn't a routing calculation to be completed in the overlay control plane. A demand-based overlay will require that the mappings in the central mapping database be updated, and it will also require that any active caches be updated. The number of active caches communicating with the mobile endpoint is limited and normally small. Thus, the impact of reconvergence for a move in a demand-based overlay is limited, and switches that are not involved in any communication with the moving endpoint will not be notified of this move. In the initial Cisco® Software-Defined Access (SD-Access) implementation, the departure switch is also notified of the move so that it can react to data streams from old caches and trigger the refresh of those old caches. Convergence times are subsecond and are not affected by the size of the network, as reconvergence events involve only signaling between relevant parties. In future implementations, a publish and subscribe mechanism may be used to trigger the refresh of map caches. This would reduce convergence times from hundreds of milliseconds to low-double-digit millisecond times.

In contrast, mobility in a traditional routing protocol involves the addition of mobility sequence attributes (this is one of the items included in Ethernet VPN (EVPN), along with Layer 2 family support). These sequence attributes must be flooded across the fabric, along with the new location information for the moved host. This means that a move causes a global reconvergence of the overlay control protocol, as all nodes maintain state for the endpoints and they must all be notified of the change in endpoint location, whether there is communication with the endpoint or not. Furthermore, the flood must happen twice, once in the form of a new routing update and a second time in the form of a withdrawal. If the routing update signal is lost, the mechanism does not converge. Convergence times are a function of the volume of moves and the size of the network. In addition, the full path-calculation machinery is exercised with each change and update, which requires a great deal of processing power on the part of the network devices involved (this, in turn, limits scalability).

Layer 2 semantics and blast radius control

In addition to the standard Layer 3 service that is provided by LISP in SD-Access, the SD-Access/LISP overlay implementation supports Layer 2 overlays in two modes:

- Enhanced forwarding
- Traditional forwarding

In the traditional forwarding mode, Address Resolution Protocol (ARP) requests and broadcasts are flooded across the footprint of the Layer 2 domain, with optimizations to limit the amount of ARP traffic that is forwarded and to rate-limit broadcast traffic as necessary.

The enhanced forwarding mode allows the switching fabric to present the endpoints with complete Layer 2 semantics, while treating all communications as IP routed communications. In other words, intra-subnet communication will be routed as if the hosts belonged to separate /31 networks (in the IPv4 case). However, hosts are configured within a subnet (of a larger size than a /31) just as they are today, and they are presented with responses to their ARP requests that are equivalent to what they would have seen if the LAN they are in had been stretched. Because intra-subnet communication is always routed, ARP requests and any type of Layer 2 broadcast are no longer flooded. This basically brings the benefits of a routed network down to the level of a single host, without having to manage a separate subnet for every host.

The two modes are provided in order to support broadcast-based applications if and when necessary. The enhanced forwarding mode also assumes that all hosts are IP hosts. Non-IP hosts would require the use of the traditional forwarding mode.

Enhanced forwarding is instrumental in restricting the Layer 2 blast radius and containing failures in environments that do not require broadcasting. The formalization of enhanced forwarding is done within the LISP standardization and hasn't been a topic in the EVPN standards track to date.

Scale

There are two main aspects of scale in a demand-based overlay:

- State on forwarding devices
- Fanout of sites

The state maintained on forwarding devices is significantly reduced in a demand-based model, because only the state relevant to active connections and sessions is cached. Furthermore, the cache is populated only on the forwarding devices connected to the source of the connection and nowhere else. This means that the forwarding state is distributed and is also scoped only to active connections. In an environment dealing with host-level reachability granularity, this level of state reduction is critical. As location-aware policies become more widely adopted, the demand-based control plane can be used to propagate policy attributes relevant to a particular connection. Only the attributes relevant to the connection are communicated and cached in a demand-based overlay. The communication of policy attributes in a push-based model (traditional routing protocols) would be impractical, as all combinations covering any possible flow would have to be pushed out to all network devices ahead of time. In a demand-based system, the propagation of the information could even be subject to a location-restricted policy that may be enforced at the time when mappings are provided (that is, upon request).

In terms of the fanout of sites, a demand-based mechanism doesn't have to maintain an active session or peering between the forwarders and the mapping database. Because the relationship between forwarders and the mapping system is stateless, there isn't a scalability concern with regard to the number of sites that the overlay may fan out to. In contrast, with a routing protocol in which the aspects of path liveness and topology calculation require an actively monitored session between participants in the control plane, the scale of the domain is limited to the number of TCP peerings that the processor on a node can maintain.

Multidomain scale and survivability

Although a demand protocol can support a very large number of sites in a single domain, it is desirable to divide the network into independent resiliency zones that could survive a failure in other zones. Multiple domains can be interconnected in a demand-based model, and the boundaries between the domains will be stateless, allowing minimal state to be cached at the borders of the domains and no remote or cross-domain state to be propagated into local domains. Only state for active connections is effectively cached at the borders and does not need to be propagated into the leaf domains where the endpoints connect. One important aspect of this scaling methodology is that it is effective for nonsummarizable address spaces. In a network supporting mobility, nonsummarizable address spaces are the norm, as hosts are disseminated across the fabric without regard to subnet boundaries. A system capable of scaling a flat host address space is a must in these environments. The scalability benefit also extends to Layer 2 services based on the MAC addressing space.

Extranets for shared services

In a segmented network, it is common to allow the different segments to access services in a shared segment (DNS, Dynamic Host Configuration Protocol, the Internet, management and monitoring tools, etc.).

The shared segment is often referred to as an extranet segment. The segments accessing the extranet are the extranet subscribers. Subscribers may access the extranet and the extranet may connect back to the subscribers, but one subscriber segment should not have connectivity to another. In traditional VRF-based networks, this is achieved by importing routes across VRFs in a controlled manner. Thus, the extranet VRF would import all routes for all the subscriber VRFs, and each subscriber VRF would have to import all the extranet VRF routes (excluding any subscriber routes the extranet may have imported, of course). The net result is a very large replication of routes. The extranet VRF would have to store all routes in all VRFs, and each subscriber VRF would have its own routes plus the extranet routes. This is clearly inefficient, but it doesn't stop there. In order to be able to forward traffic across these VRFs, all VRFs participating in the extranet must be present at every edge device of the overlay.

Since the IIDs used for segmentation in LISP are treated as extensions of the endpoint identifier namespace, and since forwarding information is obtained on demand, a demand mechanism such as LISP can provide an extranet service in which the forwarding state does not need to be replicated across VRFs and bridge domains, but only the strictly necessary mappings are cached in the VRFs that are involved in the active extranet connections. Furthermore, because forwarding information is obtained on demand, the destination IID can be provided to the requester, so that hopping between VRFs is done at the time of encapsulation. The result is that shared services can be delivered without having to hold the routing table (VRF) for the shared services at every access switch (fabric edge), and also without requiring the switches where the shared services connect to hold the subscriber VRFs.

The state efficiencies are significant, and the convenience of a normalized operational model for unicast and multicast within or across VRFs should not be underestimated. Once again, the IIDs are simply a further qualification of the endpoint identifier, and they can be used in very flexible ways in a demand-based system such as LISP.

Multicast

The LISP-based fabric uses the same model to support multicast as it does unicast. No reverse tree building, complex Reverse Path Forwarding (RPF) checks, or multicast protocols such as Protocol Independent Multicast (PIM) are at play in the LISP overlay network. Listeners simply are registered as part of a replication list that is associated with the multicast group address they are interested in. When a source sends traffic to the multicast group address, the encapsulating device (Ingress Tunnel Router [ITR]) will simply do a lookup for the multicast destination, just as it would do a lookup for a unicast destination. The mapping system will return either the list of locators where listeners have registered or the underlay multicast group joined by the locations where listeners have been "seen." The operator can opt into using head-end replication to forward its multicast traffic or can implement multicast in the underlay. The overlay mechanism does not change.

Because unicast and multicast behave in the same way, mobility and extranet for multicast traffic are simply inherited and do not require any additional machinery or protocols. Multicast sources can move freely and will simply trigger a fresh lookup for the multicast replication list or group when the next packet is sent to the new ITR. There is no need to notify the Egress Tunnel Routers (ETRs) where the interested listeners are connected. This is in contrast to traditional multicast mechanisms, where the trees are built from the listeners back to the source, and where mobility would imply that the listener attachment devices have to be notified of the move for the trees to be rebuilt.

Multicast extranet is achieved simply by providing the pertinent cross-segment IIDs when a source transmits and triggers a lookup from an IID. There is no need for the complex signaling models that had to be introduced with Multicast VPN (MVPN) in traditional routed VPN networks.

Wireless integration

Wireless systems may be integrated into the SD-Access fabric in one of three modes:

- Over the top
- Fabric-enabled wireless controller
- Fabric-enabled flex wireless

Over-the-top wireless simply provides IP transport to an overlay wireless solution. Wired and wireless systems and policy are managed independently. It may be possible to unify the workflows in the management plane using API calls, but the functionality in each domain will remain disparate.

The Fabric-Enabled Wireless controller (FEW) model integrates the forwarding and policy enforcement of the wireless clients with those of the wired clients. This is the preferred model, as operations as well as forwarding and policy enforcement behaviors are indeed unified for wired and wireless. In this model the wireless signaling is done over the top, but the wireless access points delegate forwarding to the access switches. To achieve this, the wireless control plane and the LISP control plane exchange information using their APIs. The wireless control plane will report the location of the different endpoints to the LISP mapping system so that endpoint-to-location mappings for the wireless access points can be created in the LISP mapping system. Wireless access points will send their traffic unconditionally to the access switch to which they connect (as if they were sending their traffic to their wireless controller), and the access switches will then forward the traffic according to the mappings created in the mapping system. From that point onward, the behavior for wired and wireless is indistinguishable. The wireless endpoints thus inherit all the benefits of scale, Layer 2 blast containment, policy enforcement, service chaining, etc. that the fabric provides to wired clients. The integration is based on simple API exchange and interception of data plane traffic. These two functions can be easily adapted to most wireless solutions in the industry that offer open APIs.

A third mode is flex mode, in which the wireless access points do not encapsulate their traffic but simply put it in a VLAN (generally with 802.1q framing). In these cases the integration is even simpler and follows a model similar to the FEW solution without requiring traffic interception. It simply maps VLANs to the appropriate IIDs in the fabric.

Achieving similar integration in a Border Gateway Protocol (BGP)-based fabric may be possible, but there will be challenges in terms of the adaptation of the BGP state machine to API-introduced updates and information. Such an implementation would also encounter general challenges involving scale and calculation of widespread updates in the face of mobility. No work has been done in this regard to date.

Contextual information

LISP can handle contextual information related to the endpoint in its registrations. Parameters such as the group an endpoint belongs to, its virtual network instance ID, and its geo-coordinates may be registered in the LISP mapping database. Additionally, when a request is issued, the place or context from which the request is issued may be used to determine the kind of response that should be sent. For example, different mappings may be given back to similar requests coming from different locations (such as domestic vs. international). Some of the use cases that are addressed by this include the following:

Security and policy

The SD-Access switching fabric implementation includes a more granular level of segmentation realized by the use of group tags. Group tags are included in the Virtual Extensible LAN (VXLAN) encapsulation and are used to make forwarding time decisions on the traffic. The tags can be used as the criteria to match an Access Control List (ACL) and take actions on permitting or denying traffic, but it is also possible to take actions related to Quality of Service (QoS) or path selection. In order to get a complete match, both source and destination group tags must be known. The VXLAN encapsulation accommodates the source tag, with the assumption that the destination tag is obtained at the egress edge of the fabric. The LISP demand mechanism provides a scalable manner of actually distributing the destination tags (along with the location mapping for the particular destination) so that group-based policies can be applied at ingress and therefore include meaningful network policies such as QoS and path selection, which must be enforced from the point of ingress onward (their enforcement cannot be deferred to the egress point). So group tags can be registered in the LISP mapping system along with the endpoint identifier. When a map request is issued for a particular endpoint, the mapping returned includes the group tag value for the destination. Thus, the ITR will have both source and destination group tag information and will be able to enforce group-based policy at ingress. Moving forward, ACLs may not even need to be instantiated in the data plane, and the enforcement of access policies could be done in the LISP control plane, where mappings can be provided, modified, or denied altogether based on the policy. This is possible in a demand model in which the control plane is involved as transactions occur and policy can be enforced for different connections.

Location-based policy

Connectivity to certain applications may be routed one way or another, depending on where the source of the connection resides. In a demand system, when the request comes from a location within an area considered trusted, we may route the traffic directly to the destination, but if a request for the same destination comes from a locator overseas or outside the firewall perimeter, the traffic may be routed through a series of inspection systems.

Programmable traffic paths

The demand-based process for procuring forwarding information allows a simple methodology for the construction of engineered paths. By introducing LISP-capable devices in the intermediate nodes, traffic can be steered through a series of forwarding nodes. At each forwarding node the response from the mapping system would steer the traffic to the next node in the path. The mapping system knows which stage of the path the traffic is in simply by the locator of the requesting encapsulating router. Routers in the middle of the path will decapsulate and reencapsulate and are referred to as Reencapsulating Tunnel Routers (RTRs). The criteria by which traffic may choose one path or another can be as rich as required. The traffic steering policy may be based

Summary

The problem being addressed in the overlay control plane is not a routing problem. There are many benefits to the use of a demand control plane on the overlay. Common networking services can be simplified and enhanced, while new services may be easily introduced to leverage the rich location information present in the network. We consider a demand-based protocol like LISP to be the most state-of-the-art, efficient, and, therefore, appropriate tool for the overlay control plane to support the scale, flexibility, and programmability required moving forward.

on rich contextual information when present in the mapping data structures. This is clearly challenging to do with a push protocol in which additional control plane protocols and mechanisms and the inclusion of metadata in the data plane become necessary in order to provide this functionality. In the case of LISP, this is simply a policy that can be imposed programmatically (via a northbound API) onto the mapping system.

Service insertion

One particularly common use case related to the ability to engineer traffic paths is network service insertion. When inserting a service node such as a firewall, it is possible to combine the traffic steering capabilities of LISP with the extranet functionality (previously described) to not only steer traffic to the nodes where a firewall may be connected, but also to ensure that the traffic is put through the firewall when it gets to the node where it is connected. All of this without adding any metadata in the data plane or requiring support for anything other than traditional IP forwarding from the firewall. This is particularly useful in scenarios where access control policy must be enforced in a stateful manner and the policy enforcement activity must be logged. For these scenarios it is interesting to be able to distribute the enforcement of the access control policy to the edges of the fabric while still using firewalls for this enforcement. The APIs in the LISP control plane allow the dynamic procurement, identification, and selection of the firewalls and present a framework for full automation of distributed firewall insertion.

C11-739593-00 09/17