November 14, 2018

PRESERVING CONTROL-PLANE PERFORMANCE AT SCALE IN ON-DEMAND FABRICS VIA PRIORITY QUEUEING

Marc Portoles Comeras

Johnson Leong

Alberto Rodriguez Natal

Balaji Pitta Venkatachalapathy

PRESERVING CONTROL-PLANE PERFORMANCE AT SCALE IN ON-DEMAND FABRICS VIA PRIORITY QUEUEING

AUTHORS:
Marc Portoles Comeras
Johnson Leong
Alberto Rodriguez Natal
Balaji Pitta Venkatachalapathy

ABSTRACT

Techniques are provided for establishing priorities between different signaling messages and using a priority queueing/scheduling mechanism that will preserve performance guarantees of a centralized control plane even in scaled up systems with high signaling load. This builds on the key observation that there can be established a direct relation between signaling mechanisms in on-demand overlay fabrics and their contribution to the convergence of the network after a network update.

DETAILED DESCRIPTION

In enterprise fabric deployments, which use Locator ID Separation Protocol (LISP) for centralized control, fabric edge devices learn network overlay routes on demand from a central fabric control plane entity. This process has traditionally implied using data-traffic on the fabric edge devices as a trigger to generate on-demand signaling and gather routing and policy information from the central control-plane entity.

However, as network fabrics evolve, the role of this central control-plane entity is extended with new signaling options used to support new functionality. Current state-of-art on-demand control plane solutions use three basic signaling procedures: registration services, map resolution services, and subscription services. These three types of signaling coexist and provide the necessary basis for all mechanisms implemented as part of the solution.

Despite the benefits that the different control-plane services provide, the coexistence of multiple signaling messages in the network can lead to unintended consequences on the performance of the system. In particular, it can be shown that systems that use LISP as their control-plane mechanism of choice can see their scalability properties compromised when subscription services coexist with traditional on-demand resolution services.

As the size of network overlay fabrics grow, convergence of the system under mobility triggers or network updates may rapidly degrade as the number of fabric elements using subscription services grows. A mechanism is needed to allow deployments to simultaneously use the different control-plane services while ensuring that the convergence of the overlay fabric is not negatively impacted.

Techniques described herein address this problem by establishing a priority scheme on the centralized control entity to handle signaling messages and provide scalable traffic and state convergence guarantees.

This builds on the key observation that different types of signaling used in on-demand based fabric overlays contribute differently to the restoration and convergence of network traffic in the presence of network changes.

As a result, a priority queuing scheme is used to establish a direct relation between the processing priority of a signaling message and its expected contribution to the convergence of network traffic.

As an example, the mechanism may prioritize control-plane messages that prevent drops on the data path, over control-plane messages that just optimize the data path. In other words, without the priority queues introduced herein, the system may face periods when the centralized control-plane entity is trying to dispatch messages to optimize a given data path while data packets in another data path are being dropped.

Techniques described herein relate to applying priority queues to optimize the control-plane signaling of on-demand based overlay fabrics. This may improve the fabric control-plane scalability of various deployments.

Overlay fabrics that use LISP as its control plane of choice, currently use signaling to support three differentiated type of procedures: (1) Map Registration and Notification, (2) Map Resolution, and (3) Map Subscription and Publication.

Techniques described herein introduce the use of a priority queue to drive the processing of the signaling messages listed above. A priority is given to each type of signaling procedure in direct relation to their expected contribution to convergence of the traffic after a network update/change.

Figure 1 below illustrates a priority scheduler that dispatches signaling tasks.

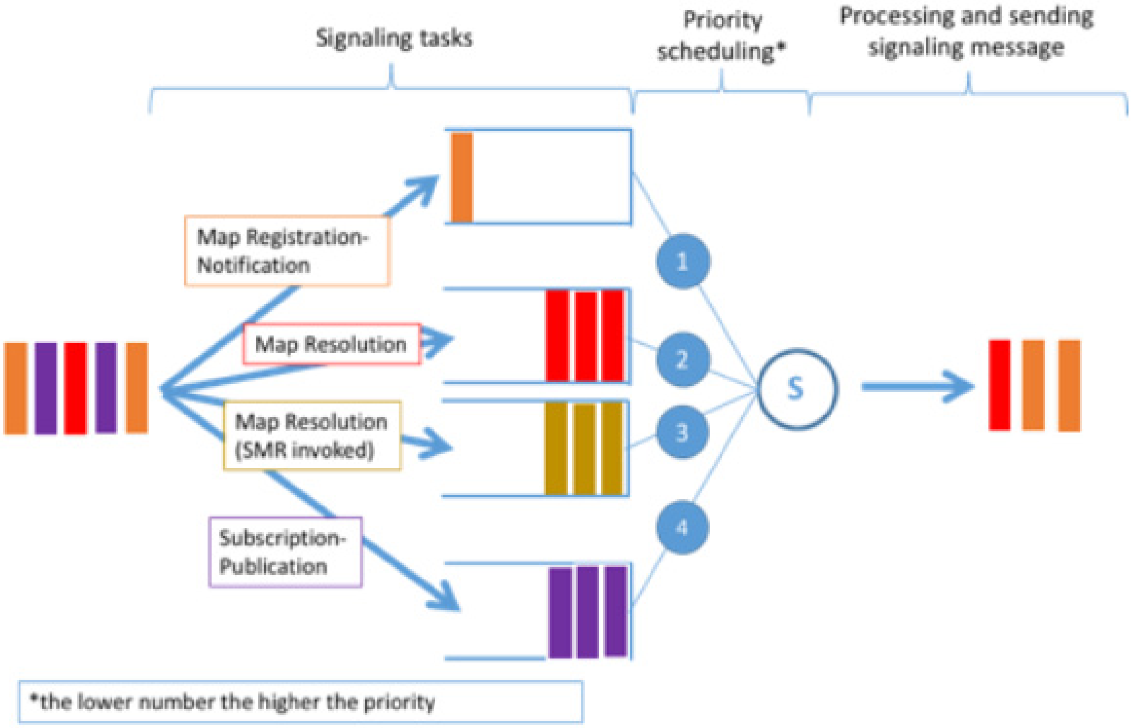
Signaling tasks
Priority scheduling*
Processing and sending signaling message
Map Registration-Notification
Map Resolution
Map Resolution (SMR invoked)
Subscription-Publication
1
2
3
4
S
*the lower number the higher the priority

Figure 1

The scheduling process may be made preemptive when possible, in order to provide an optimal time of convergence for traffic flows in the overlay fabric.

The following discussion illustrates the relation between each one of these signaling mechanisms and traffic convergence.

Figure 2 below illustrates a simplified model of an overlay datacenter network fabric topology. This model also applies to any state-of-art overlay campus network.

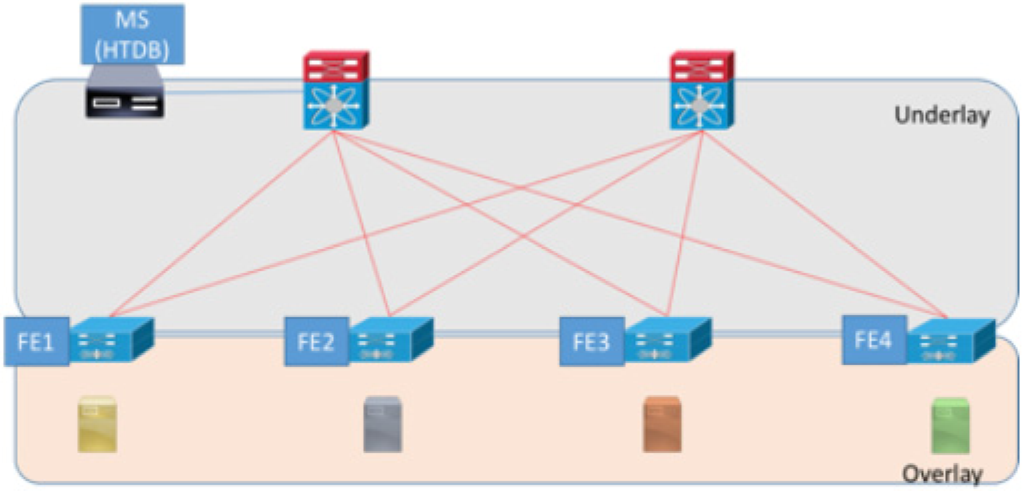

Figure 2

The network deploys a series of hosts in an overlay network which are mapped onto a routed underlay. A simplified perspective of an overlay fabric has three distinctive elements: fabric edges **(**FEx in Figure 2), end hosts (or prefixes), and a control-plane/mapping system (Host Tracking Database (HTDB) in Figure 2).

Fabric edges work as the point of interconnection (mapping) between the overlay and the underlay. Fabric edges are responsible for detecting and registering hosts as well as gathering mapping information to encapsulate traffic between host attached to different fabric edges.

End hosts are the users of the overlay network infrastructure. End hosts use fabric edges to access the network and, in many cases, can roam around the network, between different fabric edges.

The control-plane/mapping system stores the relation between hosts (or prefixes) and the fabric edge to which they are connected. In essence the mapping system stores overlay to underlay mapping information.

Described are the signaling mechanisms that are part of on-demand fabric overlays in terms of their contribution (cost/delay) to the convergence of traffic after a network update.

Figure 3 below illustrates how network events can be seen as arrivals to a centralized queuing system that needs to process and notify the rest of elements in the network.

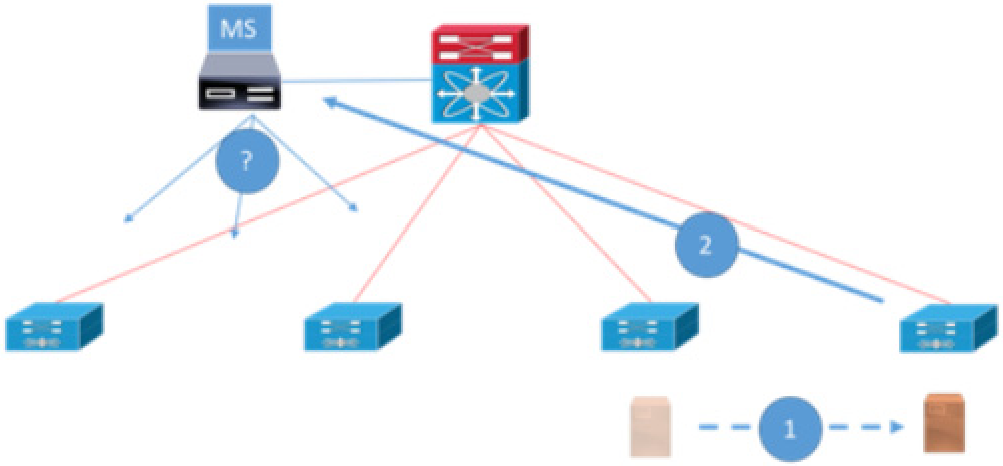

Figure 3

When using centralized control and management solutions, network updates may be seen as arrivals to a centralized queueing system. The cost of each signaling process can be understood as the cost of processing that arrival to the queue. This cost depends on the number of elements that need to be updated to guarantee traffic convergence.

When (1) a host moves between two fabric nodes, (2) the destination fabric node sends an update to the centralized control entity. The cost of each signaling process is modeled depending the amount of fabric elements that need to be notified about the update in order to guarantee that end-to-end traffic flow is restored.

Figure 4 below illustrates how restoring traffic after a host moves using map-notifications has a cost of one message to the previous fabric node where the host used to reside.

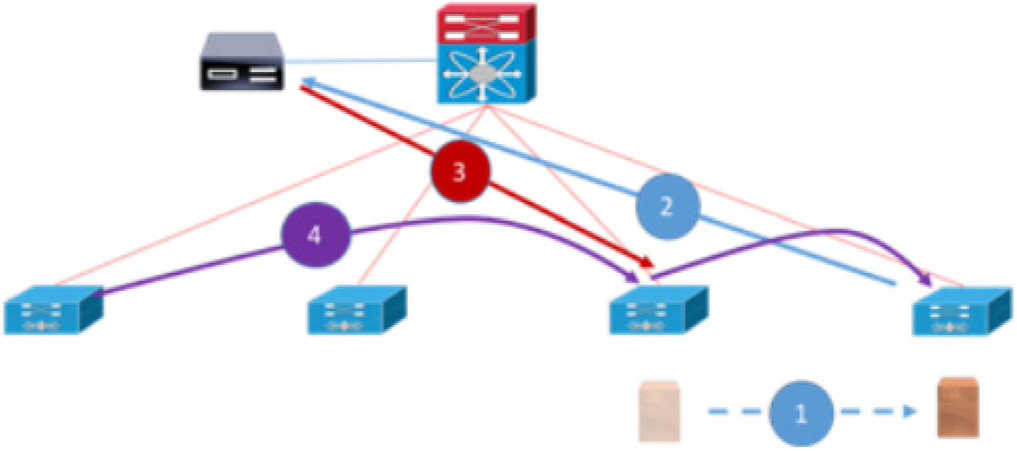

Figure 4

In a map registration/notification process, when the centralized control-plane entity receives the network update (2), it sends a notification to the previous fabric node where the host that moved used to reside (3). This notification carries full routing information with the new location of the host and the receiving fabric node can install a forwarding entry pointing to the new location.

As soon as this new forwarding route is installed, end-to-end traffic is restored for any existing flow (4). In this case packets may (temporarily) follow a suboptimal path, but they will reach their final destination and communication will be restored.

The signaling cost of this update is the cost of sending a single message to the previous fabric node.

On-demand route resolution in overlay fabrics is commonly triggered by data traffic. When a host needs to send data to a given destination, the fabric elements involved react to dynamically discover the location of the destination and program the path.

Figure 5 below illustrates the use of LISP Solicit Map Requests (SMRs) and map-requests to selectively refresh routers that are actively using the routing path that has been updated.

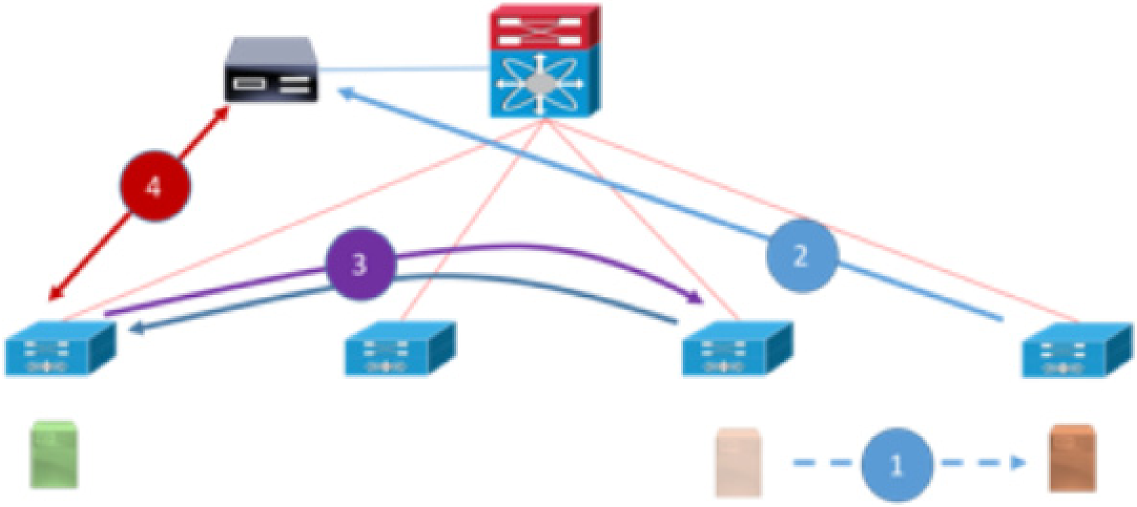

Figure 5

Figure 5 above shows the cost of on-demand route updates. After a host has moved in a fabric and the centralized control-plane is updated, traffic may still be sent to the old location (3). In this case LISP defines the use of the SMR by which a fabric edge router can send a message to another one indicating that traffic is being sent to the wrong location. This triggers (4) a selective route refresh, the map resolution process.

Once completed, the fabric node that has received the on-demand update completes the route learning and can ensure that traffic will reach the destination.

The advantage of this approach to route resolution is that route updates are selectively downloaded where they are needed. As a consequence, the signaling cost associated with a routing update is directly proportional to the popularity (in terms of use) of the route that is updated. The cost of this process increases with popularity and reduces with locality of traffic. However, after each signaling update traffic is guaranteed to follow the optimal path to the destination.

SMR triggered map-requests can be treated with lower priority than traffic on-boarding map-requests since, under stress, the system may default to guaranteeing end-to-end traffic delivery over optimal path use. This is reflected in the queueing illustration depicted above.

Finally, state-of-art overlay networks with an on-demand control-plane are extended to support subscriptions services. In this case fabric elements can subscribe to receive certain network updates regardless of whether they are actively using the routes/paths associated with these network updates.

Figure 6 below illustrates how the centralized control plane needs to notify every fabric element in the network about a route update (host move) following a subscription service.

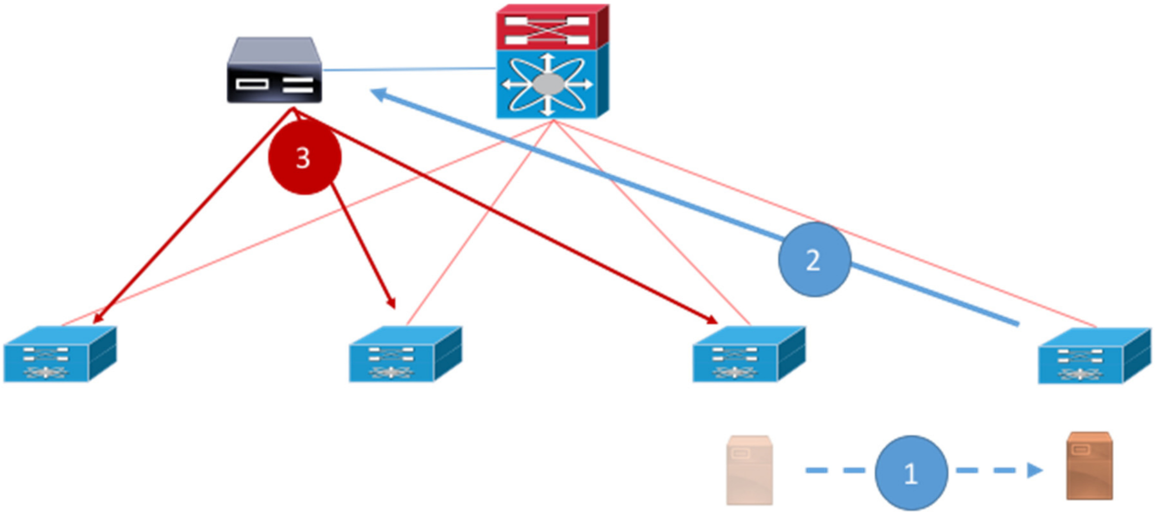

Figure 6

Once the centralized control plane receives the network update it must send a signaling message (3) to all fabric elements that have subscribed to receive this update.

Signaling associated with subscription services does not follow traffic triggers. As a consequence, in order to guarantee routing convergence using subscription associated updates, the centralized controller has to update every fabric element that has subscribed to receive network updates. In the worst case every fabric element needs to receive an update to guarantee traffic delivery.

As a result, the signaling cost to provide traffic convergence increases with the number of routers that are subscribed to a given prefix, regardless of use.

A queueing model is used to illustrate the advantages of using these techniques to provide performance guarantees in overlay network fabrics.

As illustrated in Figure 7 below, the centralized control plane is modeled as a single queue and single server queue, where network updates are treated as arrivals and each one of the three groups of signaling mechanisms are modeled as a processing cost associated to each one of the arrivals.

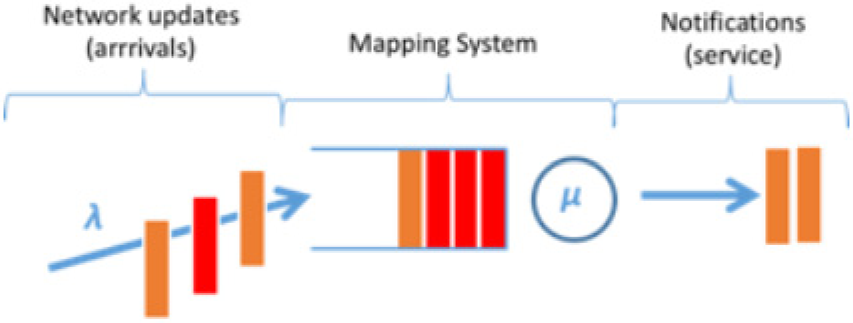

Figure 7

This model has been used to analyze the cost of processing the different signaling processes for a network overlay of 200 fabric elements (network edge routers), high prefix popularity (each fabric element is actively forwarding traffic to a 30% of the prefix destinations in the network) and where the centralized controller is able to process and generate approximately 5000 notifications per second.

Figure 8 below illustrates the cost of network updates on a centralized control plane depending on the signaling process that is used.

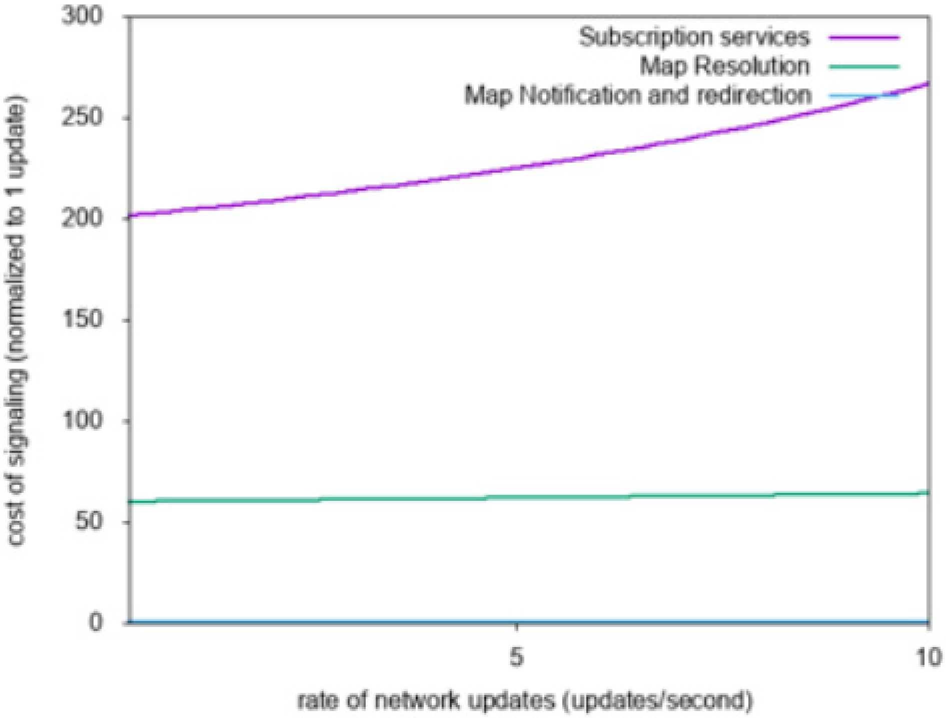

Figure 8

Figure 8 above shows the advantages of the present techniques. As the number of network updates increases, the amount of load that the server must be able to support to provide traffic convergence guarantees largely differs depending on the signaling process under consideration. Map-notifies (with redirection) achieve network convergence with minimal cost, map-resolution (triggered by first packet or SMR) incurs higher costs than map-notifies, but even at this level of prefix destination popularity (30%) its cost is considerably lower than the cost of convergence associated with subscription services.

Relating signaling priority to contribution to traffic convergence ensures that as the network grows, the network can provide guarantees of traffic convergence regardless of the amount of load that the centralized control plane needs to process.

In summary, techniques are provided for establishing priorities between different signaling messages and using a priority queueing/scheduling mechanism that will preserve performance guarantees of a centralized control plane even in scaled up systems with high signaling load. This builds on the key observation that there can be established a direct relation between signaling mechanisms in on-demand overlay fabrics and their contribution to the convergence of the network after a network update.