Routing Area Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 8, 2019

J. Heitz
K. Majumdar
A. Lindem
Cisco
November 4, 2018

Automatic discovery and configuration of the network fabric in Massive Scale Data Centers
draft-heitz-idr-msdc-fabric-autoconf-01

Abstract

A switching fabric in a massive scale data center can comprise many 10,000's of switches and 100,000's of IP hosts. To connect and configure a network of such size needs automation to avoid errors. Zero Touch Provisioning (ZTP) protocols exist. These can configure IP devices that are reachable by the ZTP agents. A method to combine BGP, DHCPv6 and SRv6 with ZTP that can be used to discover and configure an entire network of devices is described. It is designed to scale well, because each networked device is not required to know about more than its directly connected neighborhood.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on May 8, 2019.

Table of Contents

1. Introduction 2
2. Requirements 3
3. Solution Overview 4
4. Solution Details 5
5. DHCP Procedures 6
5.1. Inconsistent Endpoints 7
6. Link State Database 7
7. BGP Procedures 8
8. Segment Routing Procedures 9
9. Final Configuration 9
10. Connecting a New Controller to a Network in Production . . . 10
11. Multiple Controllers 10
12. Security Considerations 10
13. IANA Considerations 11
14. Acknowldgements 11
15. References 11
15.1. Normative References 11
15.2. Informative References 12
Authors' Addresses 13

1. Introduction

RFC7938 [RFC7938] defines a massive scale data center as one that contains over one hundred thousand servers. It describes the advantages of using BGP [RFC4271] as a routing protocol in a Clos switching fabric that connects these servers. A fabric design that scales to one million servers is considered enough for the forseeable future and is the design goal of this document. Of course, the design should also work for smaller fabrics. A switch fabric to connect one million servers will consist of between 35000 and 130000 switches and 1.5 million to 8 million links, depending on how

redundantly the servers are connected to the fabric and the level of oversubscription in the fabric. A switch that needs to store, send and operate on hundreds of routes is clearly cheaper than one that needs to store, send and operate on millions of links.

Such a network requires significant configuration on each switch and many cables to connect. This is an onerous task without automation.

2. Requirements

To configure a fabric network for massive scale data centers.

To detect every cabling error. For example, a spine switch that has a different number of links into one pod than into another pod in a Clos fabric.

Any devices should be interchangeable with another device of equivalent functionality without requiring configuration changes. That means if a device breaks, it can be replaced by any other device of equivalent functionality without any changes to its configuration. Even if a replacement device already has configuration, it should still work in its new position.

A device may have configuration, but such configuration MUST NOT depend on the location of the device in the network. Therefore, no IP addresses should be pre-configured on any devices. No fabric tier should be needed.

For scalability, every device must not need to know how to reach every other device. Only the controller should be expected to know the entire topology.

If two such auto-discovering/auto-configuring networks are connected together, the function of discovery/configuration in one network must not disturb this function in the other network.

Separate cabling for a management network must not be required.

The network should function even if the controllers are disconnected. The controller should only be needed to discover and configure devices to the network. Device and link failures and restoration should not require the controller. If a device is moved or reconnected in a way that requires reconfiguration, then the controller is required to discover the new topology and to change the configuration accordingly.

The protocol does not need to be fast.

The controller must be able to reach any device if there is any way at all to reach it, even if that is multiple hops between spine switches or any other path that may be disallowed in a normal Clos network. At the same time, normal traffic must remain restricted to allowable paths.

The routing protocol for normal traffic must be fast and efficient.

The network must scale to 1 million connected servers and 8 million links in the fabric.

3. Solution Overview

DHCPv6 [RFC3315] and ZTP are used to discover and configure devices reachable by the controller. As the controller configures devices, it configures them to be DHCP relay agents. This makes more devices reachable by the new DHCP relay agents, allowing the new devices to be configured. As this configuration process proceeds further away from the controller, it configures BGP to ensure reachabillity to all devices even if links were to fail. For scalability, each device knows only its directly connected neighbors and a route to the controller. Every device can send a packet to the controller, because every device knows a route to the controller. To send a packet from the controller to a specific target device is harder, because the devices between the controller and the target do not know how to reach the target device. The controller is the only device that knows the topology between itself and the devices it needs to reach. To send a packet to a target device, the controller builds an SRv6 (Segment Routing v6) segment list. As each device receives a packet, it will place the next segment IPv6 address into the destination IPv6 address field and forward the packet to the next device.

After the network discovery is complete, the controller will validate the discovered topology against an internal description and go back and configure application dependent state into the devices and/or report connection anomalies. An example description might be "Clos fabric connecting servers and DCI pods". Since a Clos fabric looks the same upside down, the controller needs to identify servers, switches and DCI routers. This is done with DHCP vendor class options.

In certain environments, it is required for devices to authenticate the network and for the network to authenticate devices. TCP-AO [RFC5925] can be used to authenticate BGP sessions. SZTP [I-D.ietf-netconf-zerotouch] provides for authentication during the ZTP process. Netconf can be used over SSH as described in [RFC6242].

4. Solution Details

Each device needs a unique identifier. This may be printed on the device. For easy servicability, a device must have a single identifier, visible on the outside of the device and by the controller. This will be the DUID in the DHCPv6 Client Identifier Option.

In order to discover the topology, the controller needs to know every link in the topology. This means the device ID and interface ID or interface address at each end of every link. DHCPv6 can be used to obtain that information. For each link, one end of the link is the device that requests an address. The other end of the link is either the controller itself or a DHCP relay agent. The DHCP relay agent relays all client requests back to the controller.

Configuration proceeds in waves. The wave of configuration propagates away from the controller. In the first wave, a controller allocates a routable ipv6 address to each device directly connected to the controller. These devices comprise the first wave. The controller will then configure each of these devices using a ZTP protocol, such as [I-D.ietf-netconf-zerotouch]. The configuration for each device will include the following items:

- A routable Ipv6 address for each of its interfaces that have not already acquired one by DHCP.

- A routable Ipv6 address for the loopback interface.

- Configuration to act as a DHCPv6 relay agent for the next wave of devices.

- Configuration for a BGP session to each of its connected neighbors. That BGP session will initially be down, but will establish once the neighbors are connected and configured. These sessions are single hop directly connected EBGP sessions.

- Configuration for a BGP session to the controller. This is a multi-hop EBGP session using the loopback address.

Each BGP speaker requires an AS number and a router ID. The controller should allocate a different BGP AS number for each device. There are plenty of private 4-octet ASNs available. The value of the router ID is not important.

After the first wave of devices is configured, these devices become DHCPv6 relay agents. They are now in a position to accept DHCPv6 SOLICIT messages and relay them to the controller. The controller

acts as the DHCPv6 server. As each wave is configured, the BGP sessions on each device ensure that every device has a route to the controller. In this way, each DHCPv6 relay agent can communicate with the controller. A DHCP packet relayed by a device in the second wave is not relayed again by a device in the first wave. The device in the second wave has an IP connection to the controller through which it relays the messages.

The controller will allocate a different IP address for each interface for each device in the network. When the controller receives DHCP requests from DHCP relay agents, it will recognize the DHCP relay agent end of the link from the link-address field in the relay-forward message. The controller will note the DUID in the DHCP request to keep track of the device making the request. Because it already knows the DUID of the DHCP relay agent from its IP address, it can tie the two devices together by their DUID.

The controller must keep track of the DUID in every DHCP request, so that it can recognize different interfaces on the same device. This is needed to detect looped cables and to prevent the controller attempting to use ZTP to configure a single device through multiple links at the same time.

5. DHCP Procedures

When a switch acquires an IP address on an interface, it starts sending IPv6 Router Advertisements on that interface. It includes the IP address prefix in the Prefix Information Option in the Router Advertisement. The L bit MUST be set and the A bit MUST be clear. If the switch has been configured as a DHCP relay and has a BGP route to the controller, then it will set the M bit in the Router Advertisement, otherwise it clears both the M and O bits.

If a device requires an IP address on an interface and it hears a Router Advertisement with the M bit set, it will send a DHCPv6 SOLICIT message to request an IP address. Any SOLICIT message sent must include the following items:

- Client Identifier Option with the DUID.

- User Class Option to indicate the name of the network it is attempting to join. This is to prevent the controller from configuring devices attached to the network that are not part of the network to be configured.

- Vendor Class Option to indicate the type of device.

- If the link is point-to-point, then the Rapid Commit Option.

- A single Identity Association Option. This option must be for a non-temporary address and must be for the address of the interface on which it is being sent. This allows the controller to learn the interface on which the DHCP client is sending the SOLICIT message.

When a DHCP Relay Agent receives a SOLICIT message, it encapsulates it into a relay-forward message and sends it to the controller. It puts its loopback IP address into the source IP address field in the IP header of the packet.

5.1. Inconsistent Endpoints

Two endpoints of a link may have different IP address prefixes that do not overlap. This prevents IP forwarding on the link. The controller will never assign prefixes this way. This condition may occur in the following cases:

- The controller assigned addresses to interfaces on two devices via ZTP and it did not know that these devices had a link between them. This is a normal occurrance.

- Some cables were unplugged from a device under maintenance and then plugged back in in a different way.

- A device was removed from its location in a topology and replaced in another location without having its configuration erased.

The controller can repair all these cases automatically.

If a device has an IP address on an interface and it hears a Router Advertisement that includes a Prefix Information Option, the prefix of which is different to its own prefix, then the following applies. If the Router Advertisement does not have the M bit set, then the device does nothing further. The interface will not be able to send IP packets. If the Router Advertisement has the M bit set, then it will send a DHCPv6 SOLICIT message to get a new IP address. Both sides of a link may do this and the SOLICIT messages will cross. The controller will receive both of them. When it receives the second SOLICIT, it will recognize it as being from the other end of the same link and allocate the appropriate address.

6. Link State Database

The controller will maintain a link state database of each link it learns. This is conceptual and implementations may differ.

First is the device table. Each device is associated with:

- DHCP DUID. The controller learns this from the DHCP SOLICIT message received from the device.

- Device type. This is learnt from the DHCP Vendor Class Option from the DHCP SOLICIT message received from the device. It is used to recognize the topology and match it with the description of the required topology after the complete topology is discovered.

- Loopback IP address. The controller assigns this to the device during ZTP. It is advertised to BGP sessions to neighboring devices. When those neighors receive it, they advertise it to the controller and install it. They do not advertise it to other neighbors. This address is used as the endpoint for the BGP connection between the device and the controller. When the device is acting as DHCP Relay Agent, this address appears in the source IP address field in the IP header in the relay-forward message.

Next is the endpoint table. Each endpoint is associated with:

- Reference to the device hosting this endpoint.

- IAID. The controller learns this from the DHCP SOLICIT message received from the device.

- Reference to the endpoint at the other end of the link if there is one.

- Local IP address with prefix length. The controller assigns this address either in a DHCP REPLY message or during ZTP. When the device is acting as DHCP Relay Agent, this address appears in the link-address field in the relay-forward message. This is used as the endpoint of a BGP session to the neighboring device. The host address (/128) is advertised as a network address to the BGP session across the link of this endpoint. When that neighbor receives the route, it will not install the route, but advertise it to the controller only. The controller uses that route, or rather the lack of the route, to know when the link has failed. The controller knows that the link exists from the DHCP SOLICIT message.

7. BGP Procedures

The controller will advertise its own loopback address to all the directly connected BGP neighbors with a community to identify it as the controller address. This IP address will be advertised by all

devices to their directly connected BGP neighbors. The devices will use this BGP route to forward packets to the controller.

Each device will announce its interface addresses to the BGP connections of its directly connected neighbors tagged with a community. These routes will be re-announced only to the BGP session to the controller and not to directly connected neighbors. The BGP connections can be made to fail upon interface down or BFD down. BFD should only operate on the BGP sessions to directly connected neighbors, not on the session to the controller.

The controller will host one multihop BGP session with every device in the network. This is a lot of sessions. These sessions do not need to be fast. They should have long keepalive timers.

8. Segment Routing Procedures

The devices will be segment-routing V6 (SRv6) [I-D.ietf-6man-segment-routing-header] capable. When a device receives an Ipv6 packet with its own address in the destination IP address field in the IP address header and there is an SRv6 extension header with more segments, then the device will place the next segment into the destination IP address field and forward the packet to this destination. If a device cannot replace the destination IP address from the SID list in the forwarding hardware, it can punt the packet to the control plane and do it there.

The controller, knowing the topology, will be able to send a packet to any device in the network by building the appropriate SRv6 SID list. Thus each device in the network does not need to store a route for every other device.

9. Final Configuration

Once the controller has learnt the complete network topology, or at least a large recognizable part of it, it can complete the configuration of the network. This depends on the network. The controller will be programmed with a description of the expected network and applicable constraints. As discovery proceeds, the controller will try to match the discovered topology with the programmed description. An example of a data center description is: "A number of pods. Each pod consists of 384 TORs and 32 spines. Each TOR has 32 south facing ports and 32 north facing ports. Each spine has 384 south facing ports and 192 north facing ports. Super-spines connect the pods. Some of the pods are DCI pods. The devices need aggregatable addresses and BGP sessions." The controller should be able to recognize all the switches, the servers and the DCI routers and match the discovered topology to the description. It

should then create configurations for all the devices and report inconsistencies. How the controller does this is out of scope of this document.

When a new device joins the network, the controller will detect it, because it will receive a DHCP request from it, relayed by its neighboring DHCP relay agent.

10. Connecting a New Controller to a Network in Production

A network can function without the controller present. The controller is only needed to auto-configure the network when topology changes occur. If a new controller is connected to a network that is already in production, then the controller has to discover the network before it can do anything else. The controller connects to a switch using the link-local address. The controller then uses Netconf to query the configuration of the switch.

11. Multiple Controllers

Because the controller need only be present to automate configuration changes, its absence is not likely to cause a network outage. If a device interface is incorrectly connected, then it will just not come up. Thus multiple controllers are not required for redundancy. A single controller can be connected to multiple devices in the network in such a way that unreachability of large parts of the network is unlikely even with many failures within the network.

Nonetheless, multiple controllers should be possible in a single network if they coordinate control amongst each other. Such coordination is out of scope of this document.

12. Security Considerations

When the network to be configured is used as an underlay, then it is only used to connect tunnel endpoints together within the network. The network is not accessible from outside the network. The network is accessible to directly connected devices. An adversary can connect directly to a device in the network by being plugged into a port of that device. This and all other threats listed in this section can be avoided by physical barriers to prevent access to the switching hardware.

An adversary could inject or intercept packets into tunnels that are being carried by the fabric. This can be avoided by using IPSEC tunnels for all payload traffic.

An adversary could impersonate a controller and start a netconf session. To avoid that, the real controller should use netconf over ssh to all devices.

An adversary connected to a device in the network could send a DHCP SOLICIT message and get an IP address. It can then start a BGP session with the device it connects to. To avoid the BGP session, TCP-AO is recommended.

An adversary connected to a device in the network could impersonate the controller and cause the device to request DHCP services from the adversary. To avoid damage, all DHCP services other than what are required to implement the functionality of this document should be disabled. DHCP Relay agents may use DHCP message authentication as specified in [RFC3315]. DHCP delayed authentication has been deprecated, because of operational complexity in managing shared secret keys. Alternative methods using asymmetric keys are specified in [E-DHCP] and [S-DHCP6].

An adversary that has access to the network could disrupt BGP sessions running in the network. To avoid that, TCP-AO is recommended for the BGP sessions.

13. IANA Considerations

TBD

14. Acknowldgements

The careful review and helpful suggestions of the following people significantly steered the direction of this document:

Dhananjaya Rao

Bernie Volz

Robert Raszuk

15. References

15.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC3315] Droms, R., Ed., Bound, J., Volz, B., Lemon, T., Perkins, C., and M. Carney, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6)", RFC 3315, DOI 10.17487/RFC3315, July 2003, <https://www.rfc-editor.org/info/rfc3315>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <https://www.rfc-editor.org/info/rfc5925>.

[RFC6242] Wasserman, M., "Using the NETCONF Protocol over Secure Shell (SSH)", RFC 6242, DOI 10.17487/RFC6242, June 2011, <https://www.rfc-editor.org/info/rfc6242>.

15.2. Informative References

[E-DHCP] Demerjian, J. and A. Serhrouchni, "DHCP Authentication Using Certificates", 2004, <https://link.springer.com/content/pdf/10.1007%2F1-4020-8143-X_30.pdf>.

[I-D.ietf-6man-segment-routing-header] Filsfils, C., Previdi, S., Leddy, J., Matsushima, S., and d. daniel.voyer@bell.ca, "IPv6 Segment Routing Header (SRH)", draft-ietf-6man-segment-routing-header-15 (work in progress), October 2018.

[I-D.ietf-netconf-zerotouch] Watsen, K., Abrahamsson, M., and I. Farrer, "Zero Touch Provisioning for Networking Devices", draft-ietf-netconf-zerotouch-25 (work in progress), September 2018.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

[S-DHCP6] Su, Z., Ma, H., Zhang, X., and B. Zhang, "Secure DHCPv6 that uses RSA authentication integrated with Self-Certified Address", 2011, <https://ieeexplore.ieee.org/abstract/document/6058569>.

Authors' Addresses

Jakob Heitz
Cisco
170 West Tasman Drive
San Jose, CA, CA 95134
USA

Email: jheitz@cisco.com

Kausik Majumdar
Cisco
170 West Tasman Drive
San Jose, CA, CA 95134
USA

Email: kmajumda@cisco.com

Acee Lindem
Cisco
301 Midenhall Way
Cary, NC 27513
USA

Email: acee@cisco.com