Automatic Discovery and Configuration of MSDC Fabric
draft-heitz-idr-msdc-fabric-autoconf-01

IETF 103, November 2018

Bangkok

Jakob Heitz (Cisco)

Kausik Majumdar (Cisco)

Acee Lindem (Cisco)

Requirements

- To configure MSDC with 1 million servers and 8 million links.
- No location dependent configuration.
- For scalability, no device must need complete topology information.
- Separate cabling for a management network must not be required.
- To detect every cabling error.
- Auto-Configuration should not bleed into an adjacent network.
- The network should function even if the controller is disconnected.

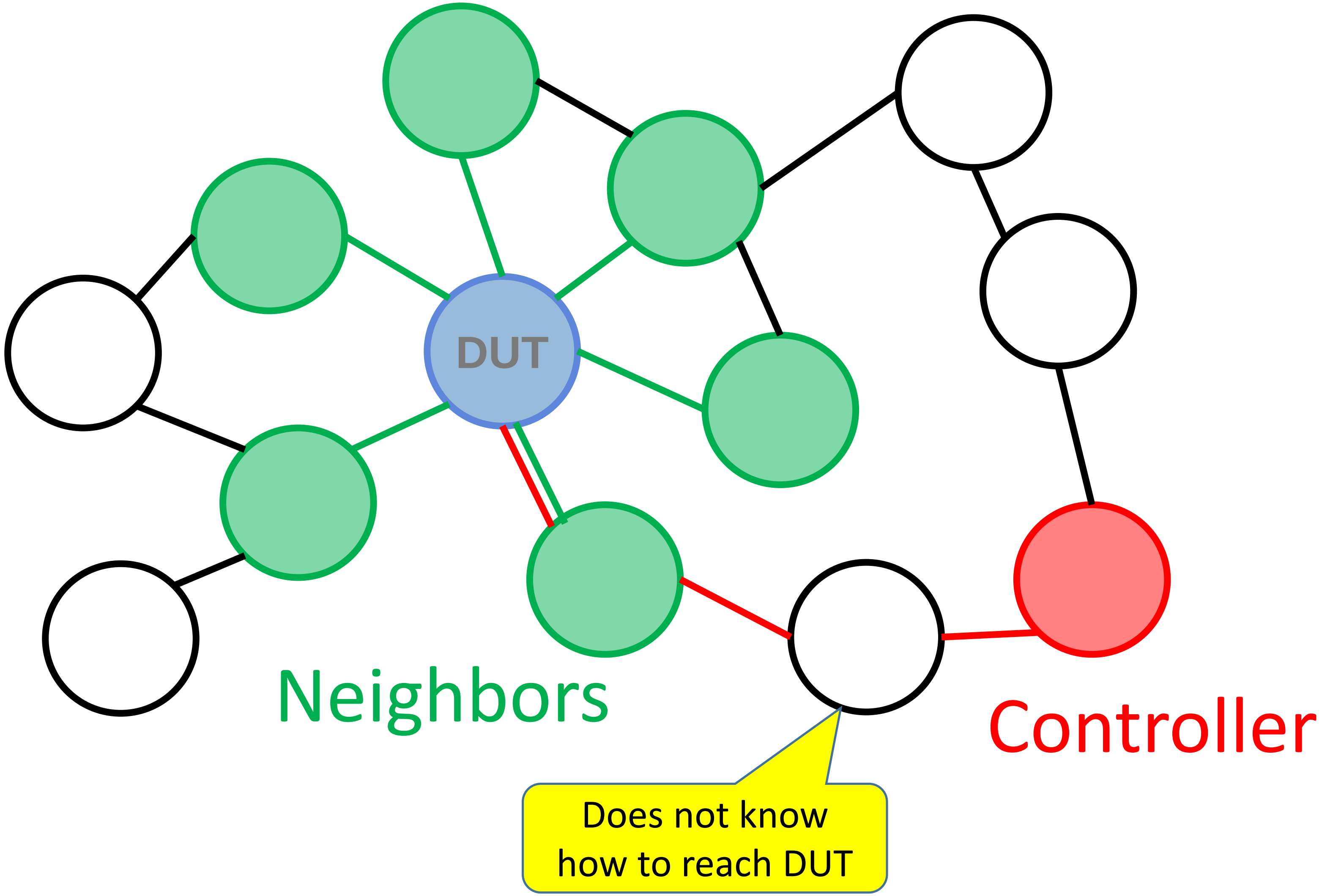
Scale

- Number of links in Clos fabric scales in the same order as the number of connected servers.
- SPF computation time scales in a higher order than the number of links.
- Data centers with a few 100,000 servers exist or are planned. A design goal to connect 1 million servers is enough for the foreseeable future
- Maximum requirement is 8 million links and 130,000 switches.
- BGP with route aggregation can do that with only 100’s of routes.
- BGP sessions and aggregatable addresses need configuration.
- Best way to auto-config is with controller that can see all.

Solution Overview

- Controller uses DHCPv6 to discover links and assign them IP addresses.
- Controller uses ZTP to complete discovery and config of each device.
- Each switch becomes a DHCP relay agent and discovers further away devices.
- Every link has a single hop BGP session.
- Single hop BGP sessions between devices distribute the controller address.
- Single hop BGP sessions between devices to learn connected neighbors.
- Controller has multihop BGP session to devices to learn link failures.
- Devices know how to reach the controller but do not know how to reach distant devices.
- Controller uses SRv6 to reach distant devices.
- Finally, the controller compares learnt topology with requirements and applies application dependent config to all devices.

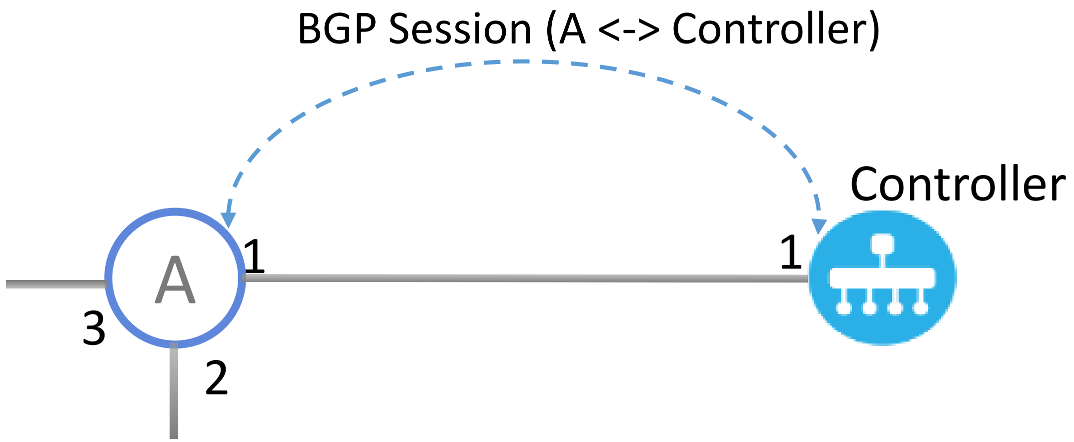
BGP



Each device knows only how to reach its neighbors and the controller

The controller learns the topology. Then uses IPv6 segment routing to reach devices.

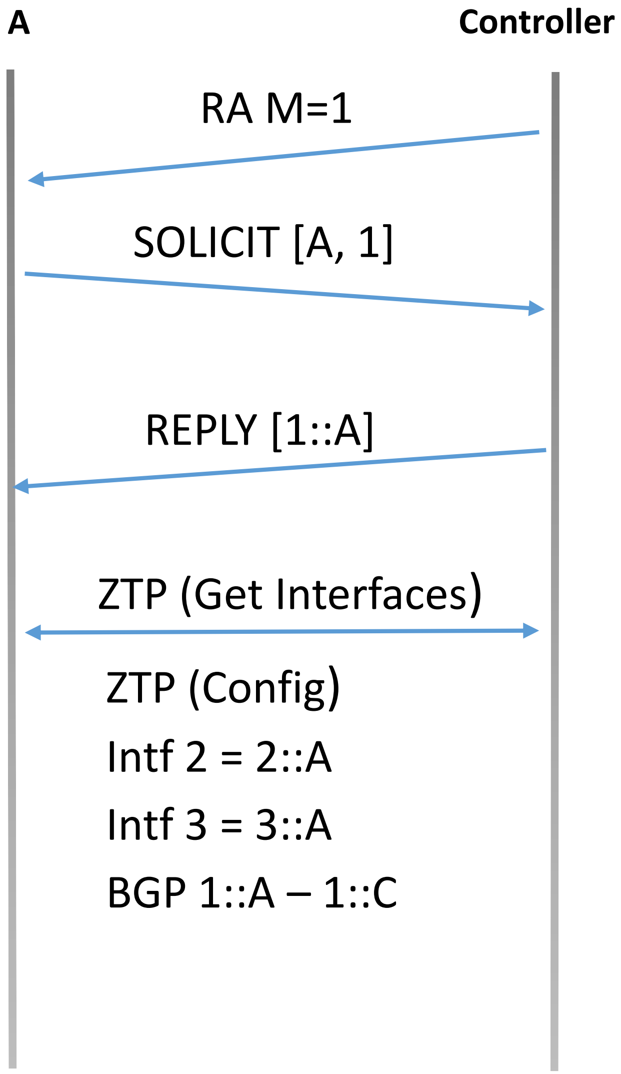
Device Directly Connected with Controller



BGP Session (A <-> Controller)

Network Endpoint Table

Index	Device	IA-ID	IP	Conn EP
1	C	1	1::C	2
2	A	1	1::A	1
3	A	2	2::A	
4	A	3	3::A	



RA M=1

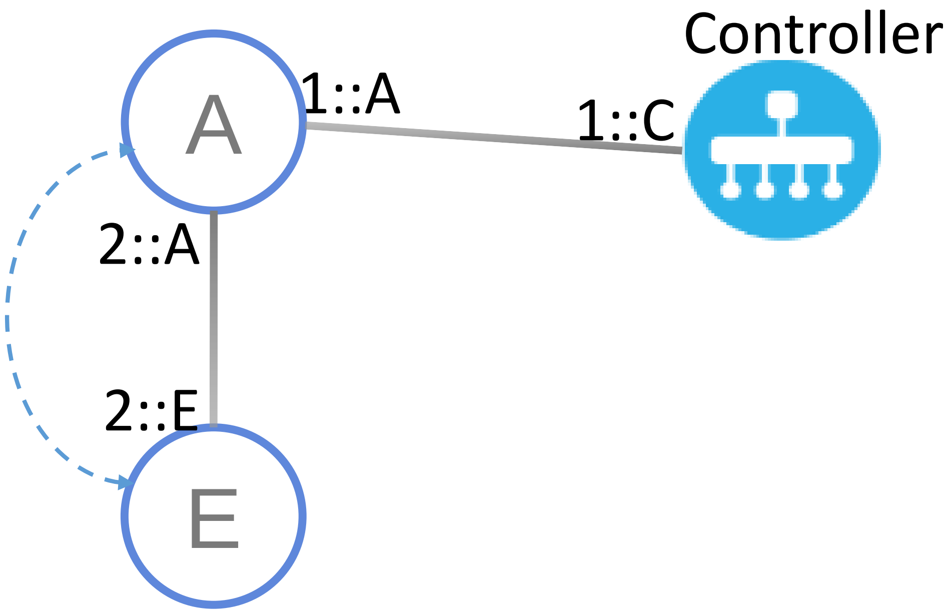
SOLICIT [A, 1]

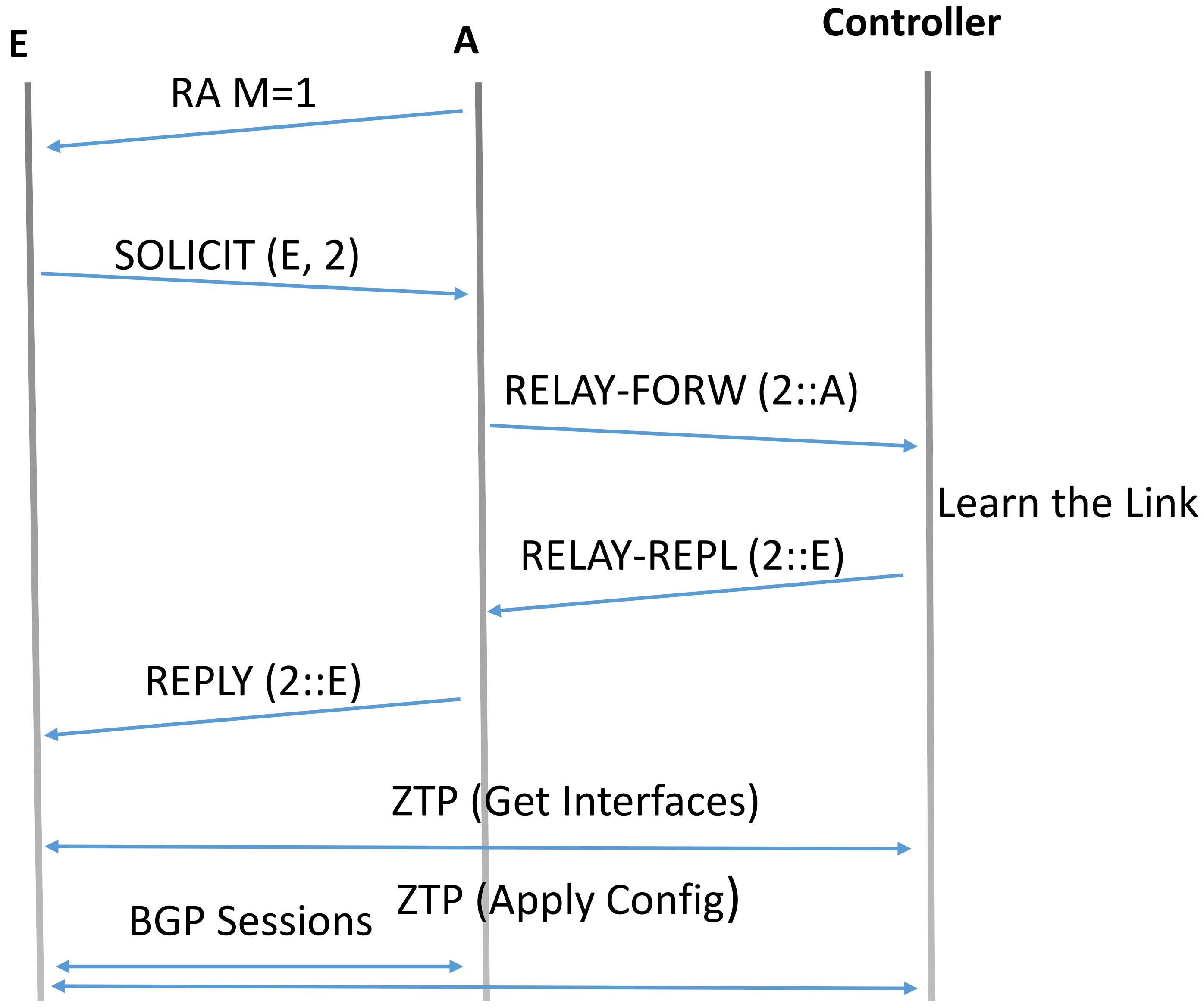
REPLY [1::A]

ZTP (Get Interfaces)

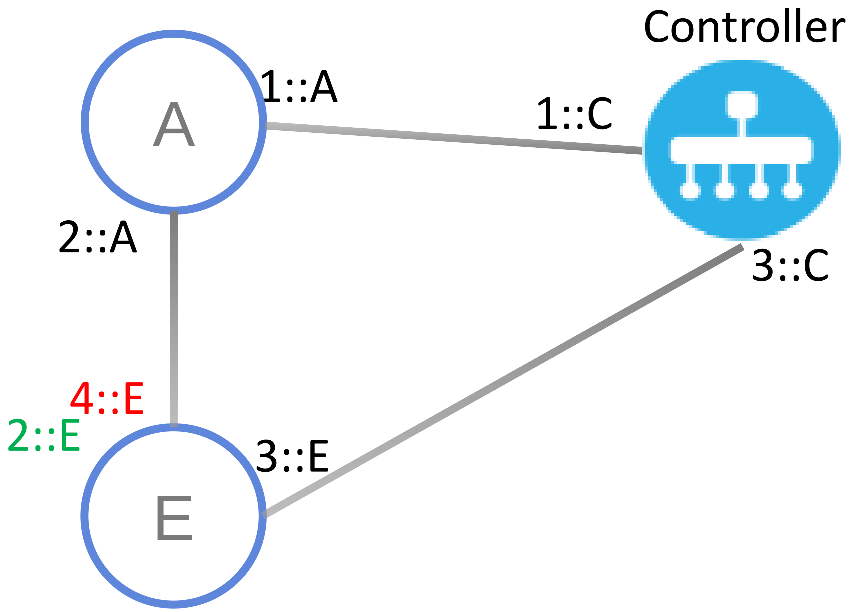
ZTP (Config)
Intf 2 = 2::A
Intf 3 = 3::A
BGP 1::A – 1::C

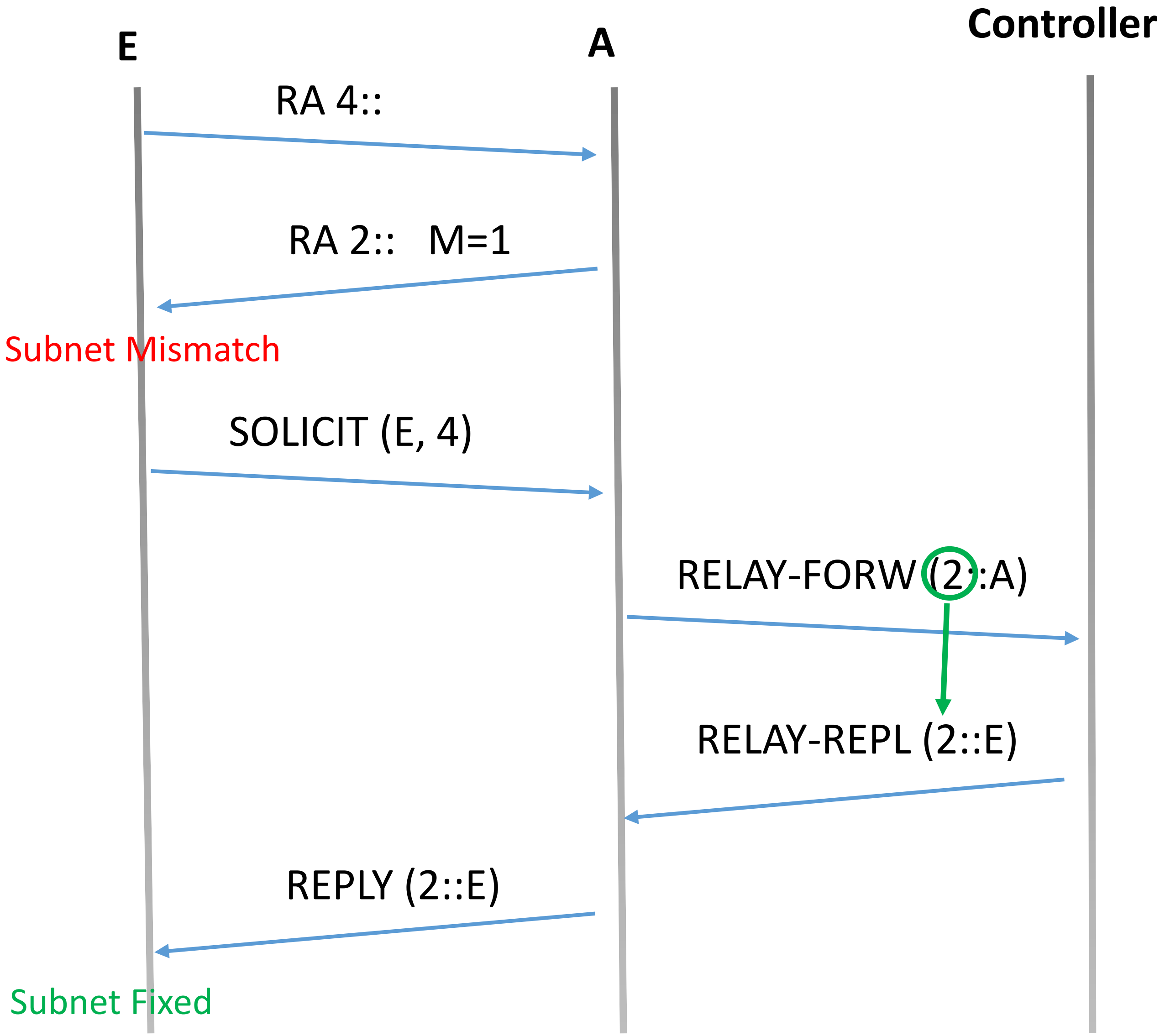
Device Not Directly Connected with Controller





Connected Link Subnet Mismatch





Security

- Fabric underlay can only be accessed by directly connected devices.
- Use IPSEC for all payload tunnels.
- draft-ietf-netconf-zerotouch-25
- RFC5925: TCP-AO
- RFC6242: Netconf over ssh
- https://ieeexplore.ieee.org/abstract/document/6058569: Secure DHCPv6 using RSA