Data Centre Routing

Adrian Farrel

Shraddha Hegde

Connections-2018

Topics

- Background
  - History and problems
  - Common DC topologies
  - The IP fabric
  - Underlay versus Overlay
  - Problems in DC topologies
- Solution space
  - E-BGP/I-BGP
  - RIFT
  - Flooding Reduction in IGPs
  - LSVR
  - EVPN Overlay

Data Centre History an Problems

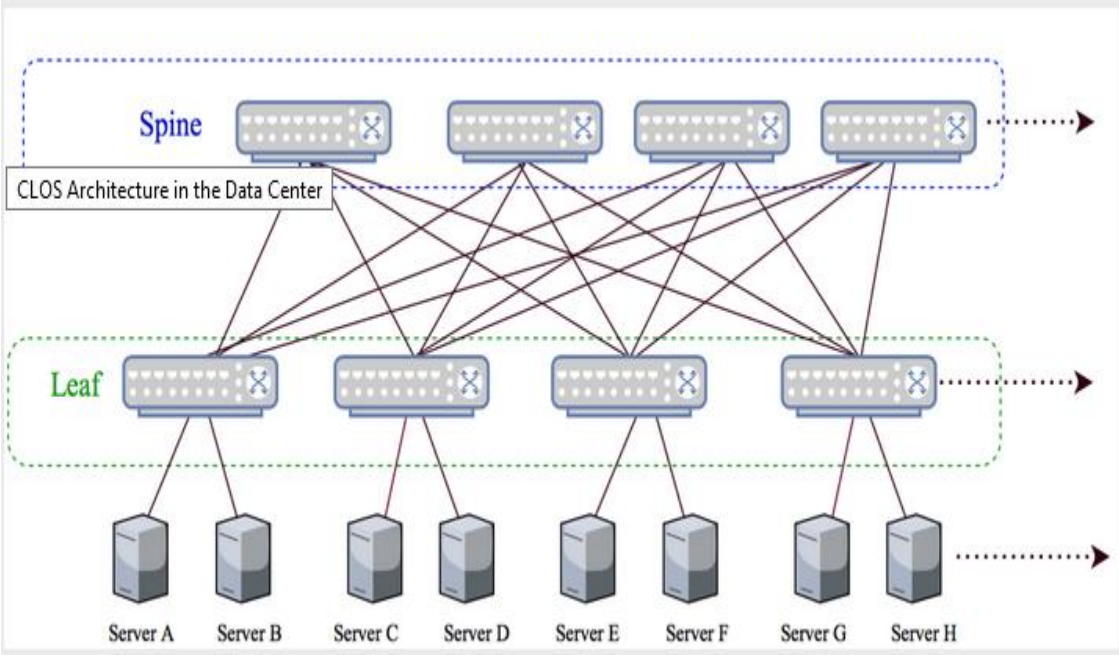
- Originally built as IEEE 802.1 Ethernet bridged networks
  - A set of Ethernet switches connected in a dynamically maintained tree topology
  - Controlled via spanning tree
- Roughly ten years ago we reached the architectural limits of such networks
  - DCs continued to grow in size
  - Addressing, convergence, and performance became issues
- Specifically
  - Scaling
    - Ethernet VLAN (IEEE 802.1q) uses a 12-bit VLAN tag, supporting 4096 tenants
    - Spanning tree means each Ethernet switch maintains forwarding state for all end-systems
  - Convergence
    - Spanning tree can take tens of seconds to converge
  - Load balancing
    - Spanning tree develops only one path leading to under- and over-used links
  - End-system mobility
    - New location has to be learned causing interruption in traffic delivery and churn in the bridged network

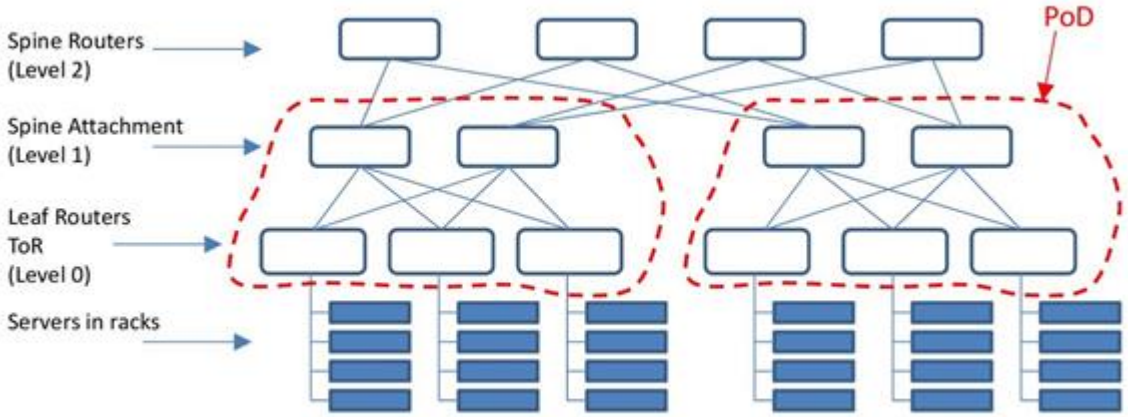
Data Centre Objectives Dictate Topologies

- Very large clusters of severs each hosting many virtual machines
- High volume of traffic in and out of the DC
- High volumes of intra-DC traffic
- Probably grow-out through bulk addition of servers
- Leads to some common approaches to DC topologies
  - Highly connected
    - No server more than a few hops from a DC gateway
    - No server more than a few hops from each other server
  - Redundant paths
    - No network failure should isolate any part of the DC
  - Modular architecture
    - Allows whole clusters to be added at once without needing significant rewiring

Common DC Topologies

- Clos networks
  - Invented in 1938 by Edson Erwin
  - Made useful in 1952 by Charles Clos
  - Valuable for wiring together telephone switches
- Clos Topology in the DC
  - 3-stage network
    - Each server is 3-hops away from the othe
    - Constant number of hops
  - 5-stage network (also known as 3-level)
    - Super spine
    - Connectivity nodes clustered into PoDs





The IP Fabric

- Set of IP forwarding nodes
  - Utilizing commodity silicon
  - Typically (but not always) lacking support for MPLS
  - Arranged in a Clos network
  - Lots of ECMP
- ToRs are given a prefix
  - In IPv4 this is typically a /24
  - Summarizes the ports (tunnel endpoints) assigned to it
  - Summarized when advertised outside the IP fabric
- IP fabric is largely static
  - Routing is to tunnel endpoints at ToR
    - These tunnel endpoints do not move within the IP fabric
    - Means IP fabric models a switch fabric where ports also do not move
  - Routing advertisements reduce to failure/repair
    - Link/nodes within the fabric
    - ToR failure means withdraw /24

Configuring the IP Fabric

- Automatic provisioning is extremely important
  - Networks are very large
  - Workload and scope for error are very high with manual provisioning
- Starts with the IP fabric topology and a naming space for the nodes within it
- Usually based on a “playbook”
  - A template for configuration of an individual node
  - Different playbook for each control plane protocol
  - Contains, for example:
    - Loopback address, IP addresses for each of the node's P2P links, an AS number, protocol configuration parameters
  - IP addresses are assigned starting with a base IP address and then moving through the IP fabric in a systematic way
- Nodes retrieve their individual copies of the playbook from the automatic provisioning system
  - DHCP (option 82), or other standard protocols
  - Routed via the nodes above it in the IP fabric
    - Implies and dependent sequence on start-up
  - Out-of-band management system LAN considered impractical
    - Network is too large
- Node initiate links
  - Use LLDP or CDP to verify its provisioned P2P connectivity
  - Bind the P2P IP addresses to each of its P2P links
  - Establish protocol sessions
- Provisioning systems need to handle nodes from multiple vendors

Overlay and Underlay Routing

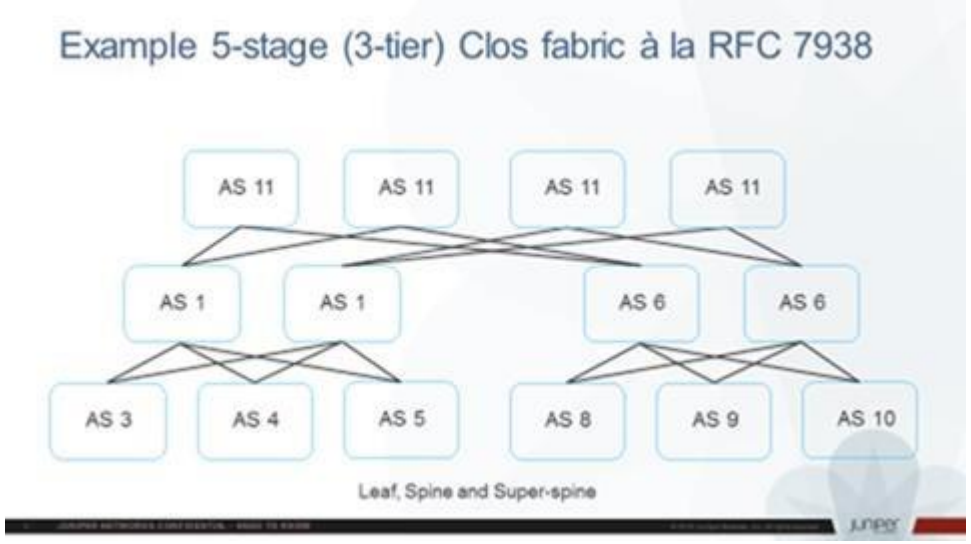
- Routing in the DC is considered as underlay and overlay
- Underlay routing is within the IP fabric
  - ToR to ToR
  - Gateway to ToR
  - ToR to Gateway
- Overlay routing is across the IP fabric
  - VM to VM
  - VM to external (other DC, Internet, etc.)
  - External to VM

Routing Problems

- L2 Technology
  - Scale issue with vlan-tag
  - Slow convergence of spanning tree protocol
  - Mac learning
  - End-point mobility
  - ECMP not supported
- IGP
  - Flooding in regular topology
- BGP
  - Configuration heavy
  - Slow convergence

BGP inside the Data Center

- Simple protocol state machines
- Reliable as it uses TCP underlay
- Minimum configuration
- Light weight policy for minimal traffic engineering
- Single hop E-BGP session
- Private ASNs from 64512-65534
- Unique ASN to Tire2 and Tier 3 devices



RIFT

- New Routing Protocol for Datacenters
- Distance vector down, link state up
- Automatic link discovery, automatic tree formation
- Minimal amount of route information on leaf/TORs
- High degree of ECMP natively (BGP needs additional knobs)
- Traffic Engineering on all links
- Fastest Convergence on failures
- Automatic disaggregation prevents blackholing

Flooding Reduction in IGPs

- Idea of flooding reduction existed in IGPs
- There are multiple proposals in IETF
- draft-li-dynamic-flooding-isis-00
  - Choose a flooding topology
  - Choose a leader in the topology
  - Use DR/DBR kind of election/Tie breaker
  - Leader advertises flooding topology
  - Flooding topology calculation is a local implementation
- draft-cc-isis-flooding-reduction
- draft-xu-isis-flooding-reduction-in-msdc-01

LSVR

- Addresses the slow convergence issue in BGP based DCs
- Uses BGP-LS Like NLRI to advertise Node/Link Prefix Info
- Computes SPF in BGP using topology info
- BGP best path computation replaced with SPF computation
- New SAFI named LSVR SAFI

EVPN as the Overlay Routing Protocol

- Mechanism to achieve end-point connectivity in DCs
- Same or different bridge domain
- Uses MP-BGP to distribute MAC addresses
- Uses concepts from traditional VPNs to separate customers/CEs
- Highly scalable for end-point connectivity
- In-built multi-homing and mobility

THANKS