Multi-site Datacenter Network Infrastructures

Petr Grygárek
Why Multisite Datacenters?

- Resiliency against large-scale site failures (geodiversity)
  - fire, wide area power outages, political reasons, law regulations ...
- Disaster recovery
- Easier handling of planned outages
  - Workload migration to unaffected site
- Traffic optimization
  - choose ingress point closer to requesting client
Interconnection of DC Sites
Traditional Requirements & Architecture

- L3
  - IP or MPLS
- Optionally L2
  - traditional design

Technically, L3 and L3 interconnection can be implemented on a single set of network devices (MPLS, EVPN, ...)

© 2009 Petr Grygarek, Advanced Computer Networks Technologies
Multi-site PoDs

Site A

WANs

Site B

L3 Interconnect
(MPLS/VPN)

Site C

ISPs

POD1

L2 Interconnect

POD2
Options of L3 extension between sites

- Dedicated core (IP-only or MPLS)
  - MPLS/VPN is beneficial for multiple tenants separation
- DMVPN over shared core (or Interner)
  - Multiple VRF instances, in tenants' VRFs
- EVPN (L3 encapsulation)
  - Tenants are separated using L3 VNID
- ...

© 2009 Petr Grygarek, Advanced Computer Networks Technologies
Why to extend L2 between sites?

- Server admins like transparent VM mobility
- Distributed clusters for better resilience
  - FWs, LBs, NASs, ...
- Server clusters
  - e.g. using Windows NLB

Be aware that usage of technologies originally developed as "local" in multi-site environments always need careful consideration

- Timers built in application software or hardware appliances (e.g. storage clusters)
L2 Extensions between DC Sites

• Dual-site (special, simplified case)
  • P2P virtual links: QinQ, EoMPLS /AToM, …
  • Multichassis Etherchannel/Virtual chassis  P2P L2 technologies
    – Cisco VPC/VSS, Dell VLT, Juniper VC, …

• General topology:
  • Redundant switched network with STP (non-recommended !), also includes QinQ
  • Distributed virtual chassis
    – if latency between sites fits into solution's limits
  • TRILL/FabricPath
  • VPLS
  • Cisco OTV
  • VxLANs (advantage: can solve both L2 and L3 extension))
  • Various SDN interconnect „clouds“ (mostly IP overlay)
Transparent Interconnection of Lots of Links (TRILL)

- IEEE 802.1aq
- L2 multipath solution
  - eliminates Spanning Tree, no stability issues
  - no blocked ports
  - reduced latency – shortest path always used
  - alternative active paths (equal-cost)
    - path selection based on data packet header hash ensures ordered delivery
- L2 frame encapsulation ("L2 over L3")
  - new header carries egress switch identity
- ISIS-like internal routing of encapsulated frames
TRILL Principles

- **Rbridge – TRILL-capable bridge**
  - Ingress, egress, TRILL cloud internal

- **Switches have identities, ISIS calculates shortest paths between switches**
  - ISIS chosen as it runs directly on L2 (no TCP/IP) and is generic enough (new TLVs)

- **2-level switching hierarchy**
  - Only Rbridge addresses have to be known in TRILL core
  - Smaller MAC address tables, better scalability

- **Data-plane MAC address learning:**
  - Conversational learning used to save space in MAC tables
    - Only src MACs of unicast frames to destinations known to be local are learned
  - Backward learning still used to learn addresses from outside of TRILL cloud
  - Ingress Rbridge maintain <MAC, egress Rbridge> or <MAC, local port> records (VLANs also supported)

- **Optional Control-plane MAC address learning**
  - End-Station Address Distribution Information (ESADI)
  - Proactive MAC address propagation (configurable per-VLAN), cryptographically secured, fast location updates
TRILL vs. Cisco FabricPath

- TRILL Rbridges may be interconnected via legacy Ethernet clouds
  - Not meaningful in DC environment, Cisco FabricPath (TRILL alternative) does not support this

- Next-hop header allows passing of TRILL frame over each legacy internal Ethernet segment (even VLAN-based) – if any
  - DST MAC in outer header specifies next-hop Rbridge
  - For each legacy Ethernet interconnect segment single Rbridge is elected (per VLAN) to avoid looping/frame duplication

- Inner header allows routing of TRILL frame to egress Rbridge

- Loop protection: Hop count (TTL) in TRILL header
TRILL Multidestination Frame Forwarding

• For broadcasts, unknown unicasts and multicasts (BUM)

• One or more distribution tree covering all egress Rbridges is calculated

• Distribution tree to be used to distribute particular data frame (destination root switch) is chosen based on “destination Rbridge address“ field of encapsulating frame
VxLANs
**VxLANs - Usage and Principles**

- **eXtensible VLANs**
- **Layer 2 overlay over a Layer 3 network**
  - Stateless tunnelling between VTEPs
  - Various payload-to-VTEP address mapping mechanisms
  - UDP encapsulation
    - well-known destination port 4789, „random“ src port to support ECMP hashing
    - Src port is actually a hash of dataframe header
    - MTU increase in transport network is needed
    - fragmentation not desirable
- Originally intended for inter-hypervisor traffic via L3 infra
- Multitenant overlay switching/routing fabrics based on IP ECMP underlay and VxLANs can be commonly seen today
  - Traditional VLANs limited to rack with particular leaf
VTEPs, bridging & routing

- Virtual Tunnel End-point (VTEP)
- Each VXLAN segment (VNID) is mapped to IP multicast group in the transport IP network to carry BUM traffic
  - alternatively, unicast replication can be used
- VLAN-VxLAN bridging
  - hardware or software gateway
  - combines together traditional L2 VLAN and VxLAN segment to single broadcast domain
- Inter-VxLAN routing
  - also routing between VxLAN VTEP and traditional L3 interfaces
- Anycast gateway – same GW IP address „present“ on all VxLAN fabric leafs
• **VxLAN header**
  - 24b segment ID
  - Reserved/Flags – for future extensibility
  - utilized by various „SDN“ technologies (e.g. Cisco APIC)
VxLANs implementation scenarios

- **MAC address learning options**
  - From data plane (+ multicast BUM traffic )
  - Separate control plane
    - EVPN
      - Proactive MAC propagation using BGP (dedicated AF) – limits or eliminates flooding
      - automatic VTEP discovery via BGP updates
      - also handles redundant attachment of the same switched segment
    - SDN
      - SDN controller / virtual machine manager has full knowledge what VMs reside on individual hypervisors/VTEPs
    - any other defined in future
MP-BGP EVPN

• Standard based
  • MP-BGP control plane, (VxLAN or MPLS data encapsulation)

• Combines intra-VxLAN bridging with inter-VXLAN routing (integrated bridging and routing - IRB)
  • Optimal for both east-west and north-east traffic (host routing)

• BGP propagates host routes (MACs and IP /32s) and site IP prefixes (/nn)
  • VxLAN dataplane flooding is still used for BUM traffic
    – /nn prefixes may help to route to IP hosts

• Knowledge of MAC/IP pairs can also limit ARP flooding (ARP suppression)

• Supports multiple attachments of the same L2 segment
EVPN Terminology

- EVPN Instance (EVI) = virtual switch (with VLANs)
Address Learning

• Local Learning
  • MAC addresses - from source MAC of data frames (mandatory)
  • IP addresses: learning from DHCP messages (optional)

• Remote Learning
  • BGP MAC+IP advertisements (type 2 routes)
EVPN (Host) Propagation

- Proactive address propagation via MP-BGP
  - L2vpn-evpn address family
    - L3 address + length, L2 address + length
    - L2 VNID (VxLAN ID), L3 VNID (VRF membership)
  - Site IP prefixes are used to limit unknown unicast flooding for silent destination hosts
  - BGP authentication mechanisms can prevent rogue VTEPs
    - Additional data plane anti-spoofing mechanism: VxLAN tunnel source (outer) header must correspond respective source's BGP next hop
EVPN Multi-Tenancy Support

- VxLAN (L3) interfaces in separate VRFs
- Route Distinguisher makes host routes with overlapping address spaces (IP, MAC) unique in BGP
- Route Targets regulate who can talk with whom
  - Route visibility, even between different VRFs
  - Same concept as in L3 MPLS/VPN
Host/site multihoming using LAG

- Ethernet segment (ES) = set of links to the same customer site/host
  - downlink LAG from respective EVIs' of 2 different leaf switches
  - Identified by End System Identifier (ESI)
    - 0=single homed
EVPN BGP routes

• **Type 1 – Ethernet AutoDiscovery**
  - Propagates ESI of each leaf SW
  - Allows leaf Sws connected to the same site (L2 network) to find each other

• **Type 2 – MAC/IP Advertisements**
  - also handles multihomed hosts/sites
  - Propagates ESI

• **Type 3 – Inclusive Multicast Ethernet Tag Route**
  - Signalling of multicast tunnel
  - Multicast group IP address

• **Type 4 – Ethernet Segment Route**
  - Used to elect designated forwarder for multicast on each ES

• **Type 5 – IP prefix route**
All-Active Multipath

• By combining Type 2 and Type 1 routes, source EVI can load-balance between multiple egress leafs with downlinks to the same ES
  • Even if all traffic from destination to source was always hashed to a single leaf SW which learnt is MAC address but other leaves had never seen it

• If a downlink from leaf switch to ES goes down, it sends BGP withdrawal of type 1 route (whole ESI) instead of withdrawing all individual host routes (type 2) => faster reconvergence after failure
Broadcast & Unknown Unicast Traffic Flooding

- Unknown hosts shouldn't exist as BGP advertised all (non-silent) host proactively
- ARP broadcasts should not be needed because of ARP proxying (info from BGP host routes)
- Flooding of unknown and broadcast traffic can be enabled/disabled administratively
  - Ingress replication / underlay (multicast) replication
Multicast routing

• Split horizon replication
  • From local source to other local segments and to remote VTEPs
  • From remote source to local segments only

• Designated forwarder on each ES has to be established to avoid multiple copies of the same frame for multihomed sites

• Prevention against hearing multihomed source's own multicast traffic back over EVPN core
  • if source VTEP IP address in received VxLAN packet advertises the same ES connectivity as local destination ES, packet is dropped
MAC Moves

• **Leaf1 does not know that MACx has moved until it hears Type 2 route from Leaf2**
  - Initiates triggered Type 2 route withdrawal
  - Sequence numbers are used to avoid race condition in case of multiple moves
EVPN: Inter-VxLAN routing options

- **Distributed anycast gateway**
  - L3 VxLAN interface on multiple switches
  - same IP and MAC on multiple switches

- **Asymmetric mode: Inter-VxLAN routing on source VTEP**
  - All VxLAN's L3 interfaces must be present on all switches
  - after packet is routed, it is then carried in destination VxLAN to egress VTEP (egress VTEP does only L2 lookup)
  - Unscalable (source VTEP must maintain host routes of all VxLANs)

- **Symmetric mode (preferred): Inter-VxLAN routing using (dynamic) virtual L3 P2P segment between source and destination VTEP switch**
  - Source and destination MAC on virtual L3 interconnect segment are rewritten according to well-known stateless procedure (corresponds to standard behaviour on L3 interconnect segment)
  - Packet arrives to destination switch from virtual L3 segment and then is routed to destination VxLAN
  - Both ingress and egress VTEP must do L3 lookup (+ L2 lookup)
EVPN External Traffic

- BGP may convert between L2VPN (fabric internal) and IPv4 unicast routes (external connections)
  - respectively to per-VRF external routes – VPNv4 AF
See other presentations for another inter-DC L2 technologies
Multisite DC without L2 extension

- VM migration across different subnets (keeping original IP address) is often required
  - Keeping VM identity (including FW openings) and established sessions
- Potential alternative solutions
  - Load-balancer frontend (L4 session termination or NAT)
  - Mobile IP (not used actually today)
  - LISP
LISP
Locator Identity Separation Protocol
LISP Motivation

• Identity and Location are mixed together in today's IP routing scheme

• Advent of server virtualization made servers (VMs) mobile
  • Mobility is advantageous for various operational reasons
  • VMs are decoupled from physical infrastructure (IP address does not imply physical location anymore)
  • VM has still to keep its identity, even if moved between subnets
    • VLAN extension over the whole world is not a solution

• Same principle may apply to Internet multihoming with provider-independent addressing
LISP Principles (1)

• Location / IP address separation
  • Endpoint Identifier (EID)
  • Routing Locator (RLOC)

• Separate address spaces for IDs and Locators
  • arbitrary values (e.g. MAC + GPS) or IP addresses (endpoints) + IP addresses (tunnel)
  • RIPE currently provides experimental registry for EID prefixes (/48)
  • additional level of indirection

• Mapping of EIDs to RLOCs is needed

• Traffic is tunneled to current position of EID (identified by one or multiple alternative RLOCs)
  • Ingress Tunnel Router (ITR): source site → LISP
  • Egress Tunnel Router (ETR): LISP → destination site
LISP Principles (2)

• Various mapping mechanisms were proposed
  • Mapping server(s)/resolver(s)
    • LISP-ALT topology
      • BGP over GRE between LISP routers (normal IPv4/v6 AFs for EIDs)
      • Map request and responded routed over it
    • DDNS-like structure
  • Standardized Query/Reply messages
  • Resolver proxies info from mapping mechanism for clients
    • Client can find suitable resolver e.g. using anycasting
  • EIDs may be advertised as whole subnets or per host
  • Multiple RLOCs may be advertised for single EID
    • Priority differentiates between alternative ETRs
    • Weight defines load share between ETRs of same priority
LISP Principles (3)

• Preconfigured EID (subnet) to RLOC mapping is useful e.g. for subnet multihoming
  • cases where EID does not move

• New EID mapping may be published dynamically if ETR detects arrival of VM to its subnet
  • ETR Informs other ITR/ETRs (multicast group) and registers new RLOC(s) with mapping server
LISP VM moves between L3 segments (distributed DC case)

• Outgoing traffic uses original VM's default GW
  • VM does not know that it has been migrated

• Proxy ARP is used for cold migration scenerio
  • VM's ARP cache is empty

• Manual synchronization of gateway HSRP/VRRP MAC addresses between DC sites (different IP segments) can solve hot migration case
  • in addition to proxy ARP