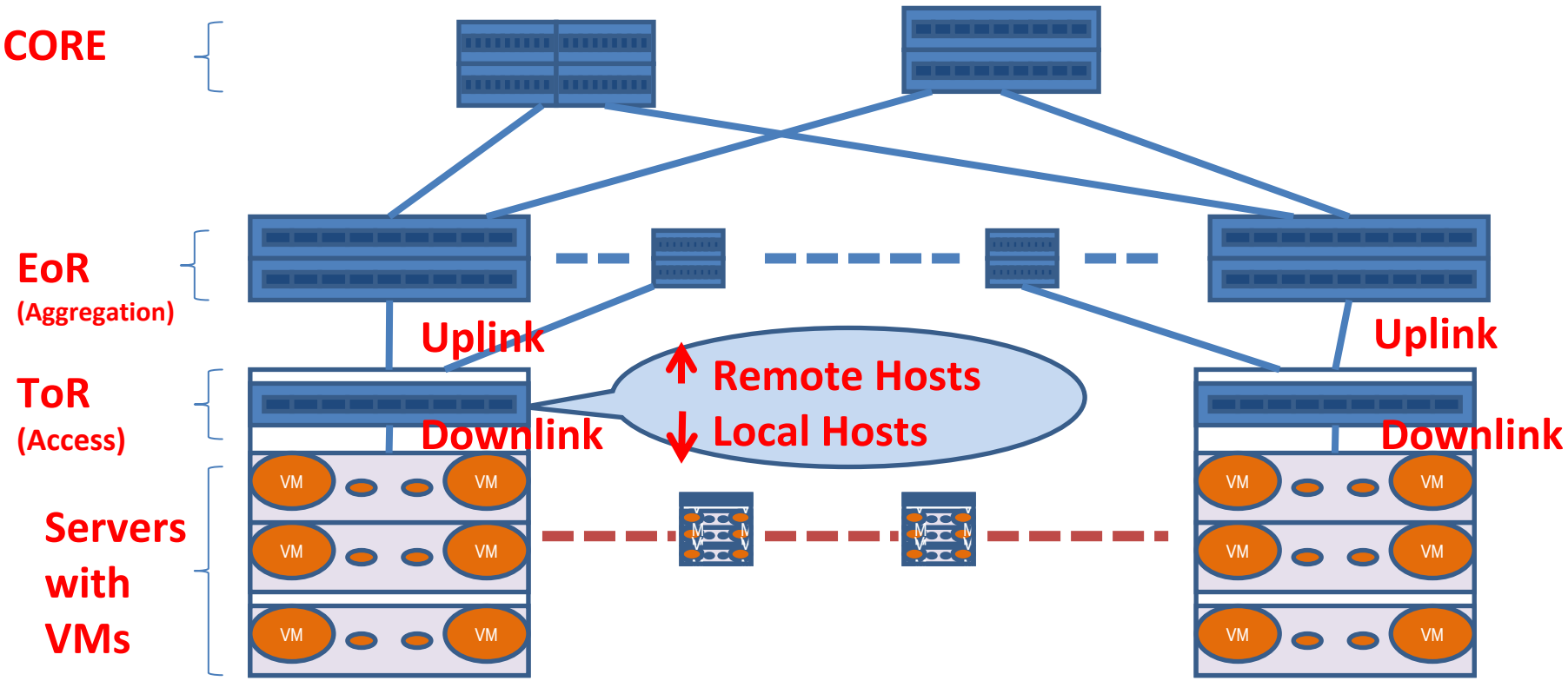
ARP Broadcast Reduction for Large Data Centers

draft-shah-armd-arp-reduction-01

Himanshu Shah
Anoop Ghanwani
Nabil Bitar

IETF-79

Typical Data Center Topology



Solution Overview

- First Hop maintains ARP tables created using learned information from transiting ARP PDUs
  - Usually this is the ToR but it may be an EoR
- Local Hosts
  - MAC<->IP entries of hosts learned from downlink
- Remote Hosts
  - MAC<-> IP entries of hosts learned from uplink
- ARP Packet Processing
  - Requests (broadcast): Sent to control plane
  - Response (unicast): Bi-casted
  - Gratuitous (broadcast): Sent to control plane
- Intend to cover Neighbor Discovery based solution for IPv6 in later revision

ARP Packet Processing

- ARP Request
  - Look up Source IP Address in ARP table
    - Learn/update/refresh
  - Look up Target IP Address in ARP table
    - Present: Send ARP Reply using information in the table
    - Absent: Forward the ARP request
- ARP Reply
  - Look up Target IP Address in ARP table
    - Learn/update/refresh
- Gratuitous ARP
  - Look up Source IP address in ARP table
    - Present: Refresh and discard
    - Absent: Learn and forward as broadcast
- Learn/update refers to MAC<->IP binding
- Refresh means restarting the aging timer for the entry

Other Considerations

- Implementations may favor Local Hosts over Remote Hosts if ARP table space in the ToR is limited
- Coordination of VM mobility between server and network is desirable
  - Use IEEE802.1Qbg or admin control
- Aging timers
  - Loss of ARP packets due to network congestion could be problematic
  - ARP aging time may need tuning
    - If too short then ARP reduction is less effective
    - If too long then potential “out-of-sync” situation could last longer
  - Local and remote aging timers could be different

Summary

- Simple, non-intrusive solution to control ARP broadcasts in large data centers
  - Uses traditional and well-known methods
- Offloads hosts' part of the ARP processing to external switches
- ARP tables can be maintained in control plane
  - No impact to existing hardware