2010-11-7

Address Resolution for Massive amount of hosts in large Data Center (ARMD)

Problem Statements

Linda Dunbar (ldunbar@huawei.com) & Sue Hares (shares@huawei.com)
Murari Sridharan (muraris@microsoft.com)
Narasimhan Venkataramaiah (narave@microsoft.com)
Ning So (ning.so@verizonbusiness.com)

Examples of Large Internet Data Center

- **None shared data center:**
  - Service provider data center which houses applications for different customers
  - Enterprise Data Centers
- **Shared Data Center:**
  - Data center not dedicated to one enterprise or service provider
    - One service provider may own part of infrastructure (servers & switches) and may need to lease VMs from the data center owner as backup plan
- **Data center whose resources, which include computing, storage, and network, are shared by many types of customer access methods, including private line, VPN, and internet.**

Special Properties of large Internet Data Center

- **Massive amount of hosts**
- **Massive amount of client subnets or Closed User Groups co-existing in cloud data center, with each subnet having their own IP addresses**
- **Hosts (VMs) migrate from one location to another**
  - Physical resource and logical hosts/contents are separated
    - applications can be loaded to any Virtual Machines on any servers,
    - VMs can be migrated to different locations for efficient server and storage management.
  - VM Migration can be in-service or out of service.
    - In both cases, VMs have to maintain the same IP and MAC addresses and same subnet
  - Sometimes VMs migrate from one VLAN to another.

ARP problems in general

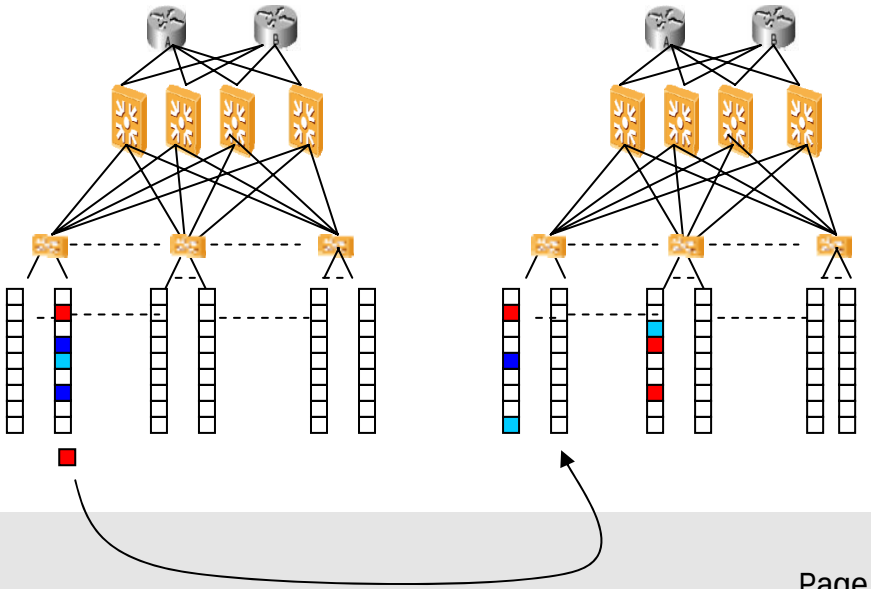
- **There are lots of ARP messages:**
  - Hosts frequently send out gratuitous ARP.
  - Hosts (applications) age out MAC to target IP mapping very frequently.
    - Usually in minutes.
    - Servers/hosts and their applications behavior are unpredictable
- **The impact of huge amount of ARP messages in one broadcast domain:**
  - Heavy impact to servers
    - Typical low cost Layer 2 switches don't have sophisticated features to block broadcast data frames or have policy implemented to limit the flooding and broadcast storm.
  - Force switches (e.g. TOR) to learn many useless source MAC addresses (newly learnt since bar BOF)
    - For a subnet with 1000 hosts, if there is only one host of the subnet residing under TOR-1, the TOR-1 has to learn all the 1000 MACs for all the hosts because of frequent ARP msgs even though the host under the TOR-1 may only need to talk to a couple of other hosts in the subnet.
    - When hosts' ARP timer is shorter than switches MAC FDB time-out value, the switches will be refreshed of all the MACs
    - When the TOR-1 has thousands of servers underneath, the MAC FDB can overflow causing more unknown flooding.

ARP Problems get worse when VMs migrate

- **Some hosts might be temporarily out of service during transition.**
  - **Lots of ARP request broadcast messages transmitted from hosts to temporarily out of service hosts.**
    - switch does not learn their path because there is no response from those target hosts,
    - causing all ARP msgs from various hosts will be broadcasted repetitively.
- **VMs are spawned automatically by VM-manager**
- **Gratuitous ARP broadcast from new location flood to all TOR switches**
  - **Why:** new TOR doesn't know where target TORs for hosts belonging to the same broadcast domain are located:
- **Most hosts don't send anything when leave one location, and some hosts don't send gratuitous ARP when emerging from the new location**

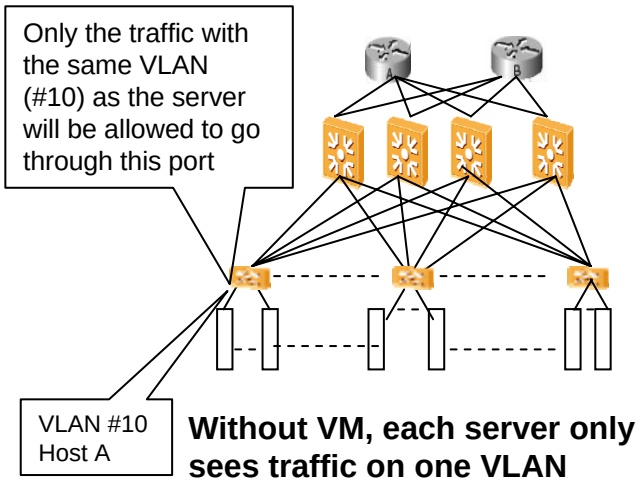
Example

- **When a host A belonging to a particular Subnet-X (or VPN-X) migrates from TOR-2 to TOR-10,**
  - For all other hosts who need to communicate with A, their corresponding TORs don't know where A is anymore.
    - Some data frames are sent to original TOR where A was
      - old TOR must either re-direct the data frame to the new location, or
      - flood the data frame to all other TORs.
    - Re-direct volume can be much more than static data center. To achieve re-direct, TOR need to have large size memory to keep track of all the hosts ever stayed in the rack.

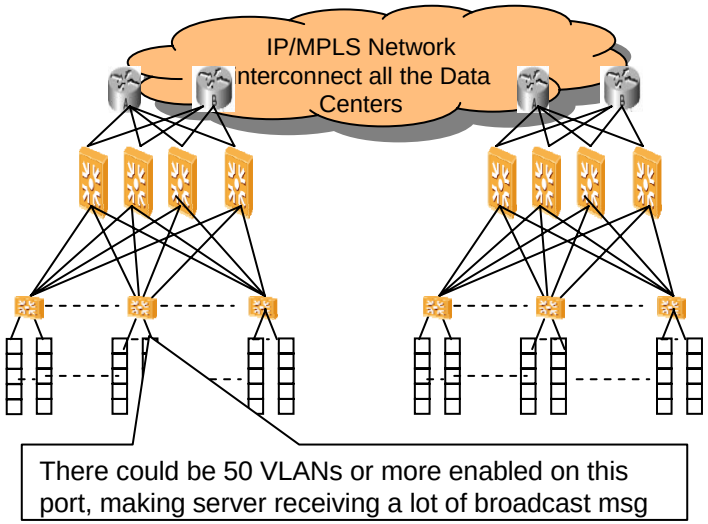


Why VLAN (or smaller subnet) alone is not enough

- **Hosts in one subnet (VLAN) can be anywhere and their locations is not fixed.**
  - TOR may not have all the information of target TOR. Under this circumstance, it has to broadcast ARP to all other TOR in order to reach the target.
- **When one physical server is supporting >100 Virtual Machines, i.e. >100 hosts, most likely the virtual hosts on one server are on different subnets (VLANs).**
  - If there are 50 subnets (VLANs) enabled on the switch port to the server, the server has to handle all the ARP broadcast messages on all 50 subnets (VLANs). The amount of ARP to be processed by each server is still too much.



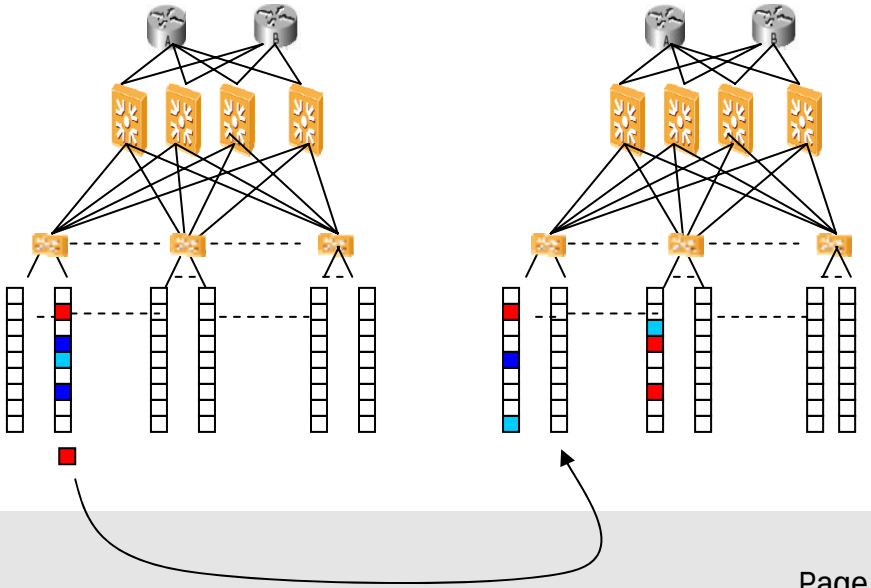
Without VM, each server only sees traffic on one VLAN



Why re-direct doesn't work well in the large IDC environment

(learned on the way to BOF)

- **VM migrates from one rack to another, or from one data center to another all the time**
- **For re-direct to work, the TOR switch need to keep track of where the host is moving to.**
- **Instead of small amount of moves in traditional network, majority of (virtual) hosts under each TOR move in and out all the time and in large quantity.**
  - In order for TOR to re-direct traffic, TOR has to have a large amount of memory to keep track of all hosts ever existed in this rack
  - The re-direct traffic volume may be very high.

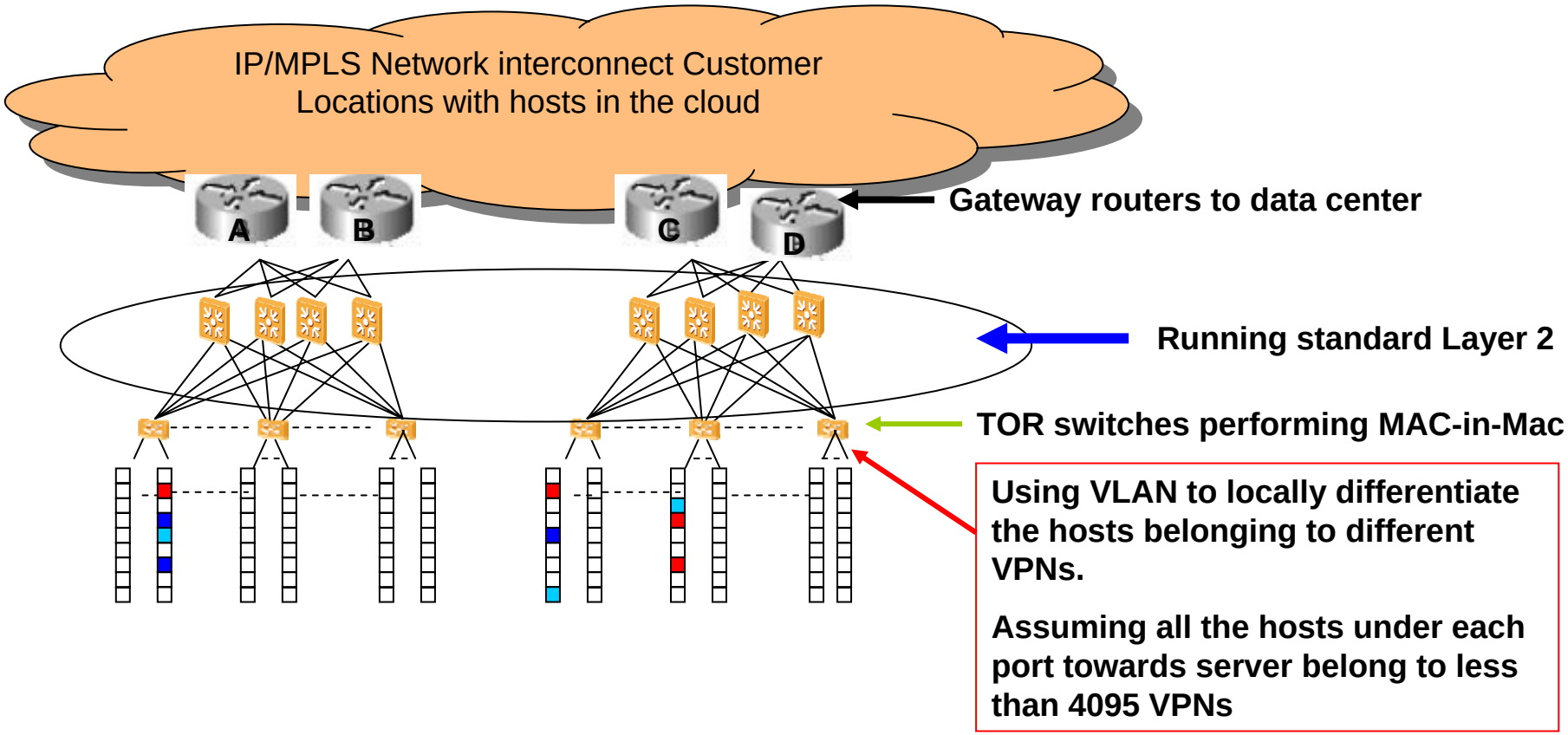


Problems with larger than 4095 subnets

- **When there are more than 4095 Closed User Groups (or VPNs) residing in one data center**
  - There is not enough VLANs to separate all the hosts belonging to different VPNs
  - Some types of encapsulation, like IP encapsulation (L3VPN, L2VPN), or MAC-in-MAC encapsulation, have to be used to further segregate traffic belonging to different customers (Closed User Groups)
    - Encapsulation can be done by TOR, or virtual switch within server

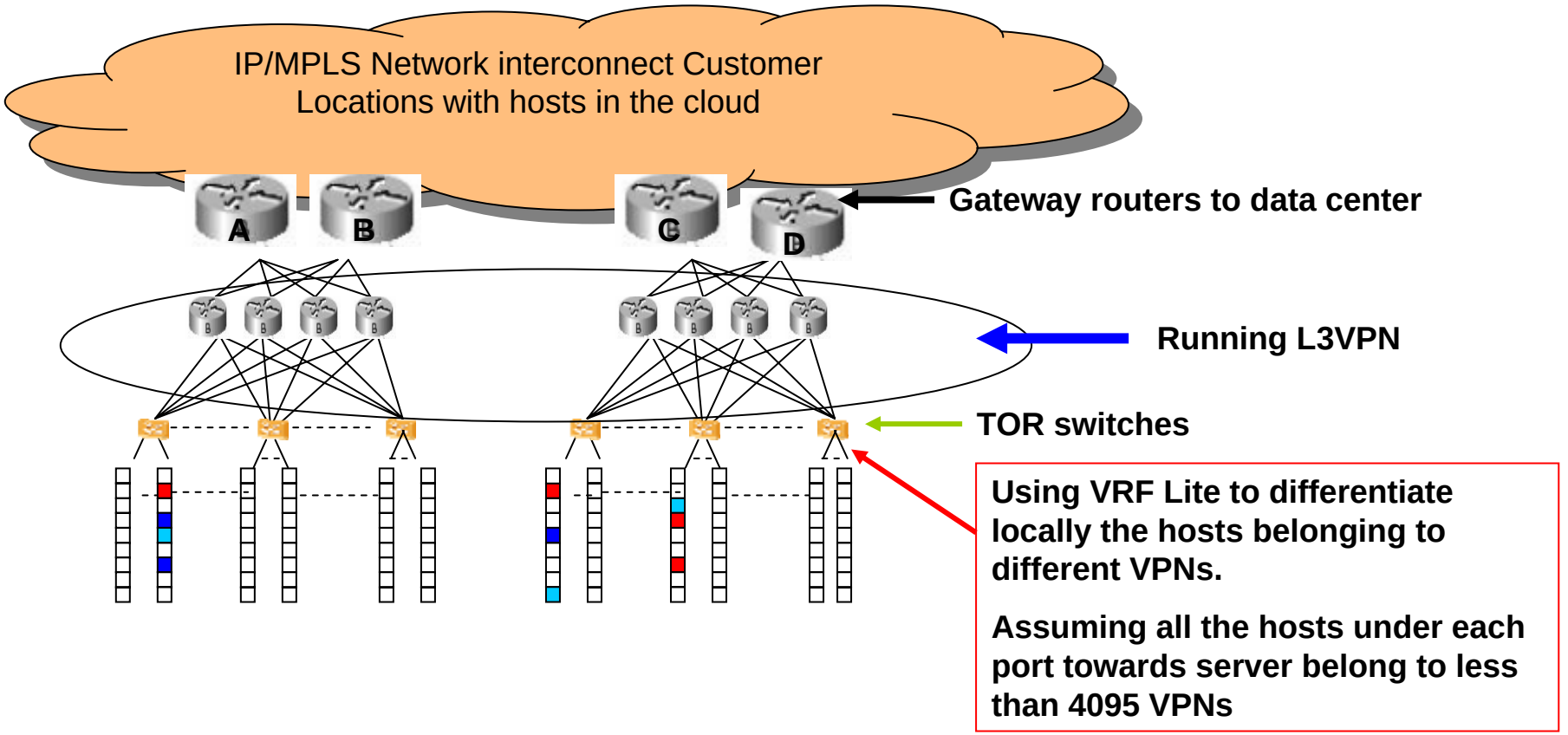
Using MAC-in-MAC from TOR switches to segregate hosts belonging to different VPNs

- **Using Service Instance ID to differentiate different VPNs**



Extending L3VPN into Data Center TOR switches

- **Extending L3VPN to TOR switches.**



Possible ways to reduce ARP storms

- **TOR ARP caching and proxy based approach**
  - This approach can alleviate some ARP storms bombarding application servers.
  - When VMs migrate, this approach has its limitation.
- **Directory based approach**
  - In the form of Address Directory or Address Location Directory
- **Others**

Proposal to IETF

- **Create a new IETF working group (http://trac.tools.ietf.org/bof/trac/attachment/wiki/WikiStart/ARMD%20charter%20and%20milestones%20v4.docx)**