2011-3-29

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

Problem Statements-01

Linda Dunbar (ldunbar@huawei.com) & Sue Hares (shares@huawei.com)
Murari Sridharan (muraris@microsoft.com)
Narasimhan Venkataramaiah (narave@microsoft.com)
Benson Schliesser(bschlies@cisco.com)

Special Properties of large Internet Data Center

- **Massive number of hosts**
  - Server virtualization makes it easier to instantiate multiple Virtual Machines (VM) on one physical server.
- **Massive number of client subnets (VLANs, or Closed User Groups)**
  - Each client may need 10 plus VLANs for various zones. Some subnets may have their own IP addresses
- **High desire to dynamically grow and shrink resources to meet the demand and to draw those resources from most optimal locations.**
  - This may lead to hosts belonging to one subnet to span across multiple shelves or locations.
  - Then, the ARP broadcast and ND multicast messages will traverse many backbone links and switches.

Amortized Cost	Component	Sub-Components
~45%	Servers	CPU, memory, storage systems
~25%	Infrastructure	Power distribution and cooling
~15%	Power draw	Electrical utility costs
~15%	Network	Links, transit, equipment

Why Layer 2 in Data Center?

- **VM migration requires Source and Destination to be in Layer 2**
  - **VMs being moved need to maintain the same IP addresses**
- **Many network services such as firewalls and load balancers must be in-line with network traffic in order to function correctly.**
  - **Some load balancing algorithms, e.g. Direct Server Return, requires all hosts in Layer 2**
  - **Layer 2 networks often provide a form of traffic engineering for steering traffic through these devices for a given subnet or segment.**
- **Active/Standby hosts need to be in one Layer 2**

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
    - 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/ND Problems get worse when subnets are not confined to one location

- **ARP/ND broadcast/multicast messages are no longer confined to smaller number of ports.**
- **Some hosts might be temporarily out of service during VM migration.**
  - **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.
- **Gratuitous ARP broadcast from new location flood to all TOR switches**

Why VLAN (or smaller subnet) alone is not enough

- **VLAN works well when all hosts belonging to one VLAN are confined to one location .**
- **When hosts belonging to one VLAN are placed at different shelves and one shelf has multiple VLAN enabled, all broadcast messages are no longer confined anymore.**
  - **The effect is same as one large VLAN.**