Network Issues in Cloud Data Center and Potential Enhancement

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

Linda Dunbar (ldunbar@huawei.com)
Sue Hares (shares@huawei.com)

Network Innovation is the Key to Enable IDC Scalability and Agility

- **Even though Network is only about 15% of total IDC cost, today's network constraints (e.g. VLANs, ACLs, load balancer, broadcast domains) creates barriers to IDC agility and increases fragmentation of resources**
  - lead to low server utilization and more power consumption
- **Agility is one of the most important requirements for Next-Gen IDC**
  - The trend of DC is to scale out (i.e. more quantity of servers) instead of scale up (i.e. higher quality)
  - The ability to dynamically grow and shrink resources to meet the demand, and
  - to draw those resources from most optimal locations.

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

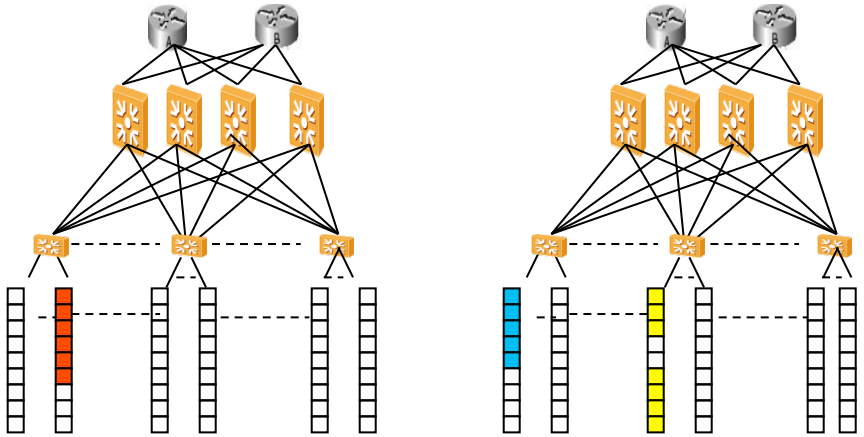
Cost Breakdown of a Typical Data Center

ARP problems in general

- **Hosts (applications) age out MAC to target IP mapping very frequently. Usually in minutes.**
  - For Microsoft Windows (Window XP or 2003), the default ARP cache policy is to discard entries that have not been used in at least two minutes, and for cache entries that are in use, to retransmit an ARP request 10 minutes. For later Window systems, the timeout value is in seconds (for ND)
- **Servers/hosts and their applications behavior are unpredictable**
  - They are from variety of vendors. Some of them frequently send ARP and other broadcast messages.
  - 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.
- **Hosts frequently send out gratuitous ARP**
  - when it does a switch over (active to standby)
  - When they have software glitch
- **That is why most Layer 2 networks have traditionally been limited to less than 200 hosts.**

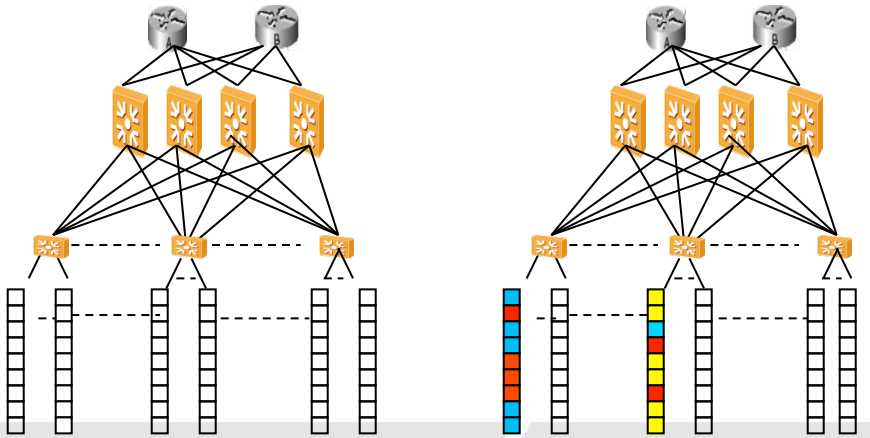
When energy efficient placements are required in data center

- Traditional network design creates silos of servers, but broadcast is confined within very limited number of switch ports



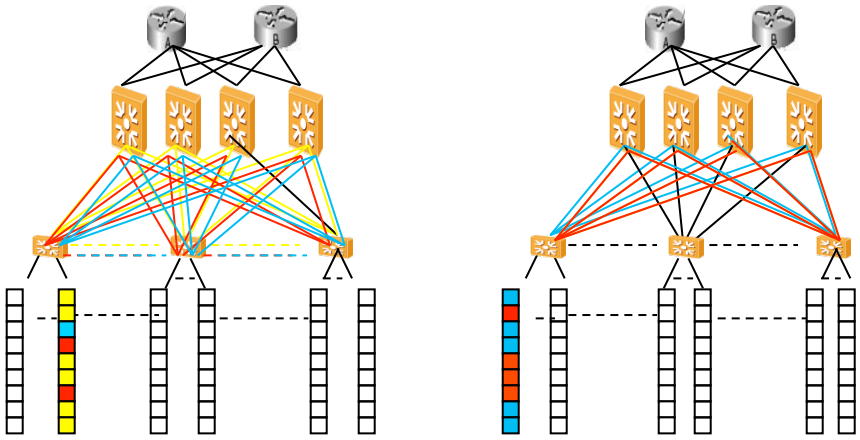
With strict partition of server resources, each server only sees traffic on one VLAN

- Energy and resource can be better utilized if all resources (servers, switch, etc) are combined to a bigger pool.
  - Resource aware algorithms can use a placement that satisfies the processing requirements of each VM but require the minimal number of physical servers and switching devices.



Small subnet is not enough:

- **When energy efficient algorithms are placed, subnets tend to extend throughout the network**
- **Then ARP/ND traffic associated with each subnet is likely to traverse a significant number of links and switches in the network.**
  - Thus, in the case of the virtualized data center, the partition of the physical network into multiple subnetworks may not confine ARP/ND to a limited number of links/switches as might otherwise be expected.



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.**
  - **Since there is no response from those target hosts**,
    - switch does not learn their path,
    - 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:
  - **Result:** Without any ARP optimization, all TOR will flood the broadcast to all servers underneath.
- **Most hosts don't send anything when leave one location, and some hosts don't send gratuitous ARP when emerging from the new location**

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

Why Directory Based Approach should be considered for large Data Center

- **Directory based approach, most likely distributed directories, can eliminate a lot of ARP broadcast messages by**
  - responding to ARP requests and terminate all Gratuitous ARP,
  - sending the target host location information to the Source TOR where requesting host reside (If ARP can be extended):
    - Result: proper Destination can be used for encapsulation from Source TOR to Destination TOR.
- **It is easier to build Address Directories for a data center than for traditional networks because**
  - there are resource management system(s) managing resources allocation, like loading guest hosts (VM) to a specific server, etc.
  - Information from those management systems can be easily transferred to Address Directories.

Proposal to IETF: Create a new IETF working group (ARMD)

- **To document the scaling characteristics of ARP (IPv4) and ND (IPv6) with respect to increasing numbers of hosts in data centers,**
- **To identify limitations of ARP and ND in such environments,**
- **To provide design recommendations intended to minimize issues associated with the identified limitations, and**
- **Potentially to develop appropriate solutions to better support data centers with large number of hosts**
  - To confine the IPv4's ARP broadcast messages to smaller zones
  - To facilitate any type of encapsulation at TOR by allowing Address Directories to send Destination TOR addresses
  - To improve re-direct process (both IPv4 and IPv6) when hosts migrate.
  - To enhance security for detecting/preventing gratuitous ARPs from malicious hosts.

We are looking for feedback!!!

- **Is it problem you've seen or anticipate at your data center?**