BCP for ARP/ND Scaling for Large Data Centers

http://datatracker.ietf.org/doc/draft-dunbar-armd-arp-nd-scaling-bcp/

Linda Dunbar: ldunbar@huawei.com
Warren Jumari: warren@kumari.net
Igor Gashinsky: igor@yahoo-inc.com

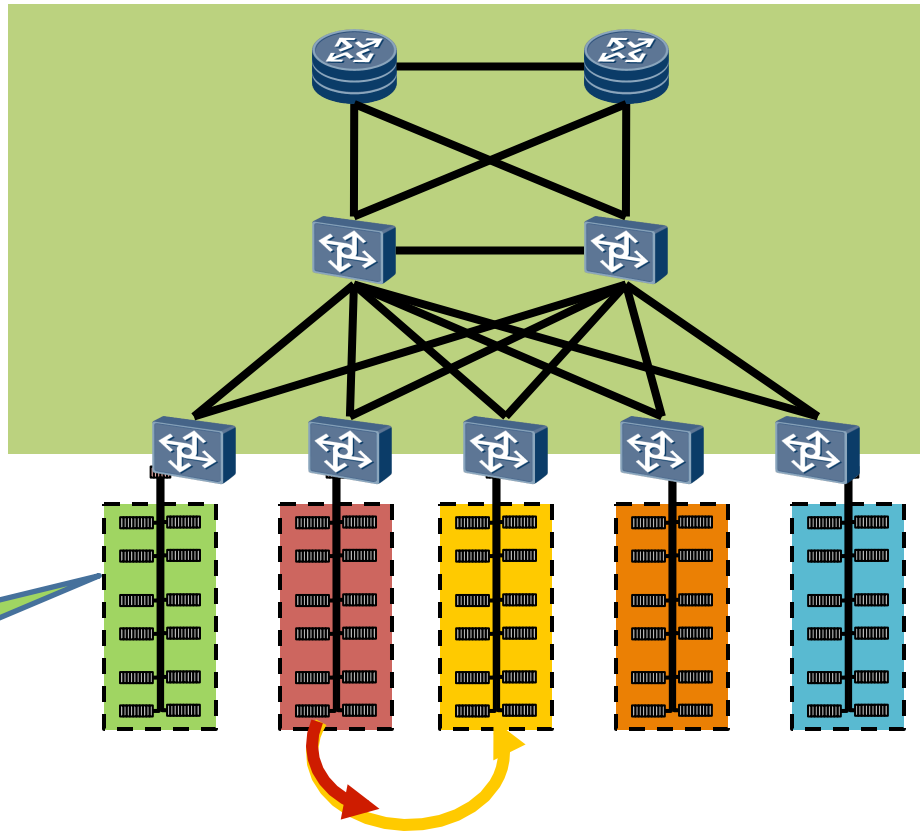
BCP #1: L3 to Access (ToR)

- A single rack is it's own L2 domain, has its own IP subnet:
  - Benefits: ARP/ND scale very well.

Practice Recommendation:

Consider overlay at ToR or at Hypervisor to hide host addresses

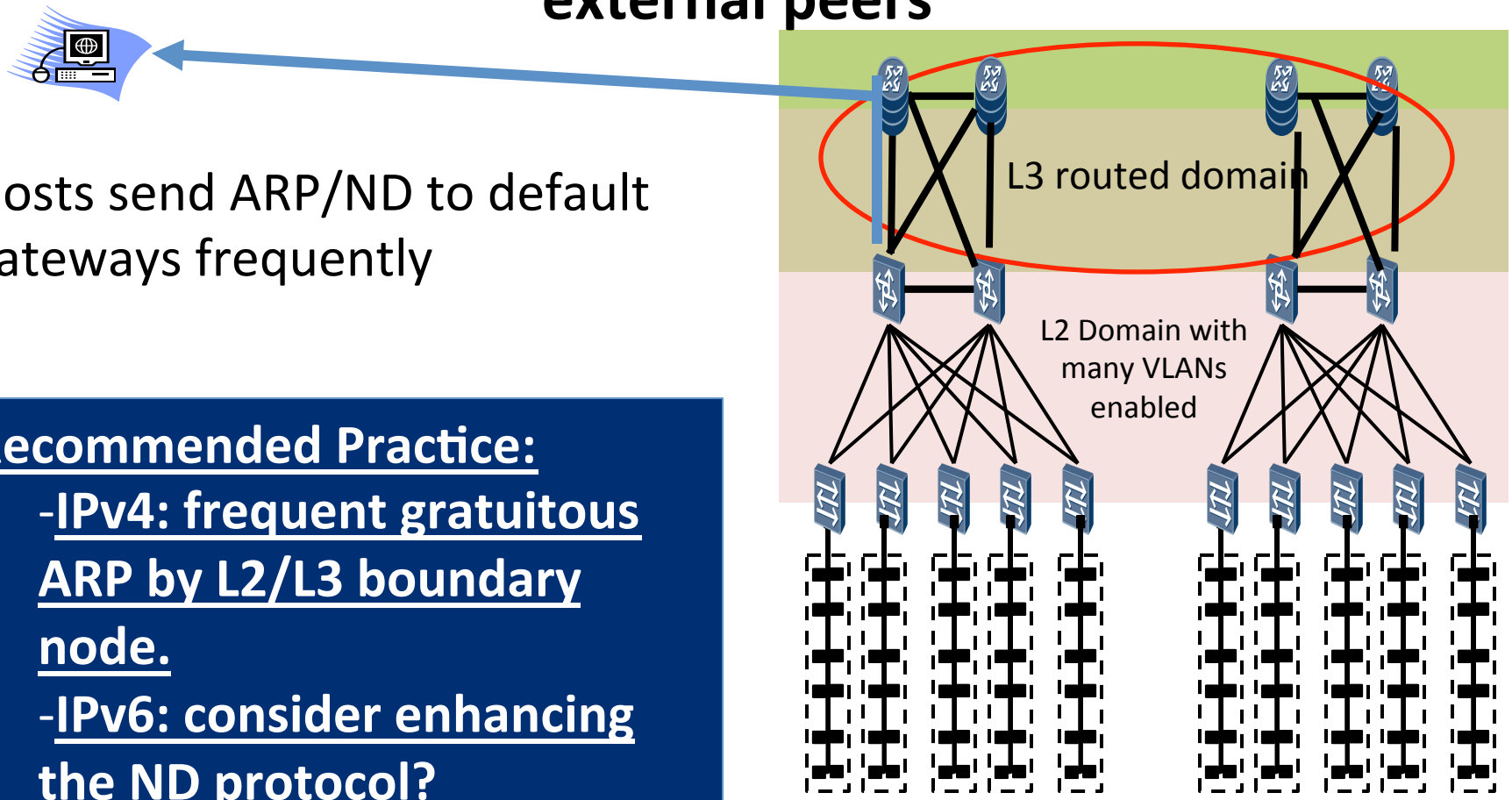
L3 Routed Access



When server is loaded with new applications, it has to inherit the same IP subnet

IP addresses have to be reconfigured when VMs move to a different rack

BCP for Scenario #2:
When internal hosts need to communicate with external peers



Hosts send ARP/ND to default gateways frequently

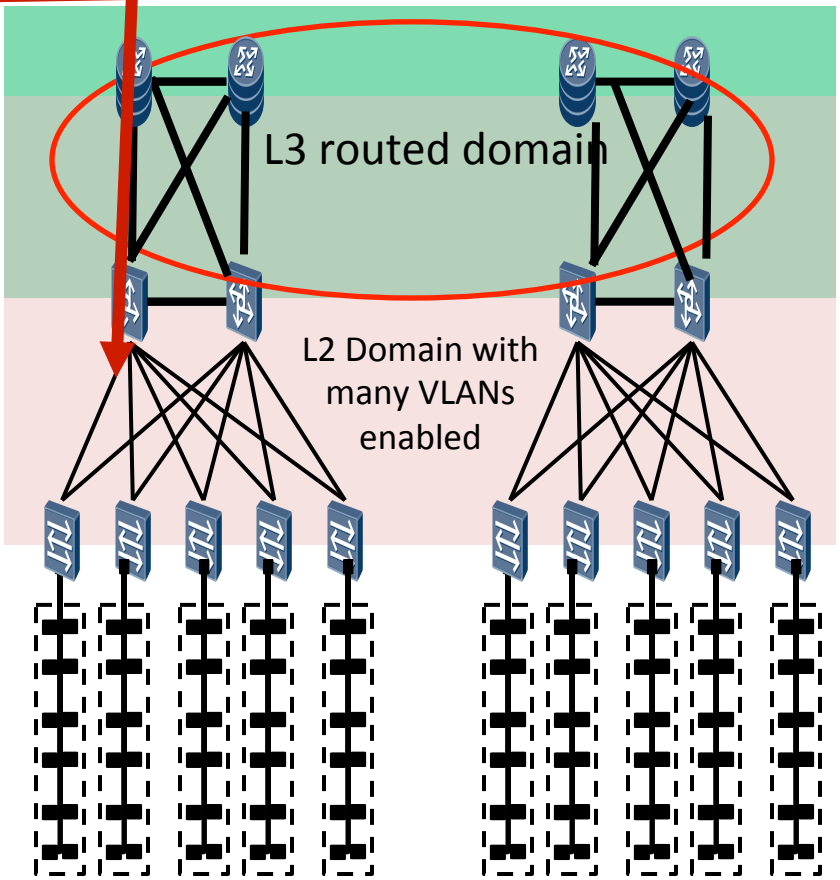
-**Recommended Practice:**
- -**IPv4: frequent gratuitous ARP by L2/L3 boundary node.**
- -**IPv6: consider enhancing the ND protocol?**

BCP for Scenario #2: When external peers initiate communication with hosts inside data center

-Issues:

- -L2/L3 boundary node needs to hold data frames,
- -Trigger ARP/ND to validate if the target exists in the L2 domain
- -When response is received from the target, send the data frames to the target

→CPU & buffer intensive.



-Recommended Practice:

- -L2/L3 boundary node proactively snoop gratuitous ARP/ND messages from local hosts.

Static Address Mapping

- In a data center, applications placement to servers, racks, and rows are orchestrated by Server (or VM) Management System(s)

-**Recommended Practice:**

-**Directory pushing down static ARP/ND mapping entries to all L2/L3 boundary nodes. Or**

-**Have access switch re-direct ARP/ND requests to Directory Server(s)**

DNS Based Solution

- Applicable to DC environment where hosts get their addresses from DNS

-**Recommended steps when a VM is to be moved to a new location:**

–Instantiate the service on a VM in a distant rack. The new VM gets a new IP address

–Change the address of the service in DNS

–Wait for the DNS TTL to expire. While you are waiting, watch the number of connections to the new VM increase and the number of connections to the old VM decrease.

–Wait a little longer. When the number of connections to the old VM reaches zero, shut down the old VM.

ARP/ND Proxy approaches

- ARP proxy defined by RFC 1027 (defined in 1987)
- “ARP Proxy” with ToR switch intercepting ARP requests and return the target hosts MAC if it knows it
- ARP/ND cache on local ToRs
- etc

-Recommendation:

-Have drafts in IETF to better define various types of ARP/ND proxy

Overlay Network

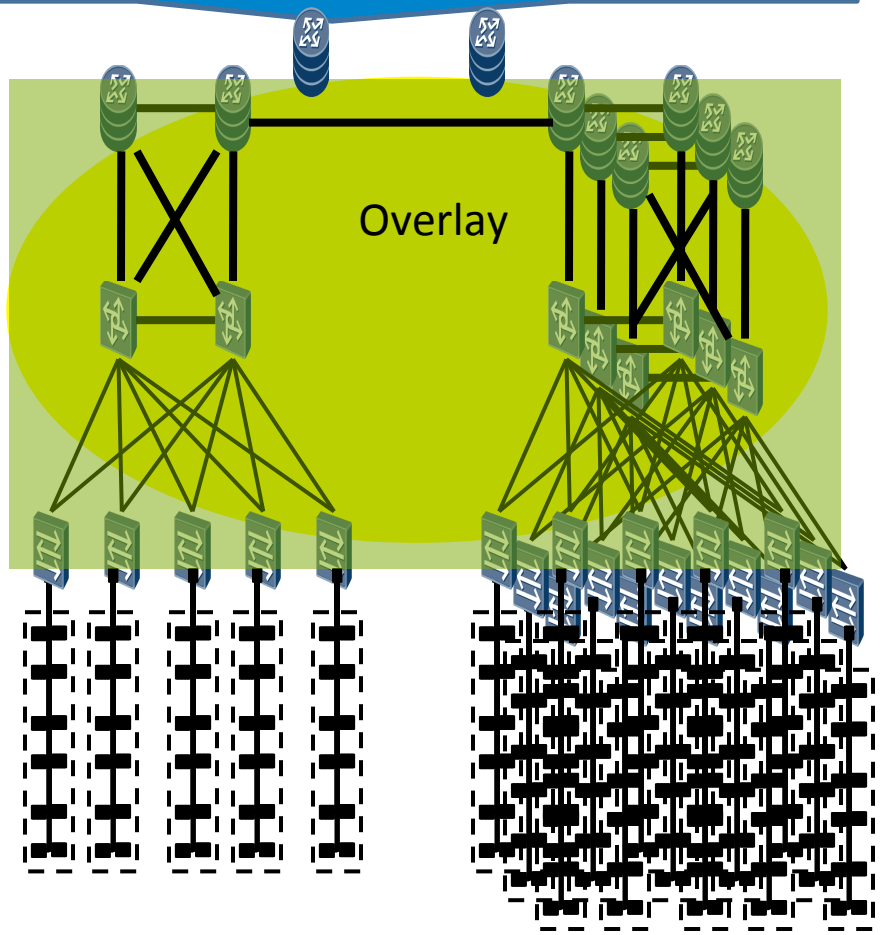
Hosts have different addresses than network addresses

When external peers communicate with internal hosts:
Gateway routers have to resolve target address, plus Network Edge node

-Recommendation:

-Static mapping for all the overlay edge nodes

-Have multiple gateway nodes to share the address resolution



Thank you!