Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: January 6, 2012

T. Narten
IBM

July 5, 2011

Problem Statement for ARMD
draft-narten-armd-problem-statement-00

Abstract

This document examines problems related to the massive scaling of data centers. Our initial scope is relatively narrow. Specifically, we focus on address resolution (ARP and ND) within a single L2 broadcast domain, in which all nodes are within the same physical data center. From an IP perspective, the entire L2 network comprises one IP subnet or IPv6 "link". Data centers in which a single L2 network spans multiple geographic locations are out-of-scope.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction 3
2. Background 3
3. Out-of-Scope Topics 4
4. Address Resolution 5
5. Summary 5
6. Acknowledgements 6
7. IANA Considerations 6
8. Security Considerations 6
Author's Address 6

1. Introduction

This document examines problems related to the massive scaling of data centers. Our initial scope is relatively narrow. Specifically, we focus on address resolution (ARP and ND) within a single L2 broadcast domain, in which all nodes are within the same physical data center. From an IP perspective, the entire L2 network comprises one IP subnet or IPv6 "link". Data centers in which a single L2 network spans multiple geographic locations are out-of-scope.

This document is intended to support the ARMD WG identify its work areas. The scope of this document intentionally starts out narrow, mirroring the ARMD WG charter. Expanding the scope requires careful thought, as the topic of scaling data centers generally has an almost unbounded potential scope. It is important that this group restrict itself to considering problems that are widespread and that it has the ability to solve.

2. Background

Large, flat L2 networks have long been known to have scaling problems. As the size of an L2 network increases, the level of broadcast traffic from protocols like ARP increases. Large amounts of broadcast traffic pose a particular burden because every device (switch, host and router) must process and possibly act on such traffic. In addition, large L2 networks can be subject to "broadcast storms". The conventional wisdom for addressing such problems has been to say "don’t do that". That is, split the L2 network into multiple separate networks, each operating as its own L3/IP subnet. Unfortunately, this conflicts in some ways with the current trend of virtualized systems.

Server virtualization is fast becoming the norm in data centers. With server virtualization, each physical server supports multiple virtual servers, each running its own operating system, middleware and applications. Virtualization is a key enabler of workload agility, i.e. allowing any server to host any application and providing the flexibility of adding, shrinking, or moving services among the physical infrastructure. Server virtualization provides numerous benefits, including higher utilization, increased data security, reduced user downtime, and even significant power conservation, along with the promise of a more flexible and dynamic computing environment.

The greatest flexibility in VM management occurs when it is possible to easily move a VM from one place within the data center to another. Unfortunately, movement of services within a data center is easiest when movement takes place within a single IP subnet, that is, within a single L2 broadcast domain. Typically, when a VM is moved, it retains such state as its IP address. That way, no changes on the either the VM itself, or on clients communicating with the VM are needed. In contrast, if a VM moves to a new IP subnet, its address must change, and clients may need to be made aware of that change. From a VM management perspective, life is much simpler if all servers are on a single large L2 network.

With virtualization, a single server now hosts multiple VMs, each having its own IP address. Consequently, the number of addresses per machine (and hence per subnet) is increasing, even if the number of physical machines stays constant. Today, it is not uncommon to support 10 VMs per physical server. In a few years, the number will likely reach 100 VMs per physical server.

In the past, services were static in the sense that they tended to stay in one physical place. A service installed on a machine would stay on that machine because the cost of moving a service elsewhere was generally high. Moreover, services would tend to be placed in such a way as to encourage communication locality. That is, servers would be physically located near the services they accessed most heavily. The network traffic patterns in such environments could thus be optimized, in some cases keeping significant traffic local to one network segment. In these more static and carefully managed environments, it was possible to build networks that approached scaling limitations, but did not actually cross the threshold.

Today, with VM migration becoming increasing common, traffic patterns are becoming more diverse and changing. In particular, there can easily be less locality of network traffic as services are moved for such reasons as reducing overall power usage (by consolidating VMs and powering off idle machine) or to move a virtual service to a physical server with more capacity or a lower load. In today’s changing environments, it is becoming more difficult to engineer networks as traffic patterns continually shift as VMs move around.

In summary, both the size and density of L2 networks is increasing, with the increased deployment of VMs putting pressure on creating ever larger L2 networks. Today, there are already data centers with 120,000 physical machines. That number will only increase going forward. In addition, traffic patterns within a data center are changing.

3. Out-of-Scope Topics

At the present time, the following items are out-of-scope for this document.

Cloud Computing - Cloud Computing is broad topic with many definitions. Without a clear (and probably narrow) scoping of what aspect of Cloud Computing to include in this effort, it will remain out-of-scope.

L3 Links - ARP and ND operate on individual links. Consequently, this effort is currently restricted to L2 networks

Geographically Extended Network Segments - Geographically separated L2 networks introduce their own complexity. For example, the bandwidth of links may be reduced compared to the local LAN, and round-trip delays become more of a factor. At the present time, such scenarios are out-of-scope.

VPNs - It is assumed that L2 VLANs are commonly in use to segregate traffic. At the present time, it is unclear how that impacts the problem statement for ARMD. While the limit of a maximum of 4095 VLANs may be a problem for large data centers, addressing it is out-of-scope for this document. L3 VPNs, are also out-of-scope, as are all L3 scenarios.

4. Address Resolution

In IPv4, ARP performs address resolution. To determine the link-layer address of a given IP address, a node broadcasts an ARP Request. The request is flooded to all portions of the L2 network, and the node with the requested IP address replies with an ARP response. ARP is an old protocol, and by current standards, is sparsely documented. For example, there are no clear requirement for retransmitting ARP requests in the absence of replies. Consequently, implementations vary in the details of what they actually implement.

From a scaling perspective, there are two main problems with ARP. First, it uses broadcast, and any network with a large number of attached hosts will result in a large amount of broadcast ARP traffic. The second problem is that it is not feasible to change host implementations of ARP - current implementations are too widely entrenched, and any changes to host implementations of ARP would take years to become sufficient deployed to matter.

5. Summary

This document outlines the scope of the problem the ARMD effort is intended to address. It intentionally begins with a very narrow scope of kind of data center ARMD is focusing on. The scope can be expanded, but only after identifying shared aspects of data centers that can be clearly defined and scoped.

6. Acknowledgements

7. IANA Considerations

8. Security Considerations

Author's Address

Thomas Narten
IBM

Email: narten@us.ibm.com