ARMD
Internet Draft
Intended status: Informational Track
Expires: January 2012

M. Karir
Merit Network Inc.
Ian Foo
Huawei Technologies

October 24, 2011

Data Center Reference Architectures
draft-karir-armd-datacenter-reference-arch-00.txt

Status of this Memo

Copyright Notice

Abstract

The continued growth of large-scale data centers has resulted in a wide range of architectures and designs. Each design is tuned to address the challenges and requirements of the specific applications and workload that the data is being built for. Each design evolves as engineering solutions are developed to workaround limitations of existing protocols, hardware, as well as software implementations.

The goal of this document is to characterize this problem space in detail in order to better understand if there is any gap in making address resolution scale in various network designs for data centers. In particular it is our goal to peel back the various optimization and engineering solutions to develop generalized reference architectures for a data center. We also discuss the various factors that influence design choices in developing various data center designs.

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 0.

Table of Contents

1. Introduction

Data centers are a key part of delivering Internet scale applications. Data center design and network architecture is an important aspect of the overall service delivery plan. This includes not only determining the scale of physical and virtual servers but also optimizations to the entire data center stack including in particular the layer 3 and layer 2 architectures. Depending on the particular application requirements and scale, data centers can be designed in variety of ways. Each design is often a representation of which aspects of the problem were and were not relevant to the purpose of that data center. In this document we attempt to generalize the various design optimizations into a common generic architecture to facilitate the discussion of potential issues under a common framework.

2. Terminology

ARP: Address Resolution Protocol

ND: Neighbor Discovery

Host: Application running on a physical server or a virtual machine. A host usually has at least one IP address and at least one MAC address.

Server: a physical computing machine

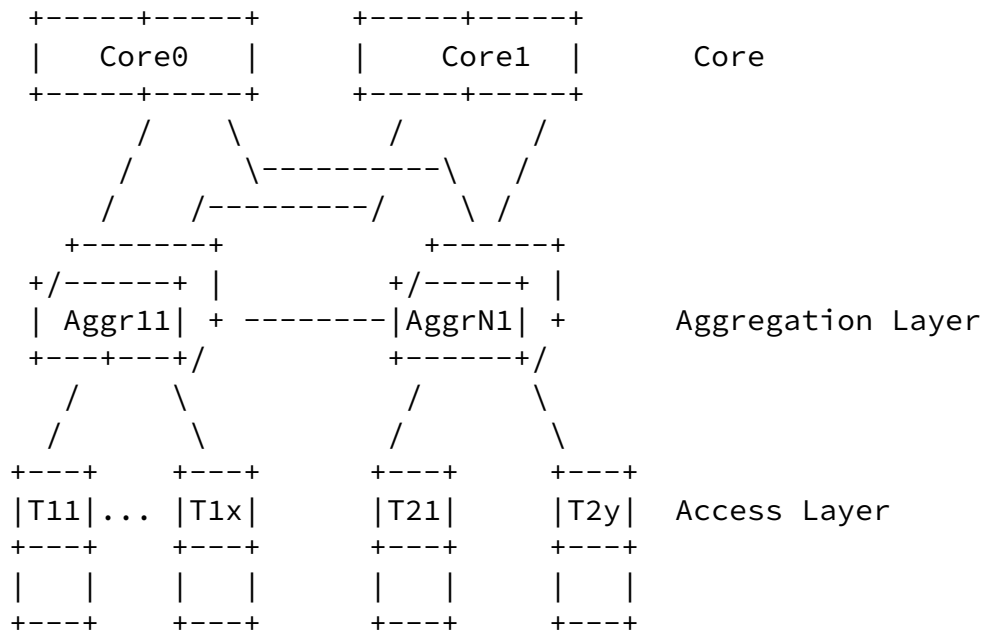
ToR: Top of Rack Switch

EoR: End of Row

VM: Virtual Machines. Each server can support multiple VMs.

3. Generalized Data Center Design

There are many different ways in which data centers might be designed. The designs are usually engineered to suit the particular application that is being deployed in the data center. For example, a massive web sever farm might be engineered in a very different way than a general-purpose multi-tenant cloud hosting service. However in most cases the designs can be abstracted into a typical three-layer model consisting of the Access Layer, the Aggregation Layer and the Core. The access layer generally refers to the Layer 2 switches that are closest to the physical or virtual severs, the aggregation layer refers to the Layer 2 - Layer 3 boundary. The Core switches connect the aggregation switches to the larger network core. Figure 1 shows a generalized Data Center design, which captures the essential elements of various alternatives.



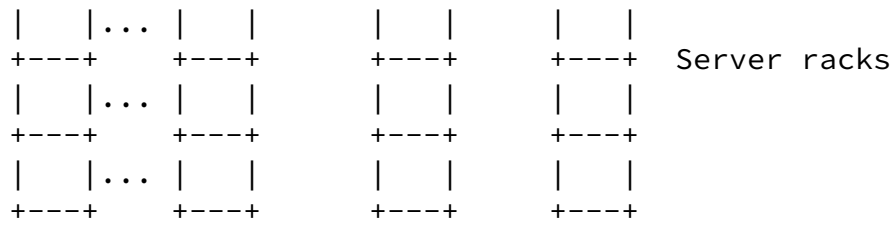


Figure 1: Typical Layered Architecture in DC

3.1. Access Layer

The Access switches provide connectivity directly to/from physical and virtual servers. The access switches might be placed either on top-of-rack (ToR) or at end-of-row(EoR) physical configuration. A server rack may have a single uplink to one access switch, or may have dual uplinks to two different access switches.

3.2. Aggregation Layer

In a typical data center, aggregation switches interconnect many ToR switches. Usually there are multiple parallel aggregation switches, serving the same group of ToRs to achieve load sharing. It is no longer uncommon to see aggregation switches interconnecting hundreds of ToR switches in large data centers.

3.3. Core

Core switches connect multiple aggregation switches and act as the data center gateway to external networks or interconnect to different PODs within one data center.

3.4. Layer 3 / Layer 2 Topological Variations

3.4.1. Layer 3 to Access Switches

In this scenario the L3 domain is extended all the way to the Access Switches. Each rack enclosure consists of a single Layer 2 domain, which is confined to the rack. In general in this scenario there

are no significant ARP/ND scaling issues as the Layer 2 domain cannot grow very large. This topology is ideal for scenarios where servers (or VMs) under one access switch don't need to be re-loaded with applications with different IP addresses or hosts don't need to be moved to other racks which are under different access switches. A small server farm or very static compute cluster might be best served via this design.

3.4.2. L3 to Aggregation Switches

When Layer 3 domain only extends to aggregation switches, hosts in any of the IP subnets configured on the aggregation switches can be reachable via Layer 2 through any access switches if access switches enable all the VLANs. This topology allows for a great deal of flexibility as servers attached to one access switch can be re-loaded with applications with different IP prefix and VMs can now migrate between racks without IP address changes. The drawback of this design however is that multiple VLANs have to be enabled on all

access switches and all ports of aggregation switches. Even though layer 2 traffic are still partitioned by VLANs, the fact that all VLANs enabled on all ports can lead to broadcast traffic on all VLANs to traverse all links and ports, which is same effect as one big Layer 2 domain. In addition, internal traffic itself might have to cross different Layer 2 boundaries resulting in significant ARP/ND load at the aggregation switches. This design provides the best flexibility/Layer 2 domain size trade-off. A moderate sized data center might utilize this approach to provide high availability services at a single location.

3.4.3. L3 in the Core only

In some cases where wider range of VM mobility is desired (i.e. greater number of racks among which VMs can move without IP address change), the Layer 3 routed domain might be terminated at the core routers themselves. In this case VLANs can span across multiple groups of aggregation switches, which allow hosts to be moved among more number of server racks without IP address change. This scenario results in the largest ARP/ND performance impact as explained later. A data center with very rapid workload shifting may consider this kind of design.

3.4.4. Overlays

There are several approaches regarding how overlay networks can make very large layer 2 network scale and enable mobility. Overlay networks using various Layer 2 or Layer 3 mechanisms enable interior switches/routers not to see the hosts' addresses. The Overlay Edge switches/routers which perform the network address encapsulation/decapsulation still however see host addresses.

When a large data center has tens of thousands of applications which communicate with peers in different subnets, all those applications send (and receive) data packets to their L2/L3 boundary nodes if the targets are in different subnets. The L2/L3 boundary nodes have to process ARP/ND requests sent from originating subnets and resolve physical addresses (MAC) in the target subnets. In order to allow a great number of VMs to move freely within a data center without re-configuring IP addresses, they need to be under the common Gateway routers. That means the common gateway has to handle address resolution for all those hosts. Therefore, the use of overlays in the data center network can be a useful design mechanism to help manage a potential bottleneck at the Layer 2 / Layer 3 boundary by redefining where that boundary exists.

4. Factors that Affect Data Center Design

4.1. Traffic Patterns

Expected traffic patterns play an important role in designing the appropriately sized Access, Aggregation and Core networks. Traffic patterns also vary based on the expected use of the Data Center. Broadly speaking it is desirable to keep as much traffic as possible on the Access Layer in order to minimize the bandwidth usage at the Aggregation Layer. If the expected use of the data center is to serve as a large web server farm, where thousands of nodes are doing similar things and the traffic pattern is largely in/out a large access layer with EoR switches might be of the most use as it minimizes complexity, allows for servers and databases to be located in the same Layer 2 domain and provides for maximum density.

A Data Center that is expected to host a multi-tenant cloud hosting service might have completely different requirements where in order to isolate inter-customer traffic smaller Layer 2 domains are

preferred and though the size of the overall Data Center might be comparable to the previous example, the multi-tenant nature of the cloud hosting application requires a smaller more compartmentalized Access layer. A multi-tenant environment might also require the use of Layer 3 all the way to the Access Layer ToR switch.

Yet another example of an application with a unique traffic pattern is a high performance compute cluster where most of the traffic is expected to stay within the cluster but at the same time there is a high degree of crosstalk between the nodes. This would once again call for a large Access Layer in order to minimize the requirements at the Aggregation Layer.

4.2. Virtualization

Using virtualization in the Data Center further serves to increase the possible densities that can be achieved. Virtualization also further complicates the requirements on the Access Layer as that determines the scope of server migrations or failover of servers on physical hardware failures.

Virtualization also can place additional requirements on the Aggregation switches in terms of address resolution table size and the scalability of any address learning protocols that might be used on those switches. The use of virtualization often also requires the use of additional VLANs for High Availability beaconing which would need to span across the entire virtualized infrastructure. This

would require the Access Layer to span as wide as the virtualized infrastructure.

4.3. Impact of Data Center Design on L2/L3 protocols

When a L2/L3 boundary router receives data packets via its L3 interfaces destined towards hosts under its L2 domain, if the target address is not present in the router's ARP/ND cache, it usually holds the data packets and initiates ARP/ND requests towards its L2 domain to make sure the target actually exists before forwarding the data packets to the target. If no response is received, the router has to send the ARP/ND multiple times. If no response is received after X number ARP/ND requests, the router needs to drop all those data packets. This process can be very CPU intensive.

When a local host under the L2/L3 Router's L2 domain needs to send a data frame to external peers, it usually sends ARP/ND requests to get the physical address (i.e. MAC) of the L2/L3 routers. Many hosts repetitively send ARP/ND requests to their default L3 gateway routers to refresh its ARP/ND cache. This requires default routers to process great number of ARP/ND requests when the number of hosts under its L2 domains is very large. For IPv4, gateway routers frequently sending out gratuitous ARP for all the hosts under its L2 domain to refresh their ARP cache for the default gateway's MAC address can mitigate this pain point. However, for IPv6 hosts need to validate bi-direction communication with the gateway router before sending any data frames. Therefore, unsolicited neighbor announcement from gateway router can't prevent hosts from sending ND repetitively.

When hosts in two different subnets under the same L2/L3 boundary router need to communicate with each other, the L2/L3 router not only has to initiate ARP/ND requests to the target's Subnet, it also has to process the ARP/ND requests from the originating subnet. This process is even more CPU intensive.

5. Conclusion and Recommendation

In this document we have described a generalized Data Center network design. Our goal is to distill the essence of different designs into a common framework in an attempt to structure the discussion regarding various scaling issues that might appear in different scenarios. Different application needs such as traffic patterns, and the role for which the data center is being designed determine various design choices, which result in various scaling issues with regards to port density, ARP/ND, VM mobility, and performance. As

expected, engineering solutions serve to tune a given design to the particular needs of the data center at the expense of other factors.

6. Manageability Considerations

This document does not add additional manageability considerations.

7. Security Considerations

This document has no additional requirement for security.

8. IANA Considerations

None.

9. Acknowledgments

We want to acknowledge the following people for their valuable discussions related to this draft: Kyle Creyts, Alexander Welch and Michael Milliken

This document was prepared using 2-Word-v2.0.template.dot.

10. References

[ARP] D.C. Plummer, "An Ethernet address resolution protocol." RFC826, Nov 1982.

[ND] T. Narten, E. Nordmark, W. Simpson, H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)." RFC4861, Sept 2007.

[STUDY] Rees, J., Karir, M., "ARP Traffic Study." MANOG52, June 2011. URL http://www.nanog.org/meetings/nanog52/presentations/Tuesday/Karir-4-ARP-Study-Merit Network.pdf

[DATA1] Cisco Systems, Data Center Design - IP Infrastructure , October 2009. URL http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_3_0/DC-3_0_IPInfra.html

[DATA2] Juniper Networks, Government Data Center Network Reference Architecture, 2010. URL www.juniper.net/us/en/local/pdf/reference-architectures/8030004-en.pdf

Authors' Addresses

Manish Karir
Merit Network Inc.

1000 Oakbrook Dr, Suite 200
Ann Arbor, MI 48104, USA
Phone: 734-527-5750
Email: mkarir@merit.edu

Ian Foo
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050, USA
Phone: 919-747-9324
Email: Ian.Foo@huawei.com

Intellectual Property Statement

Disclaimer of Validity

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.