Networking Challenges and Resultant Approaches for Large Scale Cloud Construction

David Bernstein
Cisco Systems, Inc.
daberns@cisco.com

Erik Ludvigson
Cisco Systems, Inc.
eludvigs@cisco.com

Accepted Paper

1st International Workshop on Grids, Clouds and Virtualization
WGCV 2009

held in conjunction with

4th International Conference on Grid and Pervasive Computing
GPC 2009

Geneva, 4-8 May 2009

Networking Challenges and Resultant Approaches for Large Scale Cloud Construction

David Bernstein
Cisco Systems, Inc.
daberns@cisco.com

Erik Ludvigson
Cisco Systems, Inc.
eludvigs@cisco.com

Abstract

Cloud Computing is a term applied to large, hosted datacenters, usually geographically distributed, which offer various computational services on a "utility" basis. Most typically the configuration and provisioning of these datacenters, as far as the services for the subscribers go, is highly automated, to the point of the service being delivered within seconds of the subscriber request. Additionally, the datacenters typically use hypervisor based virtualization as a technique to deliver these services. Providers who construct these datacenters run into a variety of challenges which are not common in ordinary-scale datacenters. Of specific interest is the unique demand placed on the underlying network. Many unique approaches are utilized to address these networking challenges, several of which are discussed in this paper.

1. Introduction

Construction of large scale datacenters has been a subject area of much research and industry specification. In the late 1980's when computers were accessible to departments of corporations and not just the centralized information technology (IT) groups, microcomputers were being deployed in large numbers. As IT operations grew in sophistication, especially in response to the emergence of the database and the advent of client-server computing, microcomputers (now called "servers") were relocated into the old computer rooms. The availability of networking equipment with standardized cabling further made it possible to put the servers in a specific room inside the company which became known as the corporate "data center".

The emergence of the Internet caused companies to need massive deployments of web servers with high speed Internet connectivity and accessory network devices such as firewalls and load-balancers. The complexity and expense of building these larger datacenters was not practical or affordable for many corporations so special companies started building very large facilities, called Internet data centers (IDCs), New practices were designed to handle the scale and the operational requirements of these larger datacenters.

Datacenter construction and operation has grown into a discipline with guidelines and standards published by various organizations, such as the Telecommunications Industry Association [1].

Today, the largest companies in the Internet have expanded their IDCs into planetary-sized systems for operating retail venues, all-Internet search engines, or the latest trend known as "Cloud Computing". The challenges placed on the network designers in constructing these datacenters is not well understood as the expertise is concentrated in just a few companies who have built them. Likewise, the techniques for delivering such solutions are not well understood. Cloud computing further challenges network designers as now the entire system is programmable by the subscribers so any generalized traffic pattern must be supported.

2. Size of the Challenge

Cloud datacenters, as we will call them from now on, have some of the following characteristics [2]:

Servers per Datacenter	10,000 – 100,000
No. of Datacenters	3 – 40
Internet/SP Peering Connection Points	10 – 100 x 100Mb – 1Gb via 10Gb Ethernet
Backbone Connection Points	2 – 8 x 40Gb – 80Gb via 10Gb Ethernet
Long Haul Datacenter Interconnect Capacity	N x 40Gb – 80Gb via OC192
Metro Area Datacenter Interconnect Capacity	200Gb – 600Gb via Metro DWDM

Table 1. Networking Characteristics of Cloud Datacenters

Putting this into a diagram, the Cloud Computing architecture looks like this:

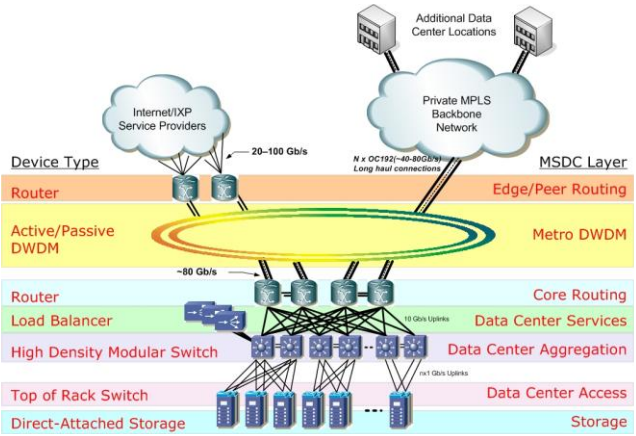

Figure 1. Cloud Computing per-location Datacenter Architecture

3. Server Interconnectivity (Top of Rack) / Data Center Access Layer Challenge

We will first look at the bottom tier of this topology, which is the Server layer. After all, for Cloud Computing, this is the “main act”, that is, to deliver computing. Overall, we must interconnect a lot of servers. As an example of how daunting a task this is at a high level, here are the total numbers of servers (estimated) in some of the larger Cloud providers:

Google	1M – 1.5M
Microsoft	500K – 1M
Yahoo	50K – 100K
Facebook	10K – 20K
Amazon	10K – 20K

Table 2. Servers in Cloud Datacenters

All of those servers, which have VM’s deployed on them, are where subscribers run their particular code. Efficiencies in Cloud Computing come from the fact that physical servers are “over-provisioned” to the extent that a single server may have many VM’s running on it. With a multi-core CPU architecture of typically two VMs/core [3], this is not a poor strategy, if the rest of the system can keep up with the generated load. The other systems which one must consider balancing are storage and memory.

As indicated in Figure 1, almost all of the Cloud Computing providers use Direct Attached Storage (DAS). Since DAS consists of in-box drive s/drive arrays, and in-box Host Bus Adapters (HBAs) storage generally scales linearly with servers, as you are added more storage every time you are adding another server.

3.1 Network Bandwidth from Servers Considerations

Networking is different, however, because it is cumulative upwards from servers to top of rack. To calculate how this must be provisioned from the network side, consider the following:

4 CPU Cores per CPU x 4 CPUs per 2U Server
= 16 Cores per Server x 2 VMs per Core
= 32 VMs per server

Table 3. VMs per Server in Cloud Datacenters

Applications running on a Cloud want to believe that they have the same throughput as if they were running on their own physical, hosted server. One might assume that we should model 1Gb networking throughput per VM as the canonical model these days for networking is a Gigabit Ethernet per server; however this is not the case as even the most streamlined OS and code path would be challenged to keep a network full to 400Mbps:

32 VMs per Server x 400Mbps per App
= 12.8Gb per server x 20 2U servers per rack
= 256Gb per rack

Table 4. Traffic per rack in Cloud Datacenters

Thus, in order to properly configure a rack of servers for Cloud Computing, one would need to install redundant 10Gb Ethernet to each server to a top of rack switch capable of 256Gb cumulative bandwidth. Usually, each rack is configured as a separate subnet, so that the top of rack switch also needs router uplink capability (Layer 3) with that aggregate throughput. This means 40 switch ports, likely 4 or 6 uplink ports, and two management ports, for 48 10Gb ports total.

3.2 Dense Network Connectivity Adds lots of Ports

Note also, we have calculated the port density based only on aggregate throughput to the servers. Our calculation has placed 4 data connections per server and although we’ve left management and uplink ports for the networking equipment, we have not included any calculation of management or “VM mobility” dedicated network. Typically [4] out-of-band network ports are dedicated for management and VM Mobility – two ports for VM Mobility and two ports for per-server management. Because one needs to copy running VMs using IP addresses not visible to the VMs, one needs addition ports for VM mobility. For management, one needs to boot machines, and so on, also adding ports. As anyone who has built a cloud knows, the physical number of NICs therefore becomes much higher than in traditional server deployments. Best practice would put 4 to 8 NICs for data depending on number of cores, 2 NICs for management, and 2 NICs for VM mobility. That's 12 NICs per server and a corresponding number of ports in the top of rack switch.

These would add 40 more ports even if we included just two of each type per server in our 20-server rack. Most Cloud providers do not support redundant, per server management and VM mobility for this reason. Clearly, a new switch with some other approach, as opposed to 96 ports per rack, would be highly desirable.

4. Server Storage Challenge

Curiously we do not see Clouds from the major providers using NAS or SAN. The typical Cloud architecture uses DAS, which is not typical of Datacenter storage approaches. This is because many of the Cloud operators have used an infrastructure which was originally developed for a specific application scenario where the data could specifically be made to take advantage of lower cost, distributed storage. For example, applications such as store catalog or search use specialized banks of servers and specialized lookup algorithms such as Map Reduce [5] which are designed to access in particular huge data sets with replicated data on a distributed topology. This is an artifact of the evolution of Cloud Computing.

4.1 The case for NAS or SAN

In a typical datacenter, applications are diverse and have many different storage requirements, including storage sharing amongst application processing elements, and management of storage lifecycle. As a result, Network Attached Storage (NAS) or Storage Access Networks (SAN) approaches are used which gives datacenter managers the highest flexibility through flexible use of pooled storage assets and connectivity. Many applications which a generalized Cloud Computing platform needs to support use smaller data sets and traditional structures such as filesystems or databases. In these cases NAS or SAN is preferred. Another use case where DAS is not appropriate is when a VM moves to another server yet still needs to access storage from the original server; in this case many of the management advantages of VM Mobility are lost because that server is still active. NAS or SAN with soft-reconfigurable WWN would be of great use here.

4.2 NAS/SAN adds More Ports

We have seen however that adding network ports such as Fiber Channel or additional Ethernet for iSCSI is impractical; this would add additional adapters with external connections per server and more switch ports. Adding 2 or 4 more Host Bus Adapter (HBA) ports for storage (for example, Fibre Channel) per server, now brings the port count per server to as much as 16.

5. Results from these Construction Techniques

Using the best datacenter construction techniques available, service providers such as Amazon have launched extremely popular services. However, in many cases the feature set, performance, and the variability of the resource or service is a disappointment as compared to the traditional enterprise datacenter.

5.1 Performance Results

Storage access and network performance are quite variable on clouds, and in terms of raw performance, not especially fast. One team [6] found that on AWS disk access (after eliminating "warm up effect" varied widely from 15mb/s to 74mb/s, although over 90% of writes occur at 40mb/s or greater, the most common rate they reported was 55mb/s, but with a wide standard deviation.

Network access varied similarly. The same team saw variations on sustained throughput (using "small" machines) from 50 Mbps to 550 Mbps but most commonly 350 Mbps. On "large" machines the network throughput was much better, varying from 550 Mbps to 900 Mbps, but most commonly near the high end of 800 Mbps. What is peculiar about this result is that the characterization of "small" and "large" machines relates to CPU capacity and is not supposed to reflect on network bandwidth.

5.2 Interconnect / Physical Construction Results

In addition to the variable performance, one has to cope with a difficult physical construction and cabling scenario. In the following examples, we've decided to model using a SAN, in addition to a VM Mobility enabled cloud. Also, we've added enough network switch capability to handle the full throughput of the multiple VMs running on the multi-core servers. To envision the repeated element here, consider the unit of the replicated server, with replicated SAN switches and SAN storage, replicated management and VM Mobility LANs, and replicated traffic LANs with enough horizontal NIC and switch ports to meet the capacity needs as outlined above. This unit of storage, network, and compute is illustrated in Figure 2.

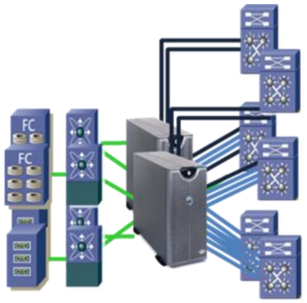

Figure 2. Cloud unit of storage, network, and compute showing “port explosion” phenomena

When one builds out the server tier in a cloud using blade server architecture, to provision 275 servers or so one needs 28 top of rack Ethernet switches, 14 storage switches, and a large rack-side aggregation switch as illustrated in Figure 3:

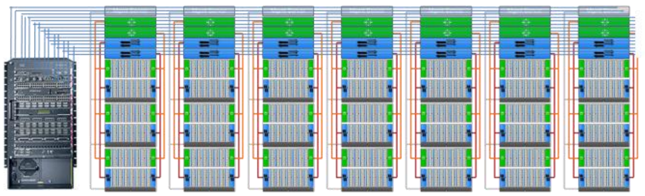

Figure 3. Cloud Blade Server tier of 7 racks of 40 blades/rack plus management blades

6. Data Center Access / Unified Fabric

Clearly a more efficient, shared networking architecture is needed for Clouds. Recently, a technology called Data Center Ethernet (DCE) has emerged [6] [7] which is a perfect fit.

6.1 Datacenter Ethernet

DCE has four main attributes.

(1) Channelization of Ethernet link: allows multiple traffic types (LAN, SAN, Management etc.) to be carried on a single link while providing partitioned bandwidth and transmission scheduling. Each channel may provide appropriate service to their apps (e.g. no-drop or low latency etc.)

(2) Congestion Management for channels: Not all traffic types expect "no-drop" behavior in the consolidated network. For traffic types which expect such reliable Layer 2 network (e.g. Storage traffic) DCE Ethernet allows congestion management.

(3) Guaranteed delivery (granular link level flow control): End to end congestion management helps network to adapt to long term (persistent) congestion. However, to deal with transient congestion, link level flow control is implemented with per-channel behavior.

(4) Increased bisectional bandwidth: Clouds are large! There is need for increased bandwidth in the core of these networks. DCE allows utilization of all the links in the network, allows shortest path bridging for the traffic and also allows Layer 2 multipathing for construction of Fat Trees.

Using a DCE based server interconnect, one can use just two physical connections to each server, and as channels on DCE, have for example 4 DCE Ethernet channels for data, 2 DCE Ethernet channels for VM Mobility, 2 DCE Ethernet channels for management, and 2 DCE Ethernet channels for storage using FCoE or iSCSI. This reduces the needed ports of our example of 20 servers in a rack from 200 ports to 40! Not only will one save in cabling and expense but in system management as well.

6.2 Fibre Channel over Ethernet

As we discovered, DAS based storage will lead to problems in a large cloud with generalized mobility and geographical dispersion. One would like to use a SAN. With a DCE substrate, one can use a particularly configured DCE channel and a software implementation of Fibre Channel for the SAN. This eliminates HBAs and SAN switches entirely.

7. Using DCE, FCoE in a Unified Fabric

We set out to utilize these main technologies, and build such a merged platform prototype. We eliminated the add-on element management server and implemented this right on the Linux system which runs the control plane of the switch All interconnectivity from the backplane fabric and extenders all the way to the network and the NICs were constructed out of DCE, using the management to set the behavior appropriate (backplane, LAN, FCoE) to the use.

The results: a drastic reduction of equipment needed to support the same computing, networking, and storage functionality – 275 servers or so with just 2 redundant top of rack switches! Additionally, it allows one to put critical security functions into a hardware switch, and created a performance profile which is much more uniform:

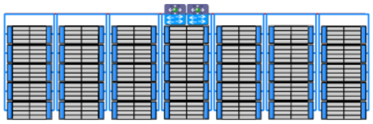

Figure 4. Merged Platform: 7 racks of 40 blades/rack w/embedded mgt – 1/3 the gear

Note that this effectively collapses the top of rack and aggregation layer. We consider this architecture a major breakthrough for Cloud builders.

8. Details of Inter-Datacenter Bandwidth / Core Routing Challenges

Moving a petabyte from one datacenter to another is no small task although it is commonly needed in a Cloud for example to replicate key databases or sets of servers geographically for Disaster Recovery purposes. Since 1 Petabyte = 8 Petabits, where in a typical architecture a datacenter supports 40Gbps of intercluster bandwidth it would take (8000 Gb /40Gb) = 200 Seconds or 3 minutes and 33 seconds of sustained 40Gbps traffic for one job under perfect conditions with no contention. However, this is never the case, links are always being used. To set up such an individual or more likely a set of jobs the current state of the network must be understood and decisions as to how to leverage under-utilized links is made by experienced network operators..

At that point, the network operators set up for the bursts manually using temporary reconfiguration of link aggregation through router Border Gateway Protocol [8] (BGP) or Multi Protocol Label Switching [9] (MPLS) traffic engineering. Once the flows are completely, the previous configuration of the network is restored. Current research is actively experimenting with autonomic backbone traffic engineering tools which can accomplish this without highly skilled network operators and manual configuration.

We do not propose a solution in this paper for this layer in the Cloud architecture but note it for later work.

9. Network Addressing

Interestingly, one area which imposes major challenges is network addressing. In a highly virtualized environment, IP address space explodes. Everything has multiple IP addresses; servers have IP addresses for management, for the physical NICs, for all of the virtual machines and the virtual NIC therein, and if any virtual appliances are installed they have multiple IP addresses as well.

9.1 Details of the Network Addressing Problem

Several areas are of concern here, on the one hand, the IPv4 address space simply starts to run out. Consider an environment inside the Cloud which has 1M actual servers. As explained above, assuming a 16 core server, each server could have 32 VM's, and each VM could have a handful of IP addresses associated with it (virtual NICs, etc). That could easily explode to a Cloud with well over 32M IP addresses. Even using Network Address Translation (NAT) [10], the 24-bit Class A reserved Private Network Range provides a total address space of only 16M unique IP addresses!

For this reason many Cloud operators are considering switching to IPv6 which provides for a much larger local address space [11] in the trillions of unique IP addresses. Switching to IPv6 is quite an undertaking, and some believe that switching from one static addressing scheme to another static addressing scheme (eg IPv4 to IPv6) might not be the right approach in a large highly virtualized environment such as Cloud Computing. If one is reconsidering addressing, one should consider the Mobility aspects of VMs in Cloud.

VM Mobility provides for new challenges in any static addressing scheme. When one moves a running VM from one location to another, the IP address goes with the running VM and any application runtimes hosted by the VM. IP addresses (of either traditional type) embody both Location and Identity in the IP address, eg, routers and switches use the form of the IP address not only to identify uniquely the endpoint, but by virtual of decoding the address, infer the Location of the endpoint (and how to reach that endpoint using switching and routing protocols). So while an addressing scheme is being reconsidered, let's consider two schemes which embody Mobility.

Mobile IPv4 [12] and Mobile IPv6 [13][14][15] mechanisms can be used in this case. Because IP addresses in either case are still provider-supplied and follow top level address allocations, we still find Vm mobility issues when a VM attempts more general mobility from one Cloud provider to another for example.

9.2 A Proposed Network Addressing Solution

In an attempt to completely generalize the addressing solution, a completely dynamic scheme where Location and Identification have been separated has been developed. This new scheme is called Location Identity Separation Protocol (LISP) [12]. LISP based systems can interwork with both IPv4 and IPv6 based networks, through protocol support on edge routers. However, internal to a Cloud, which may in itself span several geographies, LISP addressing may be used.

The basic idea behind the Loc/ID split is that the current Internet routing and addressing architecture combines two functions: Routing Locators (RLOCs), which describe how a device is attached to the network, and Endpoint Identifiers (EIDs), which define "who" the device is, in a single numbering space, the IP address. Proponents of the Loc/ID split argue that this "overloading" of functions places the constraints on end-system use of addresses that we detailed. Splitting these functions apart by using different numbering spaces for EIDs and RLOCs yields several advantages, including improved scalability of the routing system through greater aggregation of RLOCs. To achieve this aggregation, we must allocate RLOCs in a way that is congruent with the topology of the network. EIDs, on the other hand, are typically allocated along organizational boundaries.

Because the network topology and organizational hierarchies are rarely congruent, it is difficult (if not impossible) to make a single numbering space efficiently serve both purposes without imposing unacceptable constraints (such as requiring renumbering upon provider changes) on the use of that space. LISP, as a specific instance of the Loc/ID split, aims to decouple location and identity. This decoupling will facilitate improved aggregation of the RLOC space, implement persistent identity in the EID space, and hopefully increase the security and efficiency of network mobility.

To this end current experimentation is being done to assess the viability of using this protocol in conjunction with virtualization and in particular with VM Mobility. Of course, if and when LISP becomes a proven solution for the Cloud scenario, it must propagate into many forms of networking equipment which will take some time.

10. Security

In today's Cloud Computing offerings, there is very little network security. Because of the inability of the state firewall devices to scale to cloud like environments, only network attacks are protected against, for example DoS and DDoS attacks.

These are usually built into Load balancers, (Syn flood protection). ACLs are also used but become unwieldy. A better mechanism to manage and provide scalable security is required, because the firewalls and intrusion detection systems cannot handle the load/burst of a large cloud service Most clouds rely on the application security at the server level, versus network security, to put into place some kind of protection.

10.1 Details of the Security Problem

At the server or group of server level, this kind of security is usually referred to as domain isolation and policy enforcement. The idea is to separate individual customers provided specific environments (and separate connected subnets) with protected separated networks which are provisioned dynamically. Most customers are on a shared infra-structure (in other words there are no VLANs per environment in many cases). Even if the customer is provided a single environment, an ability to enforce policy on the environment, and provide VLANs per service within the environment is needed.

One popular technique which goes part way in this is to use the Netfilter/IP Tables [17] technique in the Linux operating systems which run as the base VM OS. The Netfilter framework enables packet filtering, network address and port translation and other packet manipulations. IP Tables is a generic table structure for the definition of rulesets. Each rule within an IP table consists of a number of classifiers. IP Tables are strung together in the related Linux kernels of the customers with the specified addresses allocated to that customer, thereby isolating their traffic through software, from the other operating systems of the other customers, on the same VM or same server.

Many are not happy with software domain isolation and policy enforcement, being used to VLAN and other hardware Layer 2/3 techniques in traditional datacenter architecture.

Another concern with the Netfilter approach is that each packet inspection must be done by the host CPU, thereby slowing overall system throughput. When firewall capability at Layer 4 is also desired, hardware acceleration is basically required as deeper packet inspection is performed and encrypted traffic is analyzed. Additionally, there is not Layer 2 isolation; one is using a straight shared Ethernet in the end. We would like to see a hardware switch understand and implement firewall at the VM to VM traffic level.

10.2 A Proposed Network Addressing Solution

We are experimenting with Dynamic Ingress Security Group Tag/ Role Based Access List (SGT/RBACL) filtering [18]. Here, switch hardware can be configured in an identical manner with Netfilter/IP Tables but use the Top of Rack and Aggregation switches to enforce the domain isolation with hardware.

This is a powerful answer to the immediate security issue, at least at the VM server level. On the one hand, it addresses the concerns of software-based domain isolation by having the switch hardware do the enforcing, and on the other hand addresses the performance issues by having the switch hardware do the deep packet inspection. We hope to report results of this speed-up technique soon.

11. Conclusion

Many networking challenges present themselves in constructing these new, planet-scale virtualized datacenters which are popularly called Cloud Computing. Some challenges including Data Center Access layer and Data Center Aggregation layer are readily addressed with DCE being a key enabling technology providing huge benefits in decreased port count and bandwidth increases. Some challenges in the Routing Core remain elusive as solutions are not port count or bandwidth related, they require advanced automatic traffic engineering. Other fundamental areas in IP infrastructure such as network addressing are "grand challenges" but strong work is in progress.

12. References

[1] TIA-942: Telecommunications Infrastructure Standard for Data Centers, *Telecommunications Industry Association*, http://www.tiaonline.org/standards/catalog/search.cfm?standards_criteria=TIA-942, (April 2005)

[2] Specifications from cloud providers include Amazon, Google, Yahoo, Facebook, and Microsoft (internal documents)

[3] Rodrigo Fonseca, George Porter, Randy Katz, Scott Shenker, Ion Stoica,. *X-Trace: A Pervasive Network Tracing Framework*, 4th USENIX Symposium on Networked Systems Design and Implementation -- NSDI 2007, Cambridge, MA, (April 2007)

[4] *Integrating Virtual Machines into the Cisco Data Center Architecture*, Cisco Solutions Document OL-12300-01, http://www.cisco.com/univercd/cc/td/doc/solution/vmware.pdf, (2008)

[5] Jeffrey Dean and Sanjay Ghemawat, *MapReduce: Simplified Data Processing on Large Clusters*, OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, (2004)

[6] W. Sobel, S. Subramanyam, A. Sucharitakul, J. Nguyen, H. Wong, S. Patil, A. Fox, D. Patterson. *Cloudstone: Multi-Platform, Multi-Language Benchmark and Measurement Tools for Web 2.0*. Proc.1st Workshop on Cloud Computing (CCA 08), Chicago, IL, October 2008

[7] Manoj Wadekar, *Enhanced Ethernet for Data Center: Reliable, Channelized and Robust*, Proceedings of the 2007 15th IEEE Workshop on Local and Metropolitan Area Networks (2007)

[8] The Data Center Bridging (DCB) Task Group (TG), a part of the IEEE 802.1 Working Group, a set of specifications found at http://www.ieee802.org/1/pages/dcbridges.html

[9] *A Border Gateway Protocol 4 (BGP-4)* and related other RFCs, at http://www.ietf.org/rfc/rfc4271.txt

[10] *Requirements for Traffic Engineering Over MPLS*, and related other RFCs, at http://www.ietf.org/rfc/rfc2702.txt

[11] *Address Allocation for Private Internets*, and related other RFCs, at http://tools.ietf.org/html/rfc1918

[12] *Unique Local IPv6 Unicast Addresses*, and related other RFCs, at http://tools.ietf.org/html/rfc4193

[13] *IP Mobility Support for IPv4, revised*, at http://www.ietf.org/rfc/rfc3344.txt

[14] *Mobility Support in IPv6*, at http://www.ietf.org/rfc/rfc3775.txt

[15] *Enhanced Route Optimization for Mobile IPv6*, at http://www.ietf.org/rfc/rfc4866.txt

[16] Carlos J. Bernardos, Ignacio Soto, and María Calderón, *IPv6 Network Mobility*, The Internet Protocol Journal, Volume 10, No. 2, June 2007

[17] *Locator/ID Separation Protocol (LISP)*, at http://tools.ietf.org/html/draft-farinacci-lisp-10

[18] at http://www.netfilter.org/

[19] for example, *Configuring Cisco TrustSec*, Cisco NX-OS Security Configuration Guide, OL-12914-03 found at http://www.ciscoconnects.com/en/US/docs/switches/datacenter/sw/4_0/nx-os/security/configuration/guide/sec_trustsec.pdf