8-2014

Tiered Based Addressing in Internetwork Routing Protocols for the Future Internet

Yoshihiro Nozaki

Tiered Based Addressing in Internetwork Routing Protocols for the Future Internet

by

Yoshihiro Nozaki

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctorate of Philosopy in Computing and Information Sciences

B.Thomas Golisano College of Computing and Information Sceince

Department of Computing and Information Science

Rochester Institute of Technology

Rochester, NY

August 2014

B. Thomas Golisano College of Computing and Information Sciences

Rochester Institute of Technology

Rochester, New York

Certificate of Approval

Ph.D. Degree

The Ph.D. Degree of Yoshihiro Nozaki

has been examined and approved by the dissertation committee

as satisfactory for the dissertation required for the

Ph.D. degree in Computing and Information Sciences

Approved by:

Dr. Pencheng Shi, Director of Ph.D. program Date

Committee Approval:

Dr. Nirmala Shenoy, Dissertation Advisor

Dr. Aparna Gupta, Dissertation Committee Member

Dr. Tae Oh, Dissertation Committee Member

Dr. Kaiqi Xiong, Dissertation Committee Member

Dr. Joseph Hornak, Dissertation Defense Chair

Abstract

The current Internet has exhibited a remarkable sustenance to evolution and growth; however, it is facing unprecedented challenges and may not be able to continue to sustain this evolution and growth in the future because it is based on design decisions made in the 1970s when the TCP/IP concepts were developed. The research thus has provided incremental solutions to the evolving Internet to address every new vulnerabilities. As a result, the Internet has increased in complexity, which makes it hard to manage, more vulnerable to emerging threats, and more fragile in the face of new requirements. With a goal towards overcoming this situation, a clean-slate future Internet architecture design paradigm has been suggested by the research communities.

This research is focused on addressing and routing for a clean-slate future Internet architecture, called the Floating Cloud Tiered (FCT) internetworking model. The major goals of this study are: (i) to address the two related problems of routing scalability and addressing, through an approach which would leverage the existing structures in the current Internet architecture, (ii) to propose a solution that is acceptable to the ISP community that supports the Internet, and lastly (iii) to provide a transition platform and mechanism which is very essential to the successful deployment of the proposed design.

The contribution of this work include design of the new Internet architecture that distributes the routing load across the routing domains based on the FCT concepts with new addressing scheme called Tiered Routing Address (TRA). New routing protocol called Tiered Routing Protocol (TRP) is also defined for the FCT architecture and compared with IP Routing which are both intra- and inter -domain by using simulation and testbeds to validate the

FCT architecture. In addition to design and validate the FCT concept, cost estimation model for the transition study is proposed.

Bibliographic Notes

Most of the work presented in this thesis appears in previously published journals and conference proceedings. The list of related publications are presented hereafter:

- Y. Nozaki, H. Tuncer, and N. Shenoy, "A Tiered Addressing scheme based on a Floating Cloud Internetworking model," in Proceedings of the 12th international conference on Distributed computing and networking, ICDCN'11, (Berlin, Heidelberg), pp. 382-393, Springer-Verlag, 2011.

- Y. Nozaki, H. Tuncer, and N. Shenoy, "ISP tiered model based architecture for routing scalability," In Communications (ICC), 2012 IEEE International Conference on, pp. 5817-5821. IEEE, 2012.

- Y. Nozaki, P. Bakshi, and N. Shenoy, "Tiered Interior Gateway Routing Protocol," In ICNS 2013, The Ninth International Conference on Networking and Services, pp. 68-75. 2013.

- Y. Nozaki, P. Bakshi, and N. Shenoy, "A Novel Approach to Interior Gateway Routing," International Journal On Advances in Networks and Services 6, no. 3 and 4 (2013): 208-219.

- Y. Nozaki, P. Bakshi, H. Tuncer, and Nirmala Shenoy. "Evaluation of tiered routing protocol in floating cloud tiered internet architecture," Computer Networks 63 (2014): 33-47.

- H. Tuncer, Y. Nozaki, and N. Shenoy, "Virtual domains for seamless user mobility," In Proceedings of the 9th ACM international symposium on Mobility management and wireless access, pp. 125-130. ACM, 2011.

- H. Tuncer, Y. Nozaki, and N. Shenoy, "Virtual mobility domains − A mobility architecture for the future Internet," In Communications (ICC), 2012 IEEE International Conference on, pp. 2774-2779. IEEE, 2012.

- H. Tuncer, Y. Nozaki, and N. Shenoy, "Seamless user mobility in Virtual Mobility Domains for the Future Internet," In Communications (ICC), 2012 IEEE International Conference on, pp. 5860-5865. IEEE, 2012.

Acknowledgements

I am most grateful to my dissertation advisor, Dr. Nirmala Shenoy, who expertly guided me through the time of my dissertation research. Her enthusiasm, encouragement, and faith in me throughout have been extremely helpful. She was always positive and gave generously of her time and vast knowledge. Without her guidance and persistent help this dissertation would not have been possible.

My application also extends to my committee members, Dr. Aparna Gupta, Dr. Tom Oh, and Dr. Kaiqi Xiong for serving as my committee members even at hardship. I would especially like to thank Dr. Aparna Gupta for her detailed comments and her recommendations regarding the economical study.

I would also like to thank all my colleagues, Hasan Tuncer and Yamin Al-Mousa, who shared research experiences with me from the beginning of the Ph.D. program, and Josh Watts, Alan Meekins, Arnav Ghosh, and Parth Bakshi, who contribute to the development of the FCT router in the testbed at RIT, and also to the movement of the software router to Emulab testbed.

And finally, I thank my parents for having given me the study abroad opportunity and the continuous support during my studing abroad in the US.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The current Internet has exhibited a remarkable sustenance to evolution and growth; however, it is facing unprecedented challenges and may not be able to continue to sustain this evolution and growth in the future because it is based on design decisions made in the 1970s when the earlier TCP/IP concepts were developed. The research thus has provided incremental solutions to the evolving Internet to address every new vulnerabilities, resulting in point-solution of patched work [1]. As a result, the Internet has increased in complexity, which makes it hard to manage, more vulnerable to emerging threats, and more fragile in the face of new requirements. With a goal towards overcoming this situation, which hinders significant progress, a clean-slate future Internet architecture design paradigm has been suggested by the research communities.

The alarming trends in the current Internet evolution eventually led to several clean slate Future Internet initiatives around the world such as the Future Internet Design (FIND) [2] and Future Internet Architecture Project [3] by the National Science Foundation (NSF) in the United States, the Seventh Framework Program (FP7) [4] by the European Union, AKARI project [5] by

Japan, and the *12th Five-Year Plan* projects [6] by the Ministry of Science and Technology (MOST) in China. These programs support research efforts that target challenges such as routing, scalability, mobility, security, and reliability among others - towards an ideal future Internet architecture.

While designing future Internet architecture, an important consideration in the design of Internet architectures are testing and validation of the design and scalability using realistic network scenarios in a near realistic experimental setup. The Autonomous Systems (ASs) and Internet Service Providers (ISPs) that construct the current Internet would not be willing to expose their networks to the risk of such experimentation, nor would they be willing to reveal information of their internal network topologies and implementations. Research communities have hence implemented open virtual large-scale testbeds using virtualization technologies. Large scale emulation and experimentation testbeds for this purpose are another effort sponsored and funded by major research organizations in the world; these include Global Environment for Network Innovations (GENI) [7] by NSF in the United States, the Future Internet Research and Experimentation (FIRE) project [8] a part of FP7 in the European Union, the Japan Gigabit Network 2 Plus (JGN2plus) [9] and the China Next Generation Internet (CNGI) [10] testbeds in Asia. New architectures can be evaluated and improved by testing on these testbeds before finalizing and deploying the future Internet architecture in the real world.

This research project is focused on addressing and routing for a clean-slate future Internet architecture, called the Floating Cloud Tiered (FCT) internetworking model which is supported by NSF FIND program. The major goals of this study are: (i) to address the two related problems of routing scalability and addressing, through an approach which would leverage the existing

structures in the current Internet architecture, (ii) to propose a solution that is acceptable to the ISP community that supports the Internet, and lastly (iii) to provide a transition platform and mechanism which is very essential to the ultimate successful deployment of the proposed design. Given the goals, it was decided to explicitly use the tiered relationships adopted by ISPs in their business models. ISPs have provider-customer and peer-peer relationships, and topological view of ISPs structure is rooted from tier-1 ISPs, which support several tier-2 ISPs, and tier-2 ISPs support several tier-3 ISPs and so on. Furthermore, tiered structure is also appeared within an ISP. An ISP network comprises of several Point of Presence (POPs), and three tiered clouds can be identified within each POP; a cloud of backbone routers (tier-1), which are connected to other POPs; a cloud of distributed routers (tier-2); and a cloud of access routers (tier-3), which may connect to stub networks or customer ISPs. Therefore, a tiered architecture model leverages the typical ISP tier structure.

The FCT architecture is based on the tiered models. The tiered relationships are based on a tiered structure that has the benefits of both hierarchical and distributed architectures. Each tier can have clear and well defined functionalities to incorporate and improve manageability and controllability, while still availing services from an upper tier and providing service to entities in a lower tier. A tiered architecture can help in better manageability and controllability as compared with the huge meshed structure exhibited by the current Internet. This simplicity allows us to easily manage, understand, and test networks because each tier level can have clear and well defined functionalities. Furthermore, fault isolation can be improved because it is easy to identify the abnormally behaving location in the network to help isolate possible failure points [11].

Geff Huston [12] observed the growth of BGP (Border Gateway Protocol) routing table sizes for years and showed that one of the main contributions for the increasing the table size is due to an increasing number of multi-homed ASs. Because many of the ASs are moving from a single-homed connection to multi-homing and peering, the BGP table size has rapidly increased as the result of an increasingly dense interconnected AS mesh at the edge of the Internet. Furthermore, logical links achieved by Multiprotocol Label Switching (MPLS) technology have introduced meshed topologies within an ISP. The Level 3's router topology presented in [13] is highly meshed because of this. Although flat and highly meshed network structures provide high redundancy, which comes at the cost of reduced efficiency as more and more complex routers are necessary to discover and maintain routes as the network grows in size, a fact that is apparently alarming when one notices the processing and operational conditions of core routers today [14]. Routing loops, looping packets, and high network convergence times are also the costs attributed to meshed network structures. Scalability is difficult to achieve under these conditions. Meshed networks are also hard to upgrade, troubleshoot, and optimize unless they are designed using a simple and hierarchical model [11]. Unlike mesh networks, the proposed tiered structure which optimally combines hierarchy and meshing provides a modular topology with good scalability. These advantages of tiered architectures let us recognize that it has the potential to address the issues that todays Internet has encountered.

In comparison, the tiered network structure described in this project adopts a tier-based routing and forwarding and a suitably designed tiered addressing scheme that reduces route maintenance by several magnitudes. The tiered structure does not cancel the benefits of the underlying meshed connectivity,

as they still continue to exist, operationally, optimally combining hierarchy and meshing. With the modularity introduced through the concept of network clouds and nesting, the structure affords a high level of scalability [15]. The tier concept is common among ISPs, but it can also be noted within an ISP network which comprises several POPs, inside which three tiers can be identified; tier-1 comprises backbone routers, tier-2 comprises distribution routers, and tier-3 comprises access routers. In the proposed design, each set of routers is identified as a network cloud and then associated to a tier.

Despite being proposed many more future Internet architectures from the research communities, most of them did not concern about transition and deployment mechanism. The future Internet architecture will be determined by the ability of the ISP organizations to adopt new infrastructure standards. The design takes into consideration the eventual transition and deployment through MPLS, where differentiation of tiers and tier-based forwarding can be achieved through labels and label stacking respectively. In this research, we also propose an economic model to study the adoption of the FCT architecture.

The contribution of this work include design of the new Internet architecture that distributes the routing load across the routing domains based on the Floating Cloud Tiered (FCT) concepts with new addressing scheme called Tiered Routing Address (TRA). New routing protocol called Tiered Routing Protocol (TRP) is also defined for the FCT architecture and compared with IP Routing which are both intra-domain and inter-domain by using simulation and testbeds to validate the FCT architecture. A Linux based router is implemented to support the tiered addressing and routing in a way which operates just above layer 2, and bypasses the IP layer and the IP routing protocols. The routers are deployed on the GENI testbed to run the routing protocol

on suitably designed topologies. In addition to design and validate the FCT concept, cost estimation model for the transition is proposed.

The reminder of this proposal is organized as follows. Chapter 2 describes the state of the art work in the area of future Internet initiatives and the current status of IP routing. Economic perspectives on the architecture transition are also discussed in this chapter. In Chapter 3, research questions that will be addressed as part of this study are presented. Chapter 4 describes the details of the FCT model, TRA, and operation of TRP. Details of the performance analysis and the evaluation also presented in Chapter 5. We also discussed the transition study and proposed transition cost estimation model in Chapter 6. Finally, conclusions and future plans are explained in Chapter 7.

Chapter 2

Literature Review

In this chapter, we briefly introduce research initiatives for the future Internet architecture. Next, the current Internet architecture and the main concept of intra-domain and inter-domain routing are briefly reviewed.

2.1 Research Initiatives for the Future Internet Architecture

There is a vast amount of very interesting research work conducted towards resolving the current Internet issues. We highlights only those that are closely related to our work in terms of the targeted goals and few projects that demonstrate a variety of approaches towards bringing solution to different aspects of the Internets scalability problem. We present the related work as two subsections: (1) solutions constrained by the existing Internet architecture and (2) solutions based on clean slate ideas for the future Internet [16].

2.1.1 Solutions Under the Current Internet Architecture

The Hierarchical Architecture for Internet Routing (HAIR) [17] targets limiting routing table size and decreasing churn rate by organized routing and applying a locator/identifier split approach in a hierarchical manner. The New Intern-Domain Routing Architecture (NIRA) adopts a provider-rooted hierarchy and extends the hierarchical properties to addressing, to reduce both the number of forwarding entries and convergence times [18]. The Routing Architecture for Next Generation Internet (RANGI) uses a locator/identifier split approach where a node's ID is different from its locator address to provide routing scalability [19]. The Hybrid Link State Protocol (HLP) leverages the natural hierarchy of the AS structures in route aggregation to reduce route churning [20].

An Internet Engineering Task Force (IETF) research group proposes a core-edge separation, address indirection, and a map-and-encap approach towards reduced routing table sizes [21]. The Routing on Flat Labels (ROFL) proposes a naming architecture, and a routing architecture based on flat identifiers that has no location semantics for both inter and intra-domain routing [22]. The Enhanced Mobility and Multi-homing Supporting Identifier Locator Split Architecture (MILSA) proposes a hybrid design combining the locator/ID split and core-edge separation paradigms to provide renumbering, routing scalability, and mobility support among others [23].

Another method adopted provides an overlay structure to route data packets efficiently, especially for the mobile Internet user. The General Packet Radio Service (GPRS) provides overlay routing in the local topology so that

the identity and location function of IP addresses remain the same while nodes are mobile [24]. The Internet Indirection Infrastructure (i3) Robust Overlay Architecture for Mobility (ROAM) provides a rendezvous based overlay indirection service that forwards data communication to the recent location of the mobile nodes efficiently [25].

2.1.2 Solutions Towards a Clean Slate Future Internet

Several clean slate future Internet projects were funded by NSF in the United States under its FIND program and subsequently under the FIA program. Some of these projects target routing scalability including the Floating Cloud Tiered Architecture [15] proposed by the authors. The eXpressive Internet Architecture (XIA) [26] aims to preserve the strengths of current Internet architecture while substantially improving security, and building in the ability to support evolving network functionality over time. XIA introduces a new protocol called XIP as a replacement for IP which introduces a new protocol stack, rich addressing and per-hop forwarding semantics [26]. The Mobility-First Project aims to address mobility, multihoming, connectivity robustness, context-aware routing and security by following key design principles such as separation of names from addresses, decentralized naming service, and a generalized delay tolerant network (GDTN) with storage-aware routing [27].

The FP7 European Future Internet Initiatives focused on several key areas including routing scalability [4]. Among the EU efforts, 4WARD [28] focuses on the creation of a future Internet architecture, and proposes a new connectivity paradigm called Generic Path (GP), which is mapped to a communication path for data propagation. The GP architecture relies on a new routing

scheme, Quality of Service (QoS) routing and resource Control (QoS-RRC). This mechanism is adopted to provide best network resources for reliable communication. Daidalos [29], another EU large-scale collaborative future Internet project, provides a virtual identity framework for a large number of users to access personalized services on seamlessly integrated heterogeneous network technologies with the help of an ID-Broker and an ID-Manager in a scalable manner.

The Japanese National Institute of Information and Communications Technology supports AKARI: Architecture Design Project for New Generation Network [30]. AKARI applies an ID/locator split and a cross layer design approach to support more diverse services, mobility and multihoming to a larger number of users through dynamic heterogeneous environments and devices. AKARI keeps the ID-locator mapping at the edge of the network to respond to the mobility and multihoming of the node while keeping the global locator based routing system in the core of the network for scalability purposes. For transition purposes they claim that the first 64 bits of the IPv6 address can be used as an ID and the remaining bits can be used as the locator.

2.2 Addressing in the Internet

The Internet today is an ubiquitous communications and information highway. As a result, computing devices all over the world connect to the Internet via local networks, each of which serves numerous devices and users. Among the millions of such networks that are supported in the Internet, the route discovery process is essential to establish communication links and maintain information flow between distant devices and networks. The discovery process

uses an IP address as a location identifier. However, the process becomes difficult because the IP address is a logical address that is allocated dynamically to a node and does not have any relationship with the actual location of the node. Further, the route itself is a path through the intricate mesh of networks that forms the Internet. If a network or a device fails, the connectivity information of thousands of networks and networked devices can be impacted causing very long network convergence delays and packet loss. Though robust due to redundant paths, the process of IP address allocation, combined with the high mesh connectivity has resulted in huge routing table sizes leading to routing scalability problems and its adverse impact on Internet performance such as high churn rates, high convergence times, and looping of packets amongst others.

2.2.1 IP Address

The original design of TCP/IP supported a maximum of 256 networks because it was believed that 256 networks would be sufficient [31]. IPv4 was then deployed to accommodate the growing number of networks through IP classes such as Class A, B, C, D, and E and eventually Network Address Translation (NAT) and Classless Inter-Domain Routing (CIDR) were introduced to cope with the growing demands. IPv6 was meanwhile developed to address the fast depletion of IPv4 addresses, to avoid address space fragmentation and to improve routing aggregation through hierarchical address allocation with a policy to avoid unnecessary and wasteful allocation [32].

Management of the IPv6 address space has been discussed in the Internet Assigned Number Authority (IANA) and Regional Internet Registries (RIRs). It has been recommended that the IPv6 address allocation should be done in

a hierarchical manner to avoid fragmentation of address space and to better aggregate routing information. Meanwhile, the IPv6 policy tries to avoid unnecessary and wasteful allocation [32]. It is difficult to avoid fragmentation and wasteful address allocation at the same time because future address requirements from organizations and end sites are unpredictable and most times exhibit an exponential increase [32]. Therefore, when an additional address space is required, a sequential address space may not be available and fragmentation of the address space is inevitable even with the huge IPv6 address pool.

2.3 Routing in the Internet

Internet Protocol (IP) provides best effort reachability for communication across networks and nodes connected to the Internet. In IP networks, routers use routing protocols to discover and maintain routes and also to recover from route failures. Routing tables maintained by current routing protocols increase almost linearly with increase in network size and is an unhealthy trend indicating scalability issues which can manifest as performance degradation. Also, the time taken for the network to adapt to topological changes increases with increase in network size resulting in higher convergence times during which routing is unpredictable and unstable. With more and more users connecting to networks today, this poses a serious problem. Patch and evolutionary solutions have been and are being proposed and implemented to address the problem both at the inter domain and intra domain level [33, 34].

IP Routing in the Internet is architected in two levels, intra-domain and inter-domain routing which is referred to interior gateway protocols (IGPs)

and exterior gateway protocols (EGPs) respectively. Intra-domain routing provides connectivity with in single routing domain network of a company, an organization, or an ISP, often referred to Autonomous System (AS). Inter-domain routing provides connectivity between ASs.

2.3.1 Intra Domain Routing

Interior Gateway Protocols (IGP) such as Routing Information Protocol (RIP) and OSPF were designed to work with IP. RIP is a distance vector (DV) protocol and can be used in networks with a maximum diameter of 15 hops. Large ISP networks thus use Link-State (LS) IGPs such as IS-IS or OSPF which uses the area concept to segment networks into manageable size. LS routing protocols require periodic updates and redistribution of updates to all routers in the network or in an area on link state changes. Each router running the LS routing protocol executes the Dijkstras algorithm on the collected link state information to populate routing tables. Dissemination of network-wide (or area-wide) link state information also adversely impacts scalability and convergence times in the networks using OSPF. In some cases the physical location of areas requires use of virtual links to the backbone area further limiting the versatility of OSPF.

IP based intra-domain routing protocols face two major challenges, one is scalability to increasing network diameter and the second is convergence to changes in link conditions as network size increases. RIP, the first widely used intra-domain routing protocol, is limited for operation in networks with a maximum diameter of 15 hops. First versions of RIP also had convergence issues which were overcome by Split horizon and Poison reverse. OSPF routing pro-

tocol, overcame the scalability limitations by introducing areas, where within each area a separate copy of the basic link-state routing algorithm could be run. Each area thus had its own link-state database, and the topology was invisible from outside of the area. This isolation enabled the protocol to reduce convergence times and the amount of routing traffic. Both routing protocols are IP based. RIP uses the distance vector approach, where routers record the next hop, towards other networks, and requires routers to advertise their routing tables. OSPF runs Dijkstras algorithm on network topology information collected by nodes. Advertising routing tables and requiring network wide link state information adversely impact scalability and convergence.

Significant research effort has been directed towards the reduction and optimization in IGP convergence time to link status changes in the network. In this research area, the approaches can be categorized into two: reducing failure detection time and reducing routing information update time.

To reduce failure detection time, layer-2 notification is used to achieve sub-second link/node failure detection. However this relies on types of network interfaces and does not apply to switched Ethernet [35].

Layer-3 notification is the more adopted method for link failure detection. For this purpose the Hello protocol is used. The hello protocol besides being used to disseminate neighbor information is also used to identify link/node failure in many routing protocols and is the layer-3 failure detection mechanism. OSPF sends hello packets to adjacent routers at an interval of 10 sec by default. The hello packet contains information on all links that a router is connected to. On missing four hello packets consecutively from a neighbor, OSPF routers recognize an adjacency failure with that neighbor router. Reducing hello packet interval time to sub-seconds can significantly reduce the

failure detection time, but at the expense of increased bandwidth usage due to increase in the number of periodic hello packets. Increased number of hello packets in a short interval can also increase possibility of route flaps.

Although link/node failure detection time can be reduced to sub-seconds, propagating the link status to all routers in the network takes time and is dependent on the network size.

To reduce such delays, an approach that suggests the use of several pre-computed back up routing paths was proposed. Pan et al. [36] proposed the MPLS based on a backup path to reroute around failures. However, having all possible MPLS back up paths in a network is not efficient. Multiple Routing Configurations (MRC) [37] uses a small set of backup routing paths to allow immediate packet forwarding on failure detection. A router in MRC maintains additional routing information on alternative paths. However, MRC guarantees recovery only from single failures. Liu at el. [38] proposed the use of pre-computed rerouting paths if the same can be resolved locally. Otherwise multi-hop rerouting path had to be set up by signaling to a minimal number of upstream routers. Another approach limits the propagation area of link state update after failure. Narvaez [39] proposed limited flooding to handle link failures. When a link failure occurs, the descendants of the failed link in the shortest path tree are determined and the new shortest path without the failed link is calculated. Then, the updated information is propagated in only the area of descendant nodes.

The two delays discussed above are significant. However, the SPF recalculation time can also be almost a second in large networks [35]. As packet loss/delay or routing loops occur during convergence, it is important to reduce this time. Novel routing approaches under the future Internet initiatives thus

provide the opportunity to view the routing problem from a fresh perspective and thus design solutions that are not constrained by the current architectures or implementations.

2.3.2 Inter Domain Routing

The Internet is comprised of more than 73,700 Autonomous Systems (ASs) all over the world today [40] and inter-domain routing maintains routes between those ASs. This high load in the core routers is indicative of an imbalance in the routing information handling, which could adversely impact the advantages of the meshed structure, by making the routers a potential bottleneck. Furthermore, the constant increase in routing table sizes is likely to become unmanageable in the near future. The complexity of BGP is reflected in both the exterior BGP (eBGP) and interior BGP (iBGP) as they make complex decisions that combine technical route criteria with policies and service level agreements across networks belonging to different ASs.

In general, ASs have either a customer-provider or a peer-to-peer relationship with neighboring ASs. A customer pays its provider for transit and peers provide connectivity between their respective neighbor ASes. Based on the AS relationships, the tiered structure and the hierarchy in the AS topology becomes obvious when looking at the Internet [41].

The de-facto standard inter-domain routing protocol in the Internet is the Border Gateway Protocol (BGP) ver4, which was specified in [42] on March, 1995. BGP aims at providing reachability among the Internet while supporting routing policies of ASs. For scalability reason, BGP does not maintain the entire Internets topology. BGP is a path-vector protocol which is populating

the end-to-end AS paths. This reachability information is learned by BGP sessions which are exchanging information between BGP routers in different ASs. This is referred as external BGP (eBGP) session. On the other hand, internal BGP (iBGP) sessions are established within the same AS to share the reachability information obtained by eBGP. The reachability information is built by BGP advertisement. The advertisement contains the prefix of the destination network and the complete AS path to the destination network. The simplified operation of BGP can be categorized into four components.

First, input and output of route advertisement is operated by each BGP session. The advertisement contains AS paths of each destination network. The second component is populating BGP routing table (RIB). This BGP routing table contains all possible distinct AS paths to each destination network learned by input of BGP advertisement. Next component is BGP decision process. When there is multiple AS paths entry for the same destination in the BGP routing table, BGP chooses one AS path to the destination based on the decision process. Unlike RIP and OSPF routing protocols that selects a route according to the shortest number of hops, the BGP decision process applies a sequence of rules to select the best route. Thus, BGP routing is more complex than simply choosing the shortest route. The rules in the decision process contains the local preference, the shortest AS path, the Multi-Exit-Discriminator (MED), and attributes for controlling traffic flows. The last component is building the forwarding table (FIB). This BGP forwarding table contains the best AS path of each destination network selected by the decision process, and this forwarding table is sent to the neighboring BGP routers via BGP sessions.

The area of inter-domain routing protocol has been considered as one of significant challenging research topics in the Internet today. As the Internet has grown largely, routing table size of BGP core router, the number of ASes in the Internet, and the number of connections per AS to the network are also increased significantly [40, 43]. As the result, slow convergence and lack of scalability have been recognized as main issues by researchers in inter-domain routing area [33].

The size of the BGP routing table at the core router today has exceeded 490,000 FIB (Forwarding Information Base) entries and 1,367,000 RIB (Routing Information Base) entries. Moreover, it is updated up to a million times a day [40]. In routers, TCAM (Ternary Content-Addressable Memory) and DRAM (Dynamic Random Access Memory) are used for storing the FIB and the RIB respectively. The maximum entry of the TCAM used in the routers, which are used by most ISPs, is 1 million routes with IPv4. However, TCAM is used for both IPv4 and IPv6 and a single entry of IPv6 occupies double the space of an IPv4 entry in terms of memory size. By default memory configuration on TCAM, maximum possible route entries for IPv4 and IPv6 are 512K and 256K. With current FIB entries increase rate, TCAM memory for IPv4 will be filled up very soon [44, 45]. On the other hand, a capacity growth of DRAM is faster than the growth rate of RIB entries, however, DRAM access speed grows only 10% per year. With over a million RIB entries, DRAM access speed contributes for BGP's slow convergence time. Moreover, to support high-speed packet forwarding with large routing tables, routers require high performance forwarding engine and expensive integrated circuit chips. Also, as the need for more powerful routers used in the BGP core increases, cooling

technology is more taxed. The current air cooling system is starting to be a limiting factor for scaling high-performance routers [14].

A convergence time is also one of the important performance metrics for a routing protocol. Measurements of BGP convergence time in the Internet was carried out by [46]. Their experimental measurement showed slow convergence that the BGP convergence time after a failure averaged around three minutes during the two years of their observation. The reason of the slow convergence of BGP is due to the size of the Internet. A single failure can force all BGP routers to exchange large amount of BGP advertisements (updates), while exploring alternative AS paths toward the affected destination (path exploration). To avoid exchanging massive BGP advertisements, BGP has a timer to prevent BGP routers from sending a new advertisement for a destination network if the previous advertisement of the same network was sent within 30 seconds. This timer is called Minimum Route Advertisement Interval (MRAI) [42]. MRAI can reduce the number of BGP advertisement during its convergence; however, it may introduce extra delay for the convergence time. Griffin and Presmore [47] tried to find the optimal value of MRAI, which is 30 seconds by default, with their experiments. Although optimal value of MRAI timer can reduce BGP convergence time significantly, it might be difficult to find in practical because the value is different by each network and topology. Storms of BGP advertisement can also be caused by flapping routers that regularly sending new BGP advertisement. To avoid this, BGP router ignores routes that change too often by using BGP route flap damping technique, however, it also increases BGP convergence time [48, 49]. Several solutions have been proposed to reduce the BGP convergence time while reducing the number of BGP advertisements. BGP-RCN [50] and G-BGP [51]

added location information of a root-cause into each BGP message when failure occurs. With this failure location information, distant BGP routers can avoid to select alternate AS path which is also affected by the failure.

2.4 Adoption of New Network Architecture

With the explosive growth of the Internet today, the scalability of current routing protocols has become a significant issue especially in inter-domain routing. Thus, researchers have proposed many solutions in past years. However, the replacement of the current inter-domain routing protocol, BGP, is a not realistic option due to its worldwide deployment. Furthermore, since these domains are completely autonomous entities such as ISPs, the proposed solution must be easy to deploy, which makes them appealing to ISPs, and efficiently balance the trade-off between their effectiveness and cost to implement [33]. Adoption of new product and technology has been widely studied in economics [52–55]. Economists have identified the new technology diffusion phenomena as that the diffusion will be based on increasing returns to adopters, benefit of adoption is a function of the number of current adopters in the industry, economies of scale may come out when costs decrease as volume increases, the increasing accumulated experience of using the technology will keep increasing to provide increasing returns to adoption [56–59]. Based on the above studies, Hovav et al. [60] proposed the Internet Standards Adoption (ISA) model. This model identifies two factors for an individual adoption decision that: usefulness of features (UF), how useful the technology is to organization, and environmental conduciveness (EC), how conductive the organizations environment is to adoption. The UF and EC can be represented as high and low states, and

modes of adoption based on the ISA model has four conditions that Status Quo (low UF, low EC), Niche (high UF, low EC), Replacement (low UF, high EC), and Full implementation (high UF, high EC). Using the modes, this model describes potential two paths to adoption of Internet standards: adoption through replacement and adoption through niche. However, this model does not perform simulation or an analytical study. Adoption of new network architecture is similar to the adoption of new technology. i.e. replacement costs and network benefits are important in both case. However, in case of adopting new technology, there are multiple organizations competing with each other to advance the technology. This, adoption of new technology depends on these competed organizations. For the adoption of new network architecture such as IPv6, they may not have opposing organizations. Therefore, opposition to a new network architecture may be organizations unwilling to invest the replacement cost. Joseph et al. [61] proposed an economic model to study the adoption of new network architecture. They use mathematical analysis and simulation to understand various factors on the adopting process such as network benefits, switching costs, and impact of converters. However, the proposed model is based on benefits offered by the new network architecture to a user, not to ISPs and infrastructure vendors.

Chapter 3

Research Questions

In the previous chapters, research challenges in both inter- and intra-domain routing protocols are presented in detail, both inter- and intra-domain routing are managed by those autonomous entities, which perform their own routing management based on policies that only have local significance. With this condition, new proposed solutions are difficult to implement or adopt in operational networks because it is too heavyweight to be deployed and standardization work is not well advanced, especially for inter-domain routing because of its worldwide deployment. Therefore, transition from current routing protocols to new routing protocol is not attractive to ISPs and ASs. Indeed, most of the existing proposals have never moved into a deployment stage. Thus, it is important to provide realistic and attractive scenario to the Internet service provider communities.

The questions that we intend to answer with our research are:

- How to address scalability issues facing on current routing protocols? In other words, how to decuple the dependency of routing table sizes from the network size?

  - Proposed new addressing scheme and routing protocol.

- The proposed solution is acceptable to the Internet service provider communities?

  - Implemented software-based router and evaluated it in testbeds.
  - Proposed an economic model to study the adoption of the FCT architecture.

Chapter 4

Methodology

In this chapter, we first describe the existing tiered structure among the ISPs and within an ISP, and introduce new addressing and routing scheme used in the FCT Internet architecture.

4.1 ISP Tiered Structure

The Internet is comprised of more than 73,700 ASs today that operate the major flow of Internet communication and the current IP traffic represents in a way their business relationships. Any AS must pay for transit services to get Internet connectivity. In general, ASs have either a customer-provider or a peer-to-peer relationship with neighboring ASs. A customer pays its provider for transit and peers provide connectivity between their respective neighbor ASs. Based on the AS relationships, the tiered structure and the hierarchy in the AS topology becomes obvious when looking at the Internet.

In the Internet, there are several tier 1 ISPs, who connect several tier 2 ISPs, as their customers, and the tier 2 ISPs connect the tier 3 ISPs as their

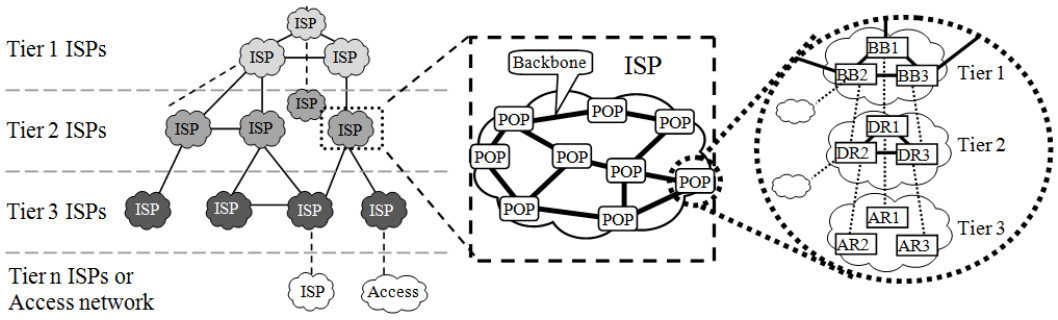


Figure 4.1: Typical ISP Tier Structure

customers. The left part of Figure 4.1 visualizes the tiered structure described above and existing among ISPs today. In Figure 4.1, we show ISPs up to tier 3, and then show access networks connecting to the tier 3 ISPs.

Inside of an ISP, there are several Point of Presence (POPs) which form the backbone of that service provider. Each POP has several routers, some of which are backbone routers that are primarily meant to connect to other backbone routers in other POPs. An interesting observation to be made at this point is the tiered structure that is also noticeable inside of an ISP POP (the dotted circle in Figure 4.1). Inside an ISP POP there is a set of backbone routers as shown in the BB cloud (we can associate them to be at tier 1 within the POP). The BB routers connect to the distribution routers (DR). The distribution routers in the DR cloud (which we can associate to be at tier 2) provide redundancy and load-balancing between backbone and access routers (AR). The ARs can then connect to customer or stub networks. The ARs and the stub network can thus be associated to tier 3.

4.1.1 Tiered Structure within an ISP

To validate the tiered approach within an ISP, we used the Rocketfuel dataset [13]. This dataset has router-level connectivity information of ISPs. From the Rocketfuel dataset, we imported the AT&T router connectivity information using Cytoscape [62] that also helps to visualize AT&Ts router-level topology on the US map (this excludes Hawaii and Alaska). The dataset contains not only the connectivity information, but also the routers location (city) information. Thus, we were able to map each router and city in the visualization shown in Figure 4.2.

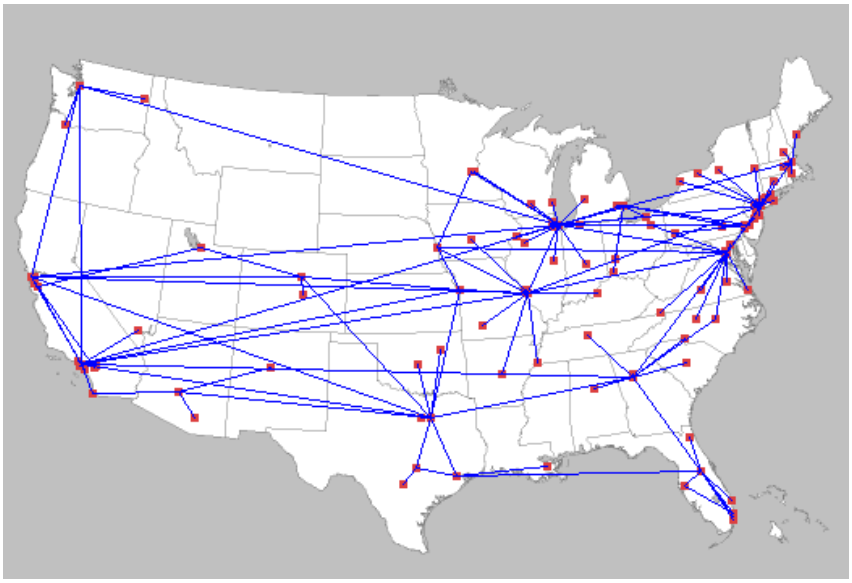
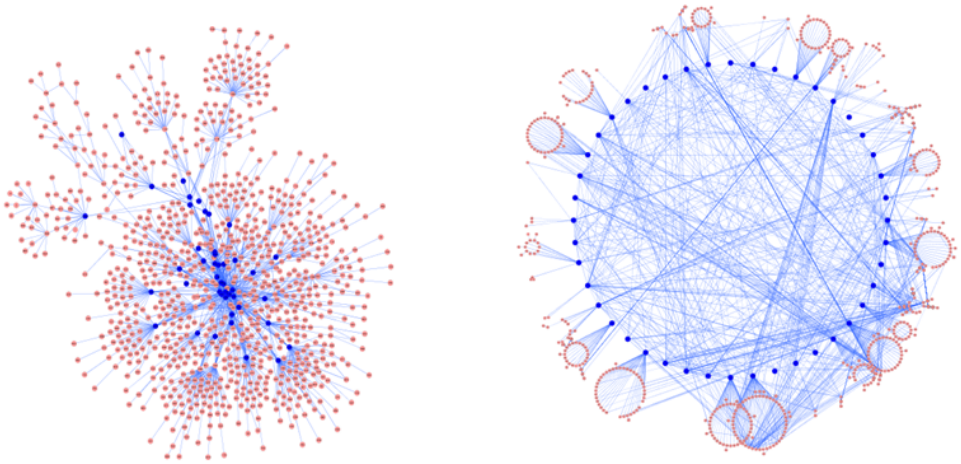


Figure 4.2: AT&T POP Level Network in the US

In total, 11,403 routers and 13,689 links interconnecting the routers were identified under this study. Each city in the topology visualization is a POP that has a large number of routers. A total 110 POPs were identified in the AT&T ISP network in Figure 4.2. In each POP, routers connecting with routers in other POPs were identified as BB routers.



a) Tree-like view b) Tiered-circle view

Figure 4.3: NY-POP Router-level network in AT&T

One of the biggest POP in the AT&T ISP network is the New York POP (NY-POP) which has 946 routers. Among these, 44 of them were identified as BB routers that have link(s) to other POPs. NY-POP router-level topology visualized as a tree structure is shown in Figure 4.3(a). The slightly large dots belong to a node (router) in the tree that has numerous branches. These routers are thus ideal candidates to be the BB routers in tier 1. Using Cytoscape, the visualization was changed to the one shown in Figure 4.3(b), where the BB routers now form the inner circle. From each BB router, routers that were one hop (or a maximum of 5 hops) were identified. These are the distribution routers can had multiple connections to the BB routers - they can be associated to tier 2. The edge routers are the access routers that were associated to tier 3 in the POP. Based on the NY-POP topology observation and the studies conducted, we could identify a total 44 BB routers, 542 DR routers, and 360 AR routers.

4.1.2 Tiered Structure among ISPs

To validate the tiered approach among ISPs, we used the Cooperative Association for Internet Data Analysis (CAIDA) dataset [63]. This dataset has AS-level connectivity information with inferred AS relationships. The dataset dated 01-20-2010 showed a total of 33,508 ASs and 75,002 AS links associated with provider-customer, peer-peer, and sibling relationships.

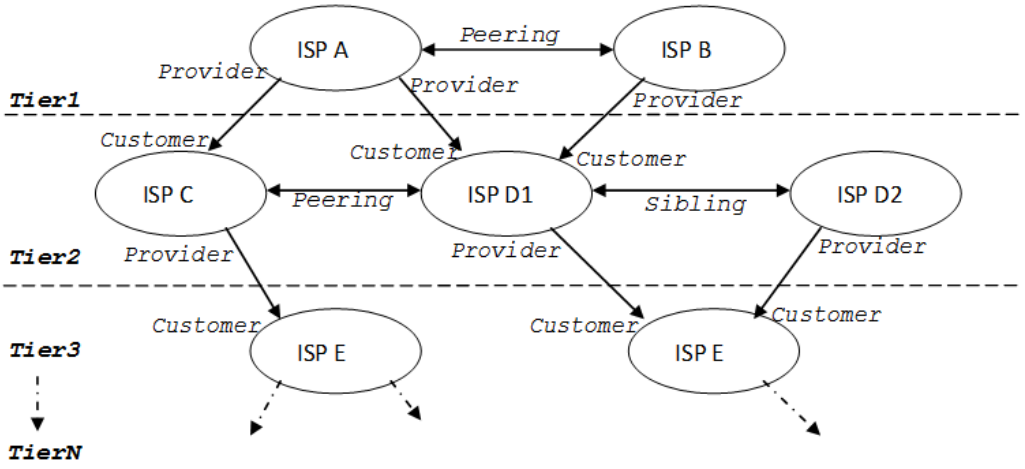


Figure 4.4: Business relationships between ISPs

Figure 4.4 depicts business relationships among ISPs. ISP A and ISP B have peering relationships. For example, ISP A is a provider of ISP C and ISP D1, ISP C is a provider of ISP E, ISP C is a customer of ISP A, ISP D1 and ISP D2 have sibling relationship that ISP D1 and D2 are same organization but having different AS numbers, and ISP D1 has more than one providers: called multi-homing.

The following strategy is applied to identify tiers among ASs:

1. Identify tier-1 AS

    - An AS which does not have any provider is recognized as tier 1 AS

2. Identify tier-N AS

   - An AS which has tier 1 AS are recognized as tier 2 AS. Then, continue the same approach till reach the last Tier-level. If an AS have multiple providers, multi-homing, (ex, tier 1 and 3 AS), the AS is recognized as lower Tier-level (ex, tier 2 AS)
   - If an AS does not have any provider but has peer relationship with tier-N, the AS is recognized as tier-N AS

3. Categorize ASs into two groups (Provider and Access AS)

   - If a tier-N AS does not have any customer, the AS is categorized as tier-N stub AS

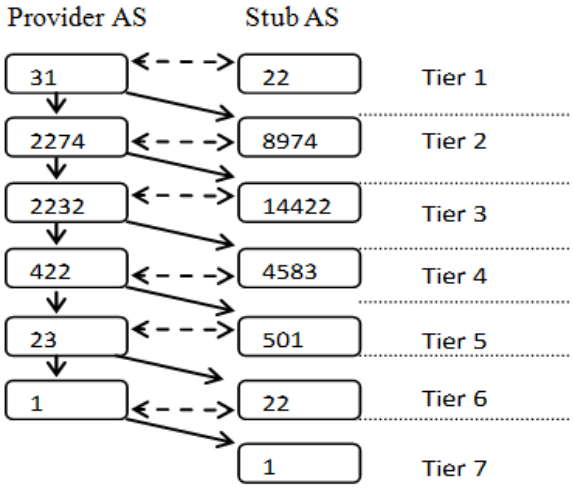


Figure 4.5: Worldwide Internet AS tiers

Figure 4.5 shows the different tiers existing in the Internet today, with the numbers of AS at each tier, with a count indicating the numbers of Provider AS and Stub AS at each tier. From Figure 4.5, there are totally seven tiers

in the ISP topology, with a single AS in tier 7. The majority of ASs are in tiers 2 and 3, accounting for nearly 83.2% of the ASs in the world. At tier 1, there are a total of 53 AS, of which 31 support customers and 22 who do not support any customer AS. This is around 0.16% of the total number of AS recorded by CAIDA. The tier 1 ASs are Level 3, AT&T and Verizon to name a few.

4.1.3 Nesting, Decoupling, and Floating Properties

Defining network clouds such as the set of backbone or border routers inside an ISP or AS network cloud is called a *nesting of clouds*. The network clouds defined within the ISP network or AS can also be associated with tiers defined within the ISP network cloud or AS cloud. For example, ISP has several POPs in Figure 4.1. In a magnified view of the POP on the right side of Figure 4.1, the network cloud comprising backbone routers can be associated with tier 1, the network cloud comprising distribution routers can also be associated with tier 2 inside the ISP network, and the network cloud comprising access routers can be associated to tier 3. Thus, a fresh set of tiers is started within a network cloud. This constitutes a *nesting of tiers*.

A network cloud can simultaneously connect to several parent or sibling clouds. If a cloud changes its relationship, only its external tiered address is changed. The internal address or structure can continue to remain the same if nesting is adopted. Nesting thus allows decoupling internal and external attributes /operations of a network cloud. The nesting feature enables network clouds to move across the tiers by simply changing or acquiring another

CloudAddr once they have an agreement with the concerned service providers. The architecture is thus named the Floating Cloud Tiered (FCT) architecture.

4.2 Tiered Routing Address (TRA)

To efficiently use the tiered structure for packet forwarding and internetworking operations a tiered routing address (TRA) was introduced. TRA allocation depends on the tier level in a network and carries the tier value explicitly as the first field. The tier levels can be assigned as described above. In an ISP, routers closer to a backbone or default gateway have lower tier value and routers near the network edge have higher tier value. TRA can be allocated to a network cloud (that comprises of a set of routers used for a specific purpose, such as backbone, distributions and so on) or a router. They are however not allocated to a network interface. Network interfaces are identified by port numbers. However, a router or end node can have multiple TRAs based on its connection to several upper tier routers or networks. This helps to support multi homing.

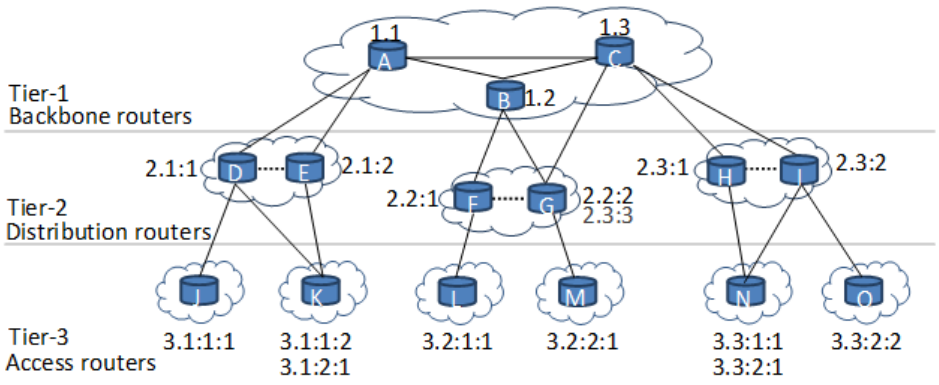


Figure 4.6: Example to Tiered Topology and Tiered Routing Address (TRA)

For example, three tiers are identified in the stub network of Figure 4.6, where each tier is allocated a *TierValue* (TV) from 1 to 3. TRA addresses are noted next to each router. TRA addresses start with a TierValue followed by : (colon) to separate the *TreeAddress* (TA). The . (dot) notation in the tiered address separates TierValue and TreeAddress. Address allocation starts from the routers at tier 1. Routers A, B, and C at tier 1 are allocated tier addresses {1.1}, {1.2} and {1.3} respectively. Note that tiered address assignment in TRP is *for a router*, and *not for each interface* in the router. This has advantages such as, reduced number of addresses, reduced routing table size, and ease in addition and removal of routers in the network.

The TreeAddresses of routers at tier 2 are allocated based on the TAs of their directly connected parent network node. In this study, it was assumed that all distribution routers have a link to one or more backbone routers, which may not always be the case. Due to the parent-child relationship between Routers A and D, Router D's TreeAddress is allocated by taking Router A's TreeAddress and appending a unique identifier for Router D. Hence, tiered address of Router D following the format {TierValue.TreeAddress} is {2.1:1}, where the first field in the TreeAddress is A's identifier and the second field is a unique identifier allocated to D by A. Likewise, Router E gets a unique identifier '2' from Router A and its tiered address is {2.1:2}. A link between routers which share a common parent is called a trunk-link. Links between Routers D-E, F-G, and H-I are trunk-links, and are represented with dotted lines in Figure 4.6.

Routers may have multiple parents and hence multiple addresses. For examples, Router K has two parents (Routers D and E) and hence has two tiered addresses {3.1:1:2 and 3.1:2:1}. Router G also has two parents and hence

has addresses {2.2:2 and 2:3:3}. When a router with multiple addresses has to allocate an address to a child router, it uses one of its addresses as a primary address and allocates an address to its child using the primary address. (This is an assumption made in this study, but can be relaxed or changed depending on the administrative policies within the AS) Thus, Router M that is a child of Router G has one TRA address {3.2:2:1}, where Router G decided to use its address {2.2:2} as the primary address. Decisions for selecting the primary address can be based on metrics of links associated with each address. Multiple addresses are useful for traffic engineering and for immediate recovery from link/node failure to reroute using the alternate address.

The logical view of tiered addressing in Figure 4.6 indicates a tree-like relationship where a tree is rooted at a tier 1 router. Hence the packet forwarding paths do not have loops and the address always refers to the shortest path to tier 1 as configured in this topology.

4.2.1 Nested TRA

As mentioned in Section 4.1.3, a tiered structure can be nested and TRA can also be nested. Figure 4.7 shows example and concept of nested TRAs. As seen in Figure 4.1, Network clouds defined within the ISP network or AS can also be associated with tiers defined within the ISP network cloud or AS cloud. In Figure 4.7, there are three ISPs, ISP1, 2, and 3, and each ISP has local tiered structure in their network. Based on ISPs relationship, ISP1 is a provider of ISP2 and 3, and global TRA addresses are assigned to each ISP. ISP1 has 1.1, ISP2 has 2.1:1, and ISP3 has 2.1:2. Each ISP has 3 tiered local router-level network and local TRAs are assigned.

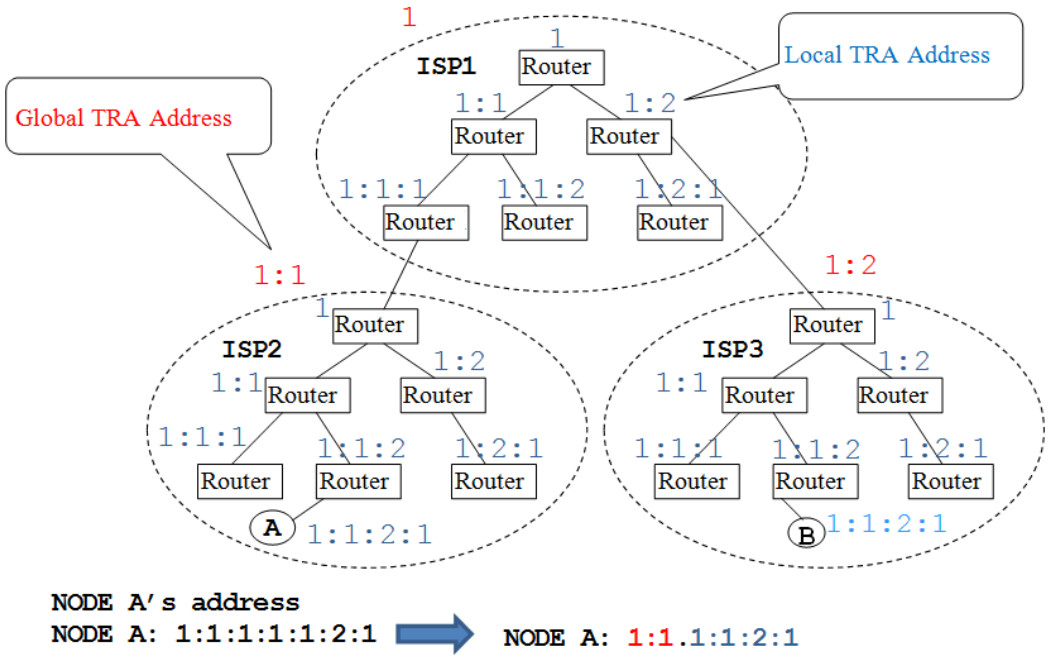


Figure 4.7: Example of Nested TRAs

Without nesting concept, TRA of Node A in ISP2 will be 7.1:1:1:1:1:2:1 that the TRA address is started from the top router in ISP1. If an address of the top router in ISP1 or topology of ISP1 is changed, it will affect all routers and nodes and required to change addresses. To avoid this situation, nested addressing can be applied. Local TRA address can be assigned from the top router of each ISP, so Node A in ISP2 has a local TRA, 4.1:1:2:1. If Node A wants to communicate with other local node in ISP2, this local TRA will be used. If Node A wants to communicate with a node outside of ISP2, combination of global TRA and local TRA are used. For a inter-domain communication, the global TRA address is used and when a packet reached the destination ISP, then local TRA address is used for the packet forwarding. With a nesting TRA concept, address change or topology change in ISP1 will

not affect to any node in ISP2 and 3 because address and topology information are summarized in the global TRA address of ISP1, which is not changed.

4.2.2 TRA Address Format

To test the FCT router and TRP, FCT packets are generated by the FCT router. The address format in a packet is shown in Figure 4.8. In the structure, the number of tiers can be very dynamic based on the network topology. We have assumed the TierValue to be 6 bits in size, which would allow for 64 levels within a topology. At each tier, the addressing scheme uses a Length Field (LF), and Address Field (AF) as shown in Figure 4.8. The AF length can be three different sizes; 4, 8, and 12 bits, and can support 16, 256, and 4096 router addresses in each tier, respectively. These different address field sizes are identified by the LF, which is located before each AF. The LF is a 2-bit static field, which can represent four cases, 00, 01, 10, and 11. The size of the address field will be 4 bits if LF = 01, 8 bits if LF=10 and 12 bit bits if LF = 11. This LF is eliminated in the Tiered Address notation. A LF value of 00 is used for special operations. It indicates the end of the address information, which allows dynamic depth of tiers and dynamic sizing of the address field. If at any level we need support for more than 4096 devices, we can add one more level and thus increase the number of devices or networks under a given tier from 4096 to 4096*4096. This is a recursive operation which can be used as required.

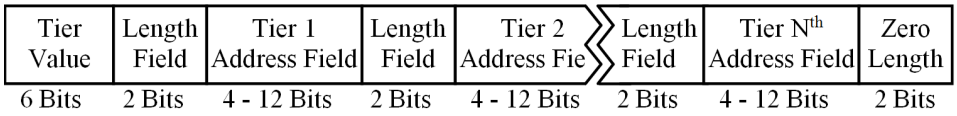


Figure 4.8: Address Format in FCT Packet

4.3 Tiered Routing Protocol (TRP)

We proposed new routing protocol that uses the tiered routing address (TRA) and adopts tiered based packet forwarding is called Tiered Routing Protocol (TRP). Operations of the TRP include TRA allocation, populating routing tables, packet forwarding, link / node failure detection and recovery.

4.3.1 TRA Allocation Process

TRP allows automatic address allocation by a direct upper tier cloud or node. Once tier 1 nodes acquire their TRAs (or have been assigned their TRAs, tier 2 nodes will get their TRA from the serving tier 1 node.

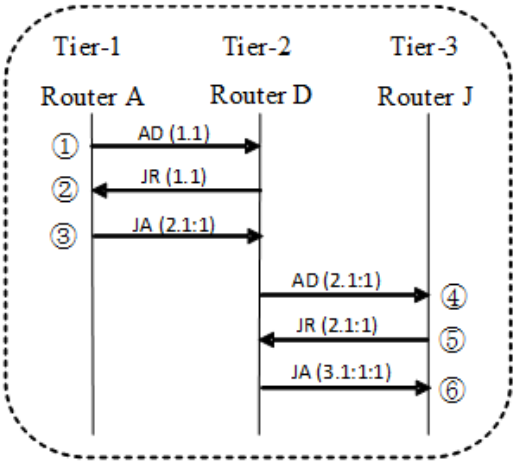


Figure 4.9: TRA allocation process

The process starts from the top tier i.e. tier 1. A tier 1 node advertises its TRA to all its direct neighbors. A node, which receives an advertisement, sends an address request and is allocated an address. For example in Figure 4.9, Router A with TRA 1.1 sends Advertisement (AD) packets to Routers B, C, D, and E. Routers D and E send Join Request (JR) to Router A because they do not have a TRA yet. Router B and C do not request address to Router A because they are at the same tier level. Router A allocates a new address (2.1:1) to Router D using a Join Acceptance (JA) packet. Another new address (2.1:2) is allocated to Router E. The last digit of the new address is maintained by the parent router i.e. Router A. Once Router D registers its TRA, it starts sending AD packets to all its direct neighbors and address assignment continues to the edge routers.

If a router has multiple parents, like Router G in Figure 4.6, it can get multiple addresses. A router with multiple addresses may decide to use one address as its primary address to allocate addresses to its children routers.

4.3.2 Populating Routing Tables

TRP maintains three routing tables based on the type of link it shares with its neighbors. In a tiered structure, links between routers are categorized into three different types: up-link which connects to an upper tier router; down-link which connects to a lower tier router; and trunk-link which connects to routers in the same tier level. A router can identify the type of link from which the AD packet arrives by comparing its tier value with the tier value in the received packet.

Router F has three different types of links to Routers B, G, and L on port numbers 1, 2, and 3 respectively. Advertisement from Router B is received at port 1 and compared with the tier level of Router B (which is 1) and its own tier level (which is 2). Since tier level of Router B is less than tier level of Router F, the link connected on port number 1 is recognized as up-link and the information is stored in the up-link table. Likewise information about Router G is stored in the trunk-link table, and information about Router L is stored in the down-link table.

Table 4.1: Routing tables of Router F from Figure 4.6

Router F {2.2:1}					
Uplink		Down		Trunk	
Port	Dest	Port	Dest	Port	Dest
1	1.2	3	3.2:2:1	2	2.2:2 2.3:3

Table 4.2: Routing tables of Router G from Figure 4.6

Router G {2.2:2, 2.3:3}					
Uplink		Down		Trunk	
Port	Dest	Port	Dest	Port	Dest
1	1.2	3	3.2:2:1	4	2.2:1
2	1.3				

In Table 4.1 and 4.2, the port column shows the port number of the router and dest column shows the TRA of direct neighbor obtained from the advertisements. There are multiple entries against a single port in the trunk-link table of Router F because Router G has two TRAs. The routing table for Router G is also shown.

TRP does not require flooding of routing information in the network, nor does it perform any calculations based on received router advertisement and

hence there is least delay in populating the routing tables. The initial convergence time in TRP is significantly lower as just one advertisement packet is required from a connected neighbor. Due to these features, the number of control packets exchanged for updating routing information is very low.

4.3.3 Packet Forwarding

For a source node to send a packet to a destination node, the source node calculates a forwarding address. First, the TierValue of a common parent between itself and the destination node is calculated. For this purpose, the source node compares its tree address with the destination address. Assume that the source node is Router L and destination node is Router M in Figure 4.6. Router L compares {3.2:1:1} with {3.2:2:1}, from left to right to find the TierValue of a common parent. The only common part in these addresses is the first field after the TierValue, thus the common parent is available only at tier 1, i.e. the TierValue of a common parent is {1}. The calculated TierValue will be the TierValue in the forwarding address. Next, the TreeAddress in the forwarding address is the TreeAddress of the destination node from the point where the common value is obtained. Thus, the forwarding address is {1.2:2:1}. As another example, a forwarding address between source Router J {3.1:1:1} and the destination Router K {3.1:1:2} will be {2.1:2} because tier 1 and 2 of source and destination address are the same {1:1}, thus the TreeAddress of the forwarding address starts with tier 2 address of the destination {1:2} and the TierValue is {2} because the TreeAddress is at tier 2, thus the forwarding address is {2.1:2}.

Algorithm 1 Packet forwarding at router R and incoming packet P

```
if R.TierValue == P.T then
    if R.TA.last_tier == P.TA.1st_tier then
        if port = find(P.TA.2nd_tier, downlink_table) then
            remove(P.TA.1st_tier)
            P.TV ← P.TV + 1
            forward(P, port)
            return true
        end if
    else if R.TV == 1 then                                  ▷ at Tier 1
        if port = find(P.TA.1st_tier, uplink_table) then
            forward(P, port)
            return true
        end if
    else if R.TV − P.TV == 1 then
        if R.TA.parent_tier == P.TA.1st_tier then
            if port = find(P.TA.2nd_tier, trunklink_table) then
                remove(P.TA.1st_tier)
                P.TV ← P.TV + 1
                forward(P, port)
                return true
            end if
        end if
    else if R.TV < P.TV then
        discard(P)                                          ▷ wrong packet
        return false
    end if
end if
if port = find(uplink_table) then
    forward(P, port)
    return true
end if
discard(P)                                  ▷ no entry in routing tables
return false
```

The TierValue in the forwarding address is used to make the forwarding decisions by TRP. The decision to forward in a particular direction, up-link, down-link or trunk-link is done by the intermediate routers as they compare the TierValue in the forwarding address with their own TierValue. The pseudo

code for the forwarding process at a TRP router is provided in Algorithm 1. In Algorithm 1, R represents the router that processes the packet, P represents the incoming packet, TV and TA represent TierValue and TreeAddress in the tiered addresses that is associated to R and P. For example, a packet containing the forwarding address {1.2:2:1} from Router L is sent to Router F. At Router F, it compares TierValue of the forwarding address {1} and its own TierValue of {2}. Since the TierValue of the forwarding address is smaller than Router F's value, a packet is forwarded upwards. A packet will be forwarded upwards until it reaches the same TierValue. In this case, a packet reaches Router B {1.2}. Then, Router B increments TierValue of the forwarding packet by 1 and removes the first TierValue {2}. The forwarding address is now {2.2:1}. From this forwarding address, Router B knows which down link port to forward the packet (which is port {2}). Thus, the packet is forwarded to Router G {2.2:2}. Router G makes the same forwarding decision of comparing the TierValue, then incrementing by 1, removing the first TierValue of the forwarding packet, which results in {3.1}. Finally the packet is forwarded to Router M. The packet takes a path of Routers L-F-B-G-M. There is another option to take the path Routers L-F-G-M because Router F could be made aware of trunk-link connection from its routing table.

4.3.4 Failure Detection and Handling

Failure detection in TRP is hello packet based, i.e. typical of layer 3 notification proposed for use with current routing protocols. In TRP, 4 missing AD packets is recognized as link/node failure. A TRP router tracks all neighbors AD packets times and if ADs from a neighbor is missing 4 consecutive

times, the TRP router updates its routing table accordingly. However, in TRP packet forwarding on link/node failure a router does not have to wait for the 4 missing AD packets. An alternate path, if it exists, can be used immediately on the missing a single AD packet irrespective of the routing table update. With the current high speed and reliable technologies, it is highly improbable to miss AD packets and redirecting packets on missing one AD packet is justified. However for a fair comparison with OSPF we adopted the 4 missing hallo packets to indicate a link/node failure.

Up-link Failure

If a node detects an uplink failure and has a trunk link, it can use the trunk link, because trunk link exists between routers that have the same parent route, or it can use an uplink is one exists. In Figure 4.10, the sibling router connected to Router F derives its address from the same parent. So, Router F knows that the uplink router on Router G will be its parent Router B.

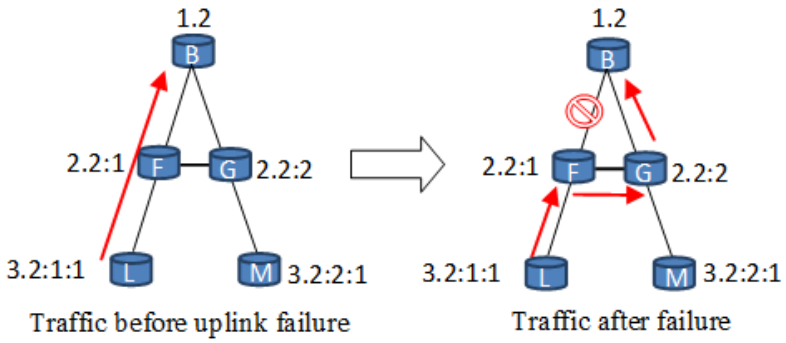


Figure 4.10: Failure handling with up-link

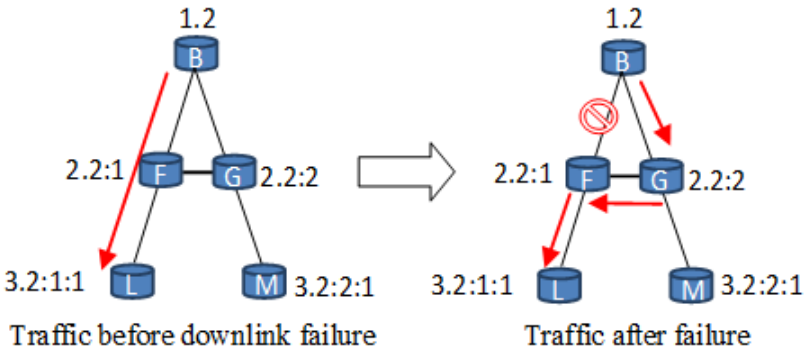


Figure 4.11: Failure handling with down-link

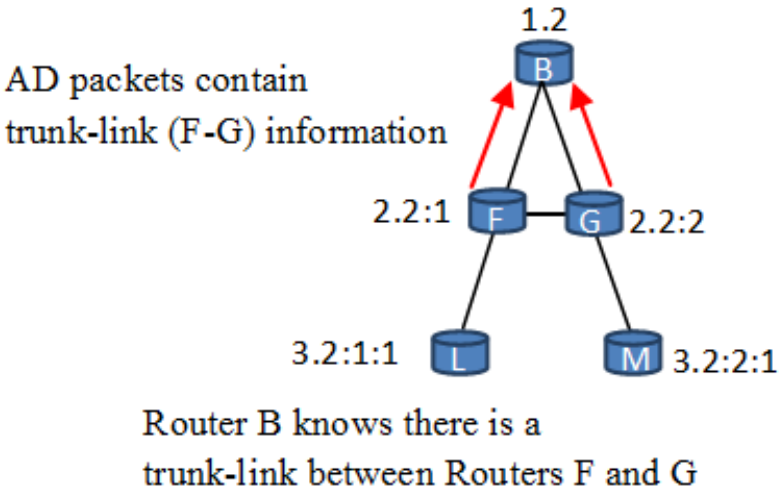


Figure 4.12: Trunk-link information sharing by the parent router

Down-link Failure

Let link failure occur between Routers B and F in Figure 4.11. To detour around the link failure, down link traffic between Router B and F needs to take a path Router B-G-F. To achieve this, Router B needs to know if there exists a trunk link between Router F and G. A parent router must know all trunk links between its children routers. The trunk link information can be set in AD packets to help a parent router maintain all trunk link information

as described in Figure 4.12. Due to inheritances, routers can assume responsibilities to forward for to their directly connected neighbors as the TRAs carry relationship information.

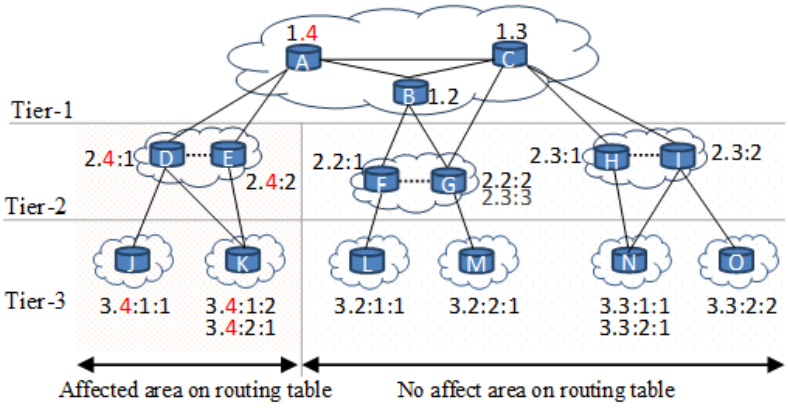


Figure 4.13: Address changes in TRP

Address Changing

Address changes can happen because of node failure, topology change, or administrative decisions. In TRP, address changes affect limited area and incur very low latency as no updates have to be propagated. For example, if Router A changed its TRA from 1.1 to 1.4 in Figure 4.13, all neighbor Routers B, C, D, and E notice the change from the AD packet sent by Router A. Router D and E will change their TRAs without notifying Router A. Therefore, children of Router A can change their addresses rapidly. The same procedure continues to Routers J and K by the next AD packet from Routers D and E. The pruning operation is triggered on change detection.

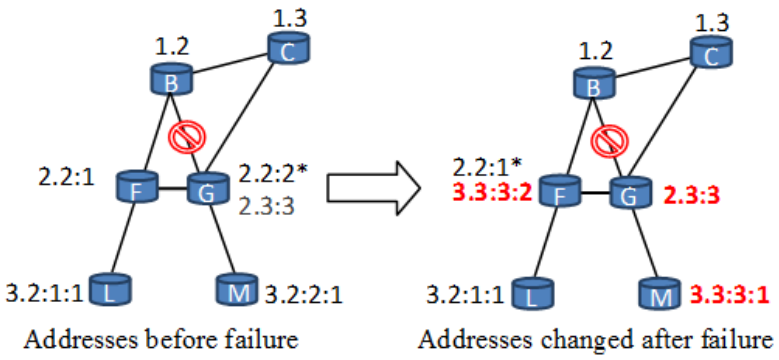


Figure 4.14: Primary address change

Primary Address Changing

If a node has multiple addresses and a link to a primary address failed, the node changes one of its secondary address to primary address and advertises the same. The child of the node also changes its address in the same manner as described in the case above and keeps the last digit. For example, Router G has two addresses and let 2.2:2 be the primary address in Figure 4.14. When a failure occurs between Routers B and G, Router G changes its primary address to 2.3:3 and then advertises it. As the result, Router M changes its address to 3.3:3:1.

4.4 Integration of Inter and Intra Domain Routing

The FCT model with TRA and TRP can be used for both intra- and inter-domain routing. Inter-domain routing protocol should be collaborated with all other ASs to support connectivity over the Internet, but choice of intra-domain routing protocol is depended on each AS and their network administrator. In

this section, we show possibility of integration between TRP and other routing protocol that TRP as inter-domain and another as intra-domain routing protocol.

4.4.1 MMT Routing in a Cloud

As part of our research we also investigated a robust routing scheme to be used for intra-cloud routing called Multi-Meshed Tree (MMT) routing [64]. Within a cloud, if we can grow trees, whose branches can be meshed and rooted at nodes that are either connected to an uplink cloud or a sibling cloud we will be able to forward the packets across the clouds. The growth of the tree is meshed yet loop free because of the numbering scheme used to generate the virtual IDs (VIDs) assigned to the nodes. The VIDs carry the branch information and also the route information to and from the root node. The meshed trees are created using local computations, in a distributed fashion and has very low complexity and overhead, hence is very robust.

	A	B	C	D	E	F
VID 1	1	11	12	111	112	123
VID 2		121	113	1211	122	1132
VID 3		1221	1123	1121	1212	1122
VID 4				1222	1131	1223
VID 5					1111	
VID 6					1231	

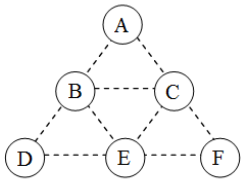


Figure 4.15: Example of MMT (Hop limit is 3)

The meshed tree can be built as shown in Figure 4.15 where the tree originates at the root node A which we refer to as Cluster Head (CH) and spread across all the nodes. Resulting VID allocations is shown in the table in Figure 4.15. Distance between CH and edge nodes is limited by 3 hops in

this example. Since each node allows having multiple VIDs, nodes in the tree can have redundant routes to the CH. To avoid loops in trees, VIDs are not assigned if there is already a child-parent relationship with particular VID. For example, if a node has VID 121, and joining node already has VID 12, the joining node will not request to get VID 1211. This VID acceptance rule applies for direct parent-child as well as any grandparents or grand children.

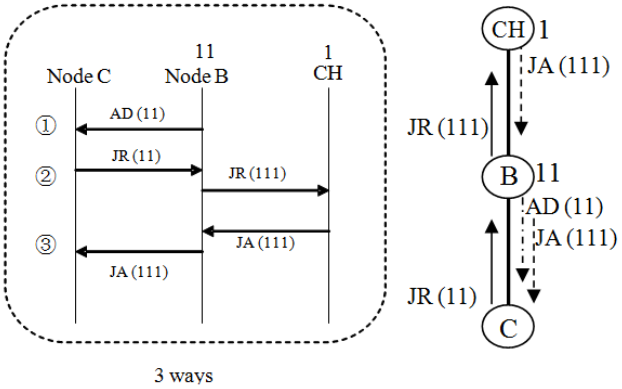


Figure 4.16: 3-ways Handshake in MMT Joining Process

Figure 4.16 describes the 3-ways handshake for meshed tree creation. Node B advertise its VID (11) through AD (Advertisement) packet to Node C, then Node C wants to join and sends JR (Join Request) packet for the VID (11) to Node B, Node B assigns new VID (111) for Node C and sends JR to Cluster Head (CH), then Cluster Head makes decision and sends JA (Join Acceptance) to node B, node B sends JA to node C, and node C updates its VID list. Now node C joins Cluster Tree. In figure 6, the ID of each node represents how MMT works to assign VIDs for nodes in one tree. Children inherent VIDs from parents and get adjunctive ID assigned by parents.

There are 4 types of nodes: Relay Node (has intra-cloud link), Up Node (has link to upper cloud), Trunk Node (has link to sibling cloud), Down Node

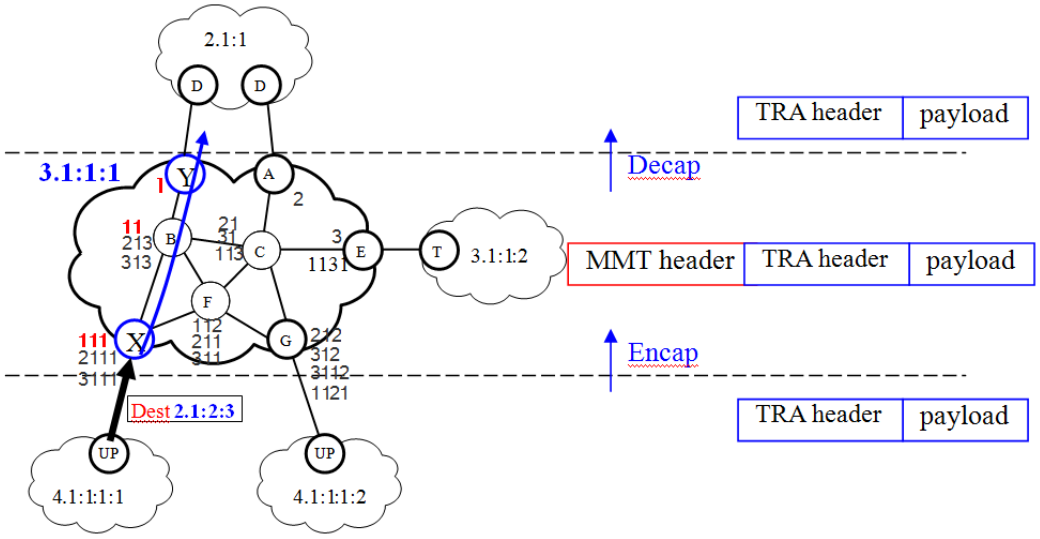


Figure 4.17: Data Forwarding with MMT

(has link to lower cloud). The nodes could also be combination of up, down, and trunk nodes, Besides Relay Node and Down Node, all other types of Nodes can be CH. Each CH connects to all other CHs. After MMT converges, every CH knows neighbor clouds information including path to the neighboring clouds. MMT is used when data packet comes to one cloud and packet is needed to be forwarded to other clouds. The edge node of this cloud which received this data packet needs to find the route to the node which connects to targeted cloud, and the path is provided by MMT routing information. In Figure 4.17, data packet is sent out from cloud ID 4.1:1:1:1 and destination is 2.1:2:3. Down Node with VID 111 receives the packet and makes decision to send to cloud 2.1:1 based on destination address. Data packet must be sent to Up Node. Down Node 111 finds path to forward packet to Up Node 1 based on VID 111. Even though MMT has very small size of routing information stored in each node (own VIDs), it provides robust and redundant routing scheme. The robustness provides correct routes, as well as redundancy.

4.5 TRP Code and Local Testbed

The TRP code was written to implement a Linux-based FCT router and evaluated using a three-tier testbed at the research lab in the organization of the authors. This local testbed used twelve commodity computers running Ubuntu Linux version 8.04.1 with kernel version 2.6.24. Each computer has total five network interfaces. One network connection is to the control network which is used for secure command line access to each machine and to serve as backbone for file transfers between machines for setting up experiments. The remaining four interfaces are direct crossover connections between computers forming the topology shown in Figure 4.18. Figure 4.19 shows example of TRA allocation and it has three tiers. Network experiments may use all network interfaces except for the control network to form connections between machines. All IP addresses on the testbed are statically assigned and are not changed. Node 1 is the only testbed machine which is accessible from outside; all other machines must be accessed through this machine. Set of bash shell scripts are written to easily copy files to all testbed machines or to run any command on all machines. The local testbed has been used to demonstrate the proposed architecture as well as to develop and debug various protocols and algorithms.

TRP code was run just above layer 2, bypassing all layers between layer 2 and the application layer. Thus TRP replaces both IP and its routing protocols. In the current study, transport layer was not included as the intent is to show the performance of the routing protocols in terms of convergence, control traffic, and packets loss during convergence. To run applications on TRP, SIPerf a modified clone of Iperf [65] which allows bandwidth and link quality measurement in terms of packet loss was used.

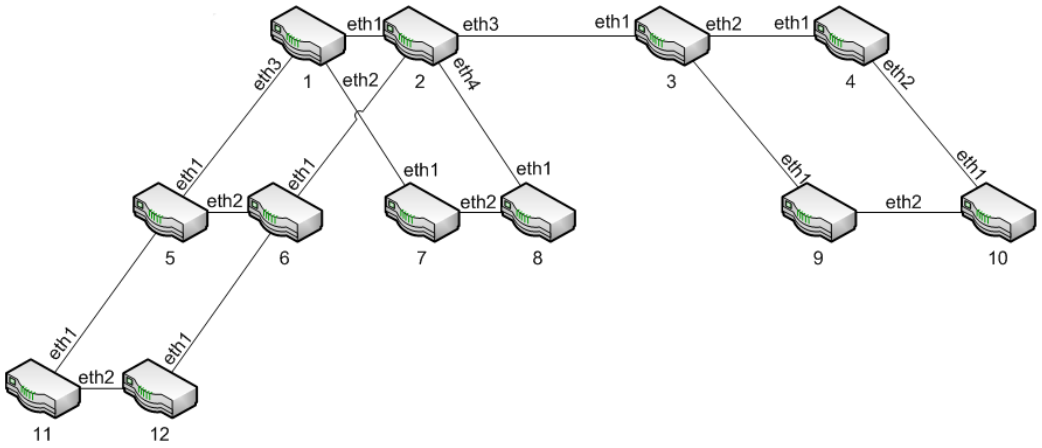


Figure 4.18: Local Testbed Topology

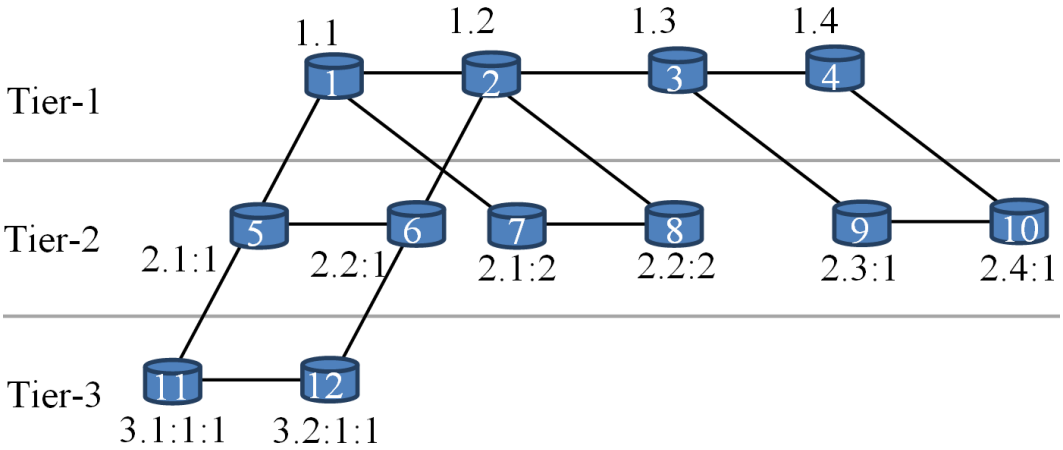


Figure 4.19: Local Testbed Topology with Example TRAs

TRP code is implemented as several functions, executed periodically or triggered on events. The code includes function for broadcasting of hello packets and tracking of missing hello packets which is activated for each active link and connection. TRP code has also a Routing function that implements the tiered based packet forwarding as a high priority process.

Chapter 5

Evaluation of TRA and TRP

To evaluate the FCT architecture, we first validated tiered stricture in an ISP and among ISPs discussed in Chapter 4.1.1 and 4.1.2 by using realistic datasets of Internet topologies. Then we first evaluate TRA with router-level topology of an ISP and AS-level topology of the Internet. Next, to validate the operation of TRP, a Linux-based FCT router is implemented. With the FCT router, the performance of the TRP is compared with both intra- and inter-domain routing protocols.

5.1 Evaluation of TRP in Intra-domain Routing

To evaluate the proposed tiered routing address scheme, we used the Rocketfuel data which has router-level topologies and maps routers to ASs, and analyzed AT&T router-level network topology.

5.1.1 Analyzing AT&T Network

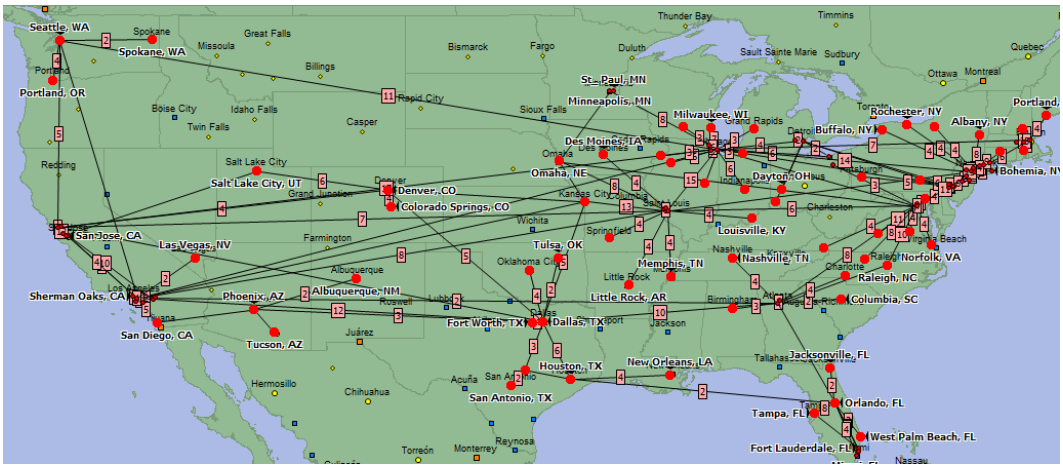


Figure 5.1: US AT&T network in OPNET imported from Rocketfuel data

With AT&T network data, we identified a total of 11,403 routers with 13,689 links interconnecting them (excluded Hawaii and Alaska). Figure 5.2 shows router-level topology of the entire AT&T network in the US. Figure 5.1 shows the topology imported into the OPNET network simulation tool [66] including the geographical locations of the different POPs (cities).

There were a total of 110 POPs in the AT&T network which can be seen as dots in the Figure 5.1. The numbers on the links represent the number of physical connections between POPs. Figure 5.3 and 5.4 are enlarged views of New York area and San Francisco area from the AT&T network. As seen in those figures, topology between cities (POPs) are look like hub and spoke topology, but inside of each city (POP) is highly connected like meshed topology.

Figure 5.5 shows node (router) degree distribution of 11,403 routers in AT&T network. Majority of routers have less than 3 links and average number of links is 2.423. Node degree can be one attribute to determine tiers in a network. For example, a node which has a large number of neighbors can be

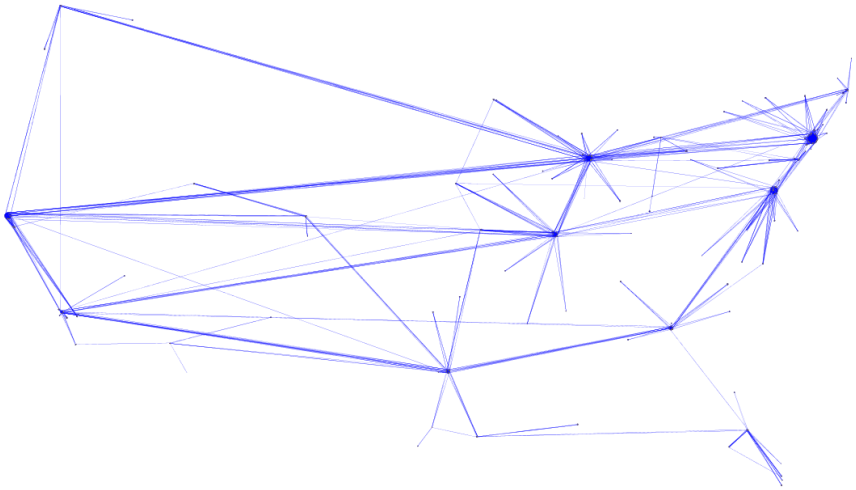


Figure 5.2: AT&T Router-Level Network Topology

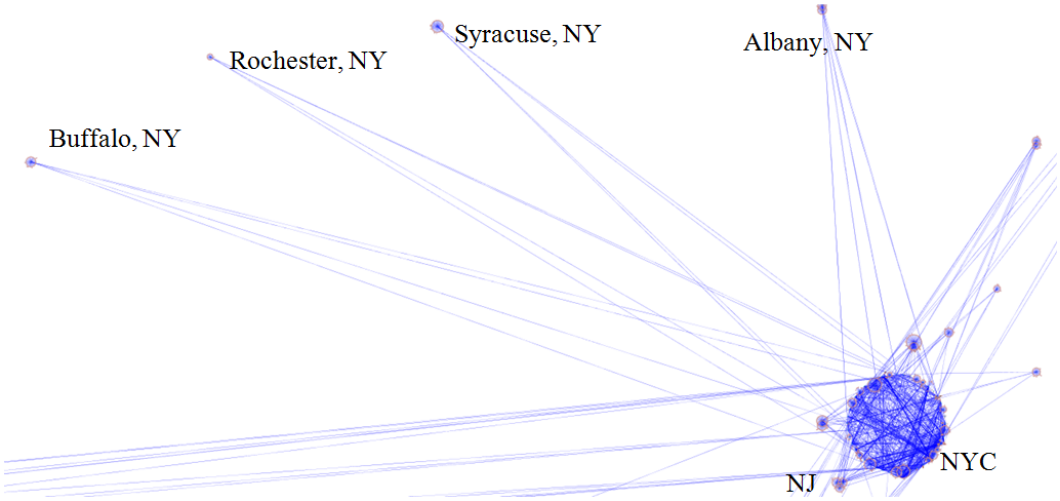


Figure 5.3: AT&T Router-Level Network Topology (NY area)

set as higher tiers because it can reduce number of tiers and hops. A shortest path length distribution of AT&T is shown in Figure 5.6. Total 137,933,376 possible shortest paths are found in the network and average of shortest path length is 7.792, and longest path length in the network is 18.

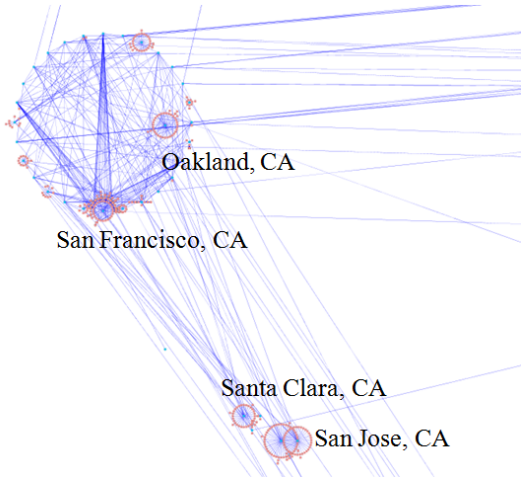


Figure 5.4: AT&T Router-Level Network Topology (SF area)

The next step was in identifying tiers and clouds inside every POP in the AT&T network. BB routers are assigned to a tier 1 cloud in the POP. The DR routers were designated to be at tier 2, and AR routers at tier 3. After assigning such categorization to all 11,403 routers, the TRA address was allocated to every router in entire AT&T network as explained in Chapter 4.2.

Figure 5.7 shows sorted distribution of routers in each POP and exact numbers are presented in Table 5.1. Only 10 % of AT&T POP has large number of routers and 90 % od them are less than 230 routers, which represents property of a hub-and-spoke topology. Those 10 % of large POPs can be recognized as a backbone of backbone in AT&T network and it can also be recognized as another tier in the network.

At tier 1 within a POP, all BB routers are assumed to belong to a single cloud, which means each POP has single tier 1 cloud. At tier 3, each AR router is recognized as a cloud because AR routers may be connected to other ASs (stub or otherwise) and networks. In this presentation we limit our con-

Table 5.1: Number of Routers at each POP in AT&T

	CITY (POP)	NUM		CITY (POP)	NUM		CITY (POP)	NUM
1	Chicago, IL	1010	41	Pittsburgh, PA	68	81	San Bernadino, CA	32
2	New York, NY	946	42	Harrisburg, PA	67	82	Des Moines, IA	31
3	Washington, DC	576	43	Wayne, PA	66	83	Dunwoody, GA	31
4	Atlanta, GA	499	44	Nashville, TN	66	84	San Antonio, TX	30
5	Dallas, TX	495	45	Hartford, CT	65	85	Ojus, FL	30
6	San Francisco, CA	485	46	Oklahoma City, OK	65	86	Bridgeport, CT	28
7	Seattle, WA	393	47	Rochelle Park, NJ	58	87	Portland, ME	27
8	Orlando, FL	368	48	Galva, IL	57	88	Fort Worth, TX	26
9	Cambridge, MA	368	49	Santa Clara, CA	57	89	Memphis, TN	26
10	Los Angeles, CA	337	50	Tampa, FL	56	90	Camden, NJ	26
11	Denver, CO	321	51	Omaha, NE	56	91	Madison, WI	25
12	St Louis, MO	226	52	Silver Springs, MD	55	92	Manchester, NH	25
13	Philadelphia, PA	205	53	Syracuse, NY	51	93	Rochester, NY	23
14	Phoenix, AZ	181	54	Cincinnati, OH	50	94	Norfolk, VA	23
15	Detroit, MI	178	55	Baltimore, MD	47	95	Dayton, OH	22
16	San Diego, CA	174	56	Birmingham, AL	44	96	Colorado Springs, CO	22
17	Houston, TX	159	57	Florissant, MO	44	97	Louisville, KY	19
18	Cleveland, OH	131	58	Tulsa, OK	43	98	Brookhaven, MI	18
19	Austin, TX	126	59	Spokane, WA	43	99	Freehold, NJ	16
20	New Brunswick, NJ	115	60	Richmond, VA	43	100	Akron, OH	16
21	White Plains, NY	107	61	Hamilton Square, NJ	42	101	Little Rock, AR	16
22	Salt Lake City, UT	106	62	Greensboro, NC	42	102	Madison Heights, VA	15
23	Anaheim, CA	100	63	Buffalo, NY	42	103	Worcester, MA	15
24	Arlington, VA	98	64	Plymouth, MI	40	104	Bridgeton, MO	14
25	San Jose, CA	94	65	Fort Lauderdale, FL	40	105	West Palm Beach, FL	10
26	Charlotte, NC	91	66	Oakland, CA	39	106	Abingdon, VA	5
27	Indianapolis, IN	85	67	Jacksonville, FL	39	107	Champaign, IL	2
28	Cedar Knolls, NJ	85	68	Providence, RI	39	108	Palo Alto, CA	1
29	Miami, FL	82	69	Albuquerque, NM	39	109	Newark, NJ	1
30	Riverside, CA	81	70	Columbia, SC	38	110	Tucson, AZ	1
31	Minneapolis, MN	81	71	Davenport, IA	38			
32	Milwaukee, WI	81	72	Stamford, CT	36			
33	Portland, OR	81	73	Oak Brook, IL	36			
34	Kansas City, MO	79	74	South Bend, IN	35			
35	Albany, NY	75	75	Bohemia, NY	35			
36	Framingham, MA	75	76	Grand Rapids, MI	34			
37	Raleigh, NC	72	77	Las Vegas, NV	34			
38	New Orleans, LA	71	78	Gardena, CA	33			
39	Sherman Oaks, CA	71	79	Springfield, MO	33			
40	Rolling Meadows, IL	71	80	St. Paul, MN	33			

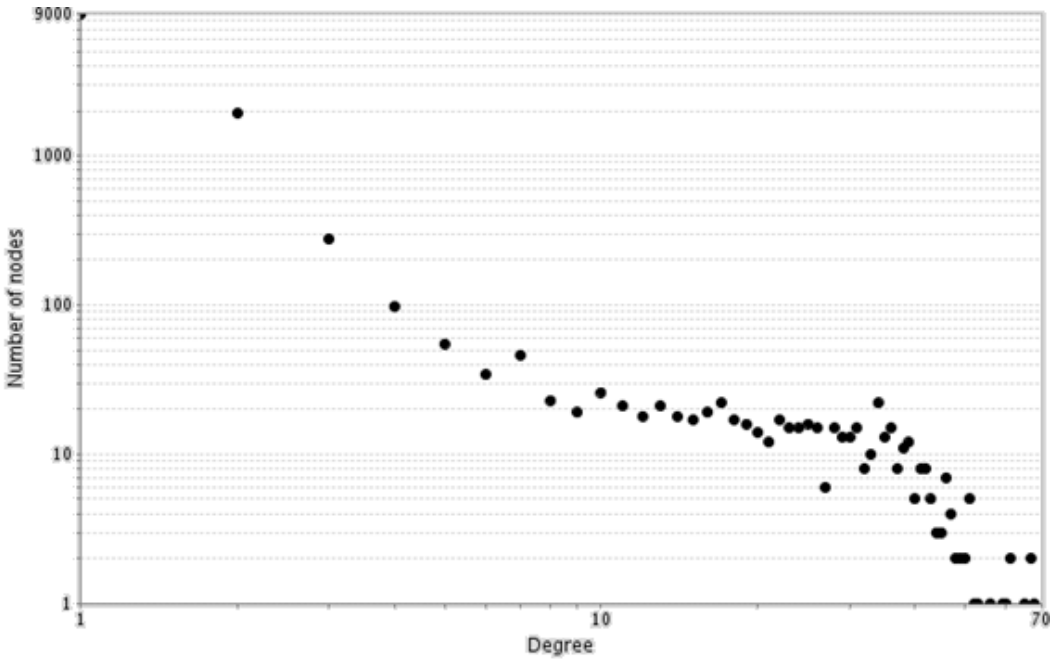


Figure 5.5: AT&T Router Degree Distribution

nectivity and address allocation study to the POP level within an ISP i.e. the AT&T ISP.

Figure 5.8 shows a distribution of BB routers in each POP. Correlation coefficient between POP size and number of router is 0.89, and relationship between number of BB routers in a POP and POP size is shown in Figure 5.9.

At tier 2, however, DR routers, which provide connectivity between BB and AR routers, should have redundancy and hence each set of DR routers is considered as a cloud. Since we did not have link weight information, the shortest path knowledge between BB and AR routers was used to identify a cloud of DR routers. Based on the shortest path between BB and AR routers, DR routers, which are on the shortest path to the same BB router, were assumed to belong to one cloud. For example, if DR router A and B are on the shortest path to BB router C, DR router A and B will belong to one DR

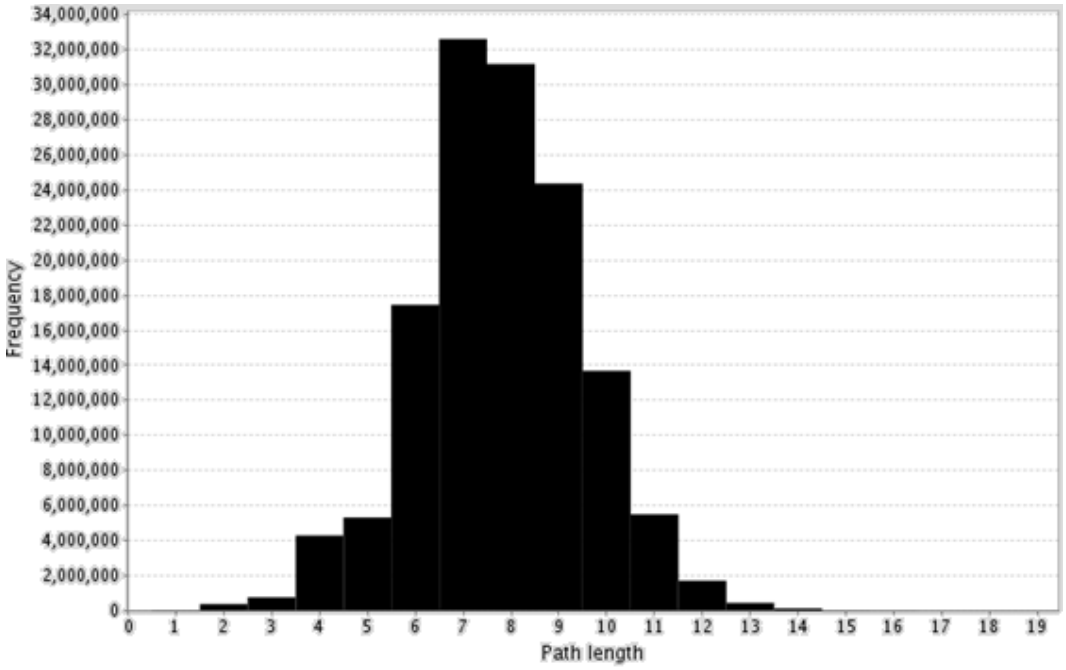


Figure 5.6: AT&T Router Shortest Path Length Distribution

cloud. If a DR router is on the path to different BB routers, then the DR router chooses the shorter hop to a BB router, and is considered to belong to the distribution cloud under that BB router.

5.1.2 Tiered Structure and TRA allocation

After identified tiers in the AT&T network, we allocated TRA address based on the tiers. Figure 5.10 shows original view of the Seattle POP of AT&T network. There are 393 routers and 437 links in the Seattle POP and each dot in Figure 5.10 represents a single router. Blue dot represents a backbone router. Figure 5.11 shows result of actual tiered address allocation to the Seattle POP in the AT&T network. In the Seattle POP, 6 BB routers were identified based on their connections to other POPs in the AT&T network and

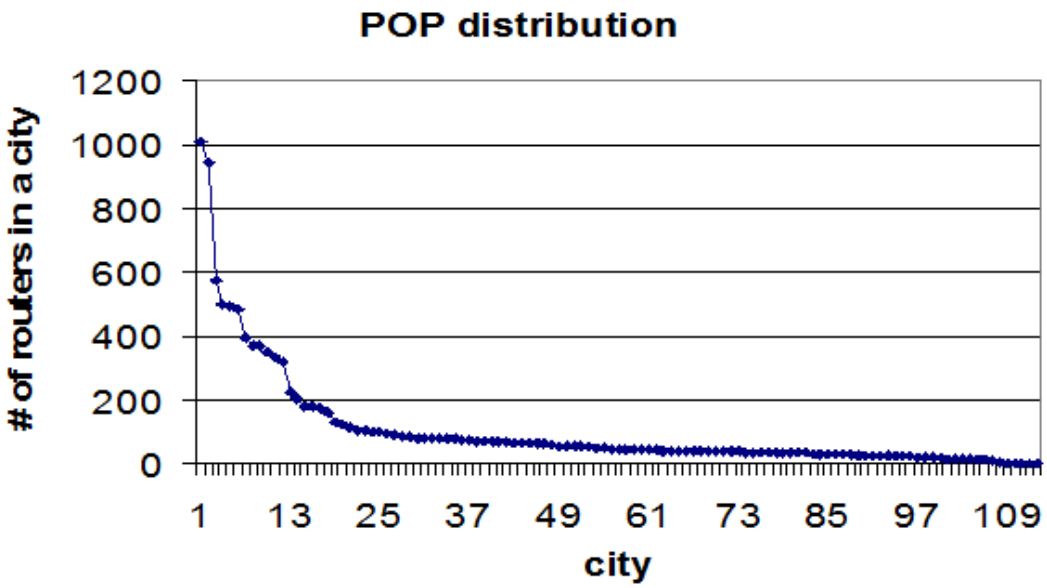


Figure 5.7: US AT&T POP Distribution

all BB routers thus belong to the cloud, which has TRA {1.7}. As per our study we used integers between 1 and 110 to uniquely identify each POP in the AT&T network and 7 is the Seattle POP ID assigned by us (one could use any other numbering strategy). At tier 2, there are 94 DR routers and 17 clouds as identified. Each block of dots (i.e. routers) at tier 2 in the figure represents a cloud. At tier 3, there are 293 AR routers and hence 293 clouds because each AR router is recognized as a cloud for reasons stated earlier, hence each dot is a cloud.

Table 5.2 shows interesting address statistics for the entire AT&T network using the tiered addressing scheme with only 3 tiers. There are a total of 110 POPs, 11,403 routers and 13,689 links in the AT&T network in the US (excluded Hawaii and Alaska). From this we identified 389 BB routers, 6,395 DR routers, and 4,619 AR routers. The Chicago POP is the largest POP based on the number of routers in the POP. The New York POP has the

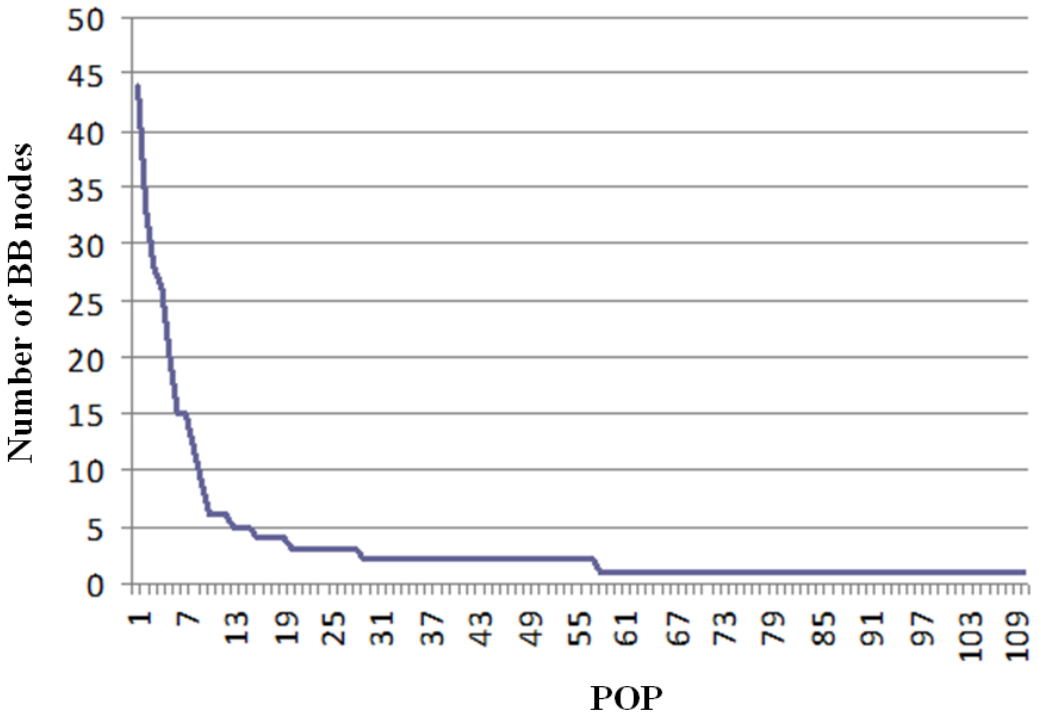


Figure 5.8: BB Routers Distribution of AT&T Network

largest BB cloud and the highest number of DR routers. The Dallas POP has the largest distribution cloud with 56 routers. The Seattle POP has the maximum number of AR routers which are connected to the same distribution cloud. These statistics are provided to show that they can be used to identify and optimize the proper size of a single cloud, help in nesting cloud decision.

5.1.3 Address Length and Numbers

We also compared TRA and IP address used in the US AT&T network. Based on a TRA address format mentioned in Figure 4.8, an address length of TRA

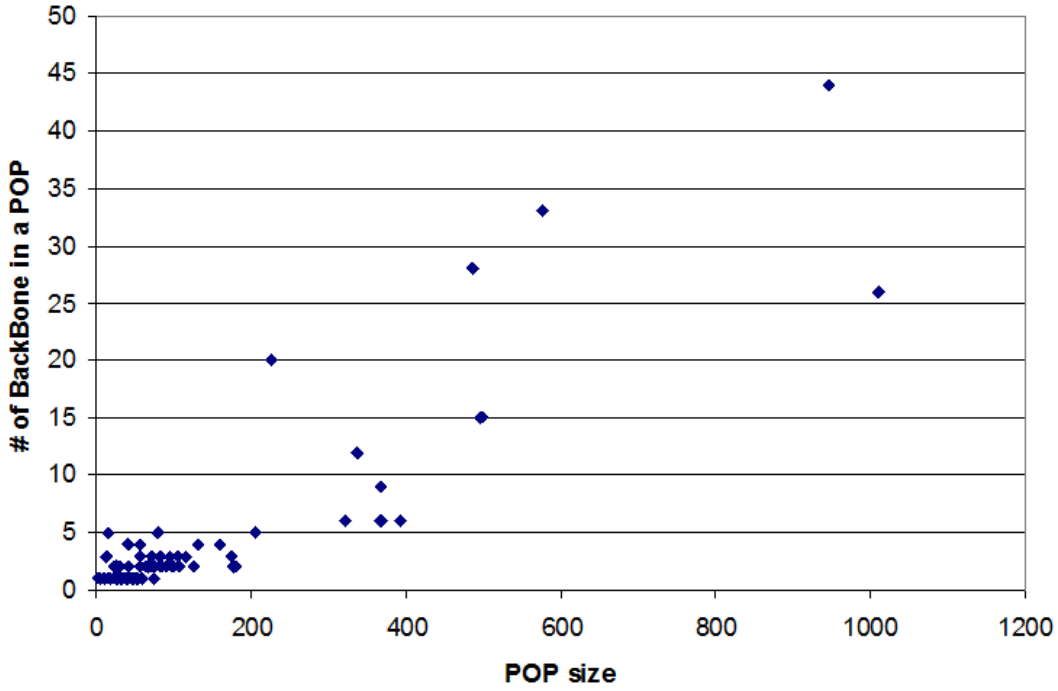


Figure 5.9: Correlation between BB routers and POP size

can be calculated by the following equation 5.1:

$$TRA_{len} = TV_{len} + LF_{end} + \sum_{i=1}^{TV} (LF_{len} + TA_{len}i) \tag{5.1}$$

$$TA_{len}i = \begin{cases} 4 & (1 < TA < 16) \\ 8 & (16 < TA < 256) \\ 12 & (256 < TA < 4096) \end{cases}$$

where TV is tier value of a TRA address, TA_{len} is a length of TA filed at the tier i level, TV_{len} is a length of a TV field, which is fixed 6 bits, and LF_{len} is a length of a LF filed, which is fixed 2 bits size.

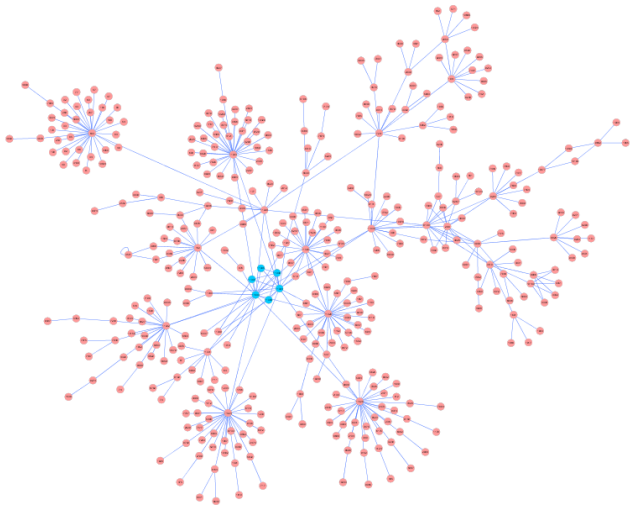


Figure 5.10: Seattle POP Topology of AT&T Network

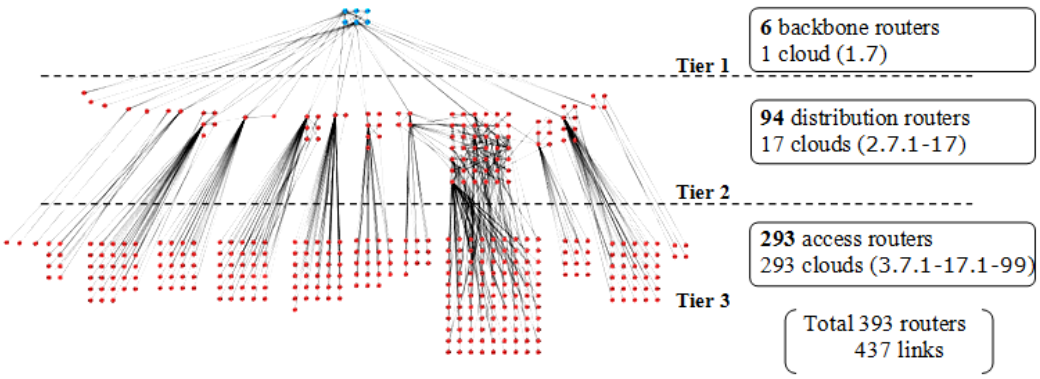


Figure 5.11: Seattle POP Topology of AT&T with FCT model

The pie chart in Figure 5.12 shows the length of the addresses that will be required if using the TRA addresses. Due to the flexibility in address sizes, less than 1 percent of addresses would exceed 32 bits, which is the length of IPv4 address, and 83.93% of addresses would be less than or equal to 28 bits. Moreover current IPv4 and IPv6 based routers requires a different address on each of its routing interfaces. In contrast the tiered address will use only one

Table 5.2: AT&T Network Statistics based on Tiered Routing Addresses

Total number of routers	11,403
Total number of links	13,689
Total number of POPs	110
Total number of BB routers	389
Total number of DR routers	6,395
Total number of AR routers	4,619
Maximum TreeAddress at tier 1	110 POPs
Maximum TreeAddress at tier 2	429 (New York)
Maximum TreeAddress at tier 3	99 (Seattle)
Maximum POP size	1,010 (Chicago POP)
Maximum BB cloud size	44 routers (New York POP)
Maximum DR routers in a POP	542 routers (New York POP)
Maximum distribution could size	56 routers (Dallas POP)

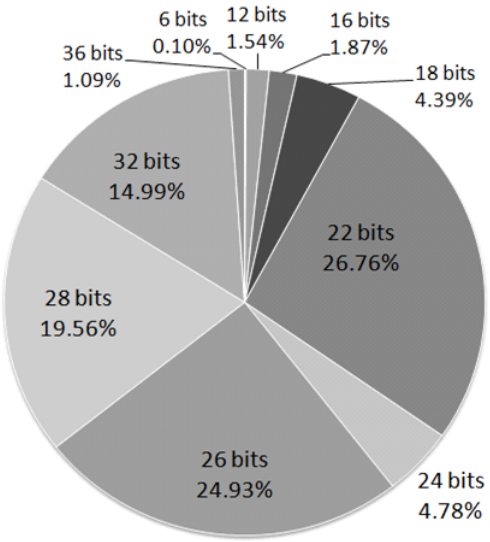


Figure 5.12: TRA Address Length distribution across AT&T network

address per router similar to Network Service Access Point (NSAP) addresses in Intermediate System to Intermediate System (ISIS).

The bar graph in Figure 5.13 estimates total bit sizes of TRA, IPv4, and IPv6 used in AT&T network. Address length of IPv4 and IPv6 are 32 bits and 128 bits per an address. Figure 5.14 shows the number of addresses required

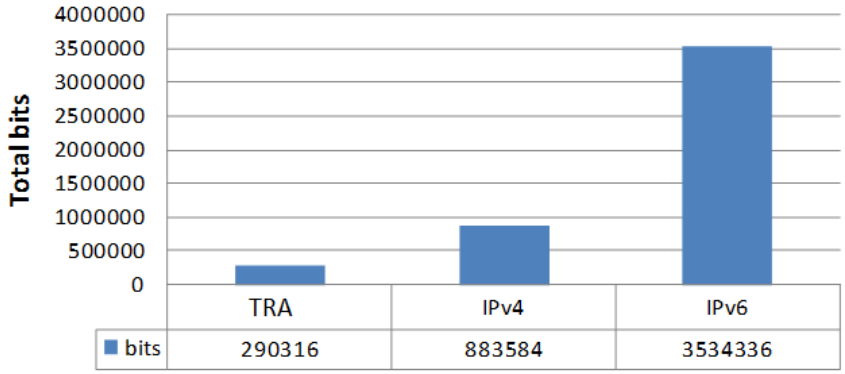


Figure 5.13: Total Size of TRA and IP Addresses

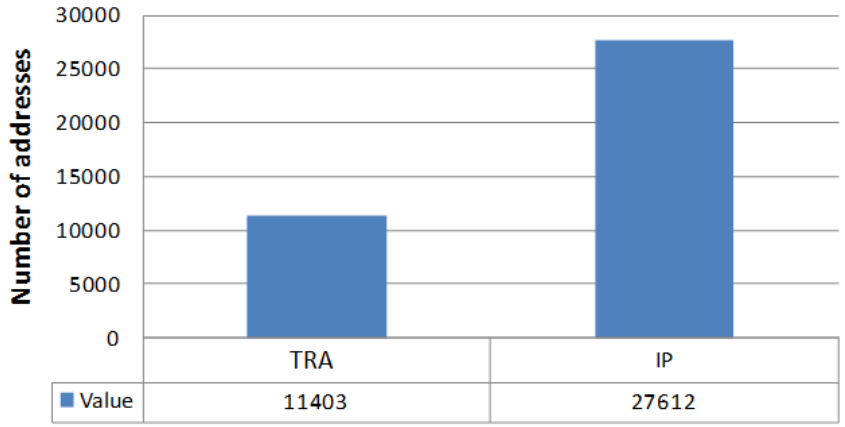


Figure 5.14: Total Number of Allocated TRA and IP Addresses

for all the routers in the AT&T using IP (v4 or v6) addresses and the tiered addresses. Both statistics shows that TRA address requires less number of address and size, which can also reduce traffics in the Internet.

5.1.4 HD Ratio for Address Allocation Efficiency

The tiered addressing scheme allows a maximum of 2^{12n} addresses at tier level n. The addresses can include ISP, AS cloud, network or device addresses within a network. The maximum address length of the entity in the network

can be calculated using Equation 5.2.

$$AL = 14n + 6 \tag{5.2}$$

where

- AL: Maximum address length
- n: Total number of the tiers in the network

In the current Internet, the efficiency of the IP address assignment was analyzed with the H_{ratio} as given by Equation 5.2 [67].

$$H_{ratio} = \frac{\log_{10}(N_{AO})}{N_{AB}} \tag{5.3}$$

where

- N_{AO}: Number of allocated objects
- N_{AB}: Number of available bits

However, since Equation 5.3 did not count the multiplicative affect of the loss of efficiency at each level of a hierarchical plan, we decided to use the Host Density ratio (HD_{ratio}), which is adopted to analyze IPv6 address allocation efficiency by IETF [67] and is given in Equation 5.4.

$$HD_{ratio} = \frac{\log_{x}(N_{AO})}{\log x(MAX\ N_{AO})} \tag{5.4}$$

where

- N_{AO}: Number of allocatable objects

- x: Any integer value bigger than 0

In [68], a HD_{ratio} of 0.94 is identified as the utilization threshold for IPv6 address space allocations. Equation 5.4 can this be rewritten to actually find out the N_{AO} as in Equation 5.5.

$$N_{AO} = (MAX\ N_{AO})^{HD_{ratio}} \tag{5.5}$$

According to Equation 5.5, IPv6 reaches HD_{ratio} of 0.9 when 1.65931E+36 addresses are allocated to the objects. At this, point new address space will be required for the new nodes.

Table 5.3: Number of Nodes in each Tier Level

Tier value	Max address length	Address capacity at the tier	Total capacity of network	Network capacity at HD:0.94
1	20	4096	4096	2486.671123
2	34	16777216	16781312	6184952.337
3	48	68719476736	68736258048	15379943237
4	62	2.81475E+14	2.81544E+14	3.82449E+13
5	76	1.15292E+18	1.15320E+18	9.51024E+16
6	90	4.72237E+21	4.72352E+21	2.36488E+20
7	104	1.93428E+25	1.93475E+25	5.88069E+23
8	118	7.92282E+28	7.92475E+28	1.46233E+27
9	132	3.24519E+32	3.24598E+32	3.63634E+30
10	146	1.32923E+36	1.32955E+36	9.04239E+33
11	160	5.44452E+39	5.44585E+39	2.24854E+37
12	174	2.23007E+43	2.23062E+43	5.59139E+40
13	188	9.13439E+46	9.13662E+46	1.39040E+44

In Table 5.3, for the tiered address scheme, the maximum number of entities, such as ISPs, POPs, networks or devices that can be accommodated, and in turn the available address space at a given tier is given in column 3. The first column gives the TierValue. The second column gives the maximum address length as calculated using Equation 5.2 at any given tier assuming maximum

address fields of 12 bits each. The total number of supported addresses, for a given TierValue including all of the addresses within the tiered hierarchy is given in the fourth column. Let us explain this with an example: at tier 2 we have a maximum address space given by 16,777,216 (=2^{12n}, where $n = 2$). However there are addresses supported in tier 1 under which we have tier 2. So the total number of addresses that can be supported in a system that has 2 tier levels will be given by 16,781,312, which is 4096 (at tier 1) + 16,777,216 at tier 2. So the values in column 4 are a cumulative count of addresses from all tiers above a given tier, including that tier. The total number of addresses that can be supported by the network till it reaches the HD ratio of 0.94 was calculated using Equation 5.5 and is given in the last column.

As it is seen in Table 5.3, the tiered routing addresses reach the IPv6 address allocation threshold capacity at tier 11 with 160 bits of address length at most. However, the threshold in the tiered address is not fixed as for IPv6; it is flexible and can be extended as needed by increasing the tier value. The only restricting factor could be the address length. As explained under the packet forwarding section and along with the nested concepts, the maximum address length that any router has to deal with is determined by the first address field in the tiered addresses, which knows how to direct or forward a packet.

Another major concern that can arise if the address length increases, is use of tiered addresses in wireless networks which are bandwidth constrained. However, in such case only the nested address used within the wireless network will be used for forwarding within that network. With a tier 3 address, this would be 48 bits maximum. At this point it has to be further noted that the use

of the tiered address would preclude MAC addresses and that all forwarding whether inter or intra-cloud can be supported by the tiered address.

In this TRA validation, the main goal was to support for future growth in an unrestricted manner, whether it is in terms of address space or networks. We highlighted the efficient use of address space with the tiered address scheme. We provided some operational aspects of the internetworking model to explain the application of the tiered addresses. We also illustrated tier based address aggregation with examples and applied the same to the AT&T network in the US. Using this application and the HD_{ratio} we then analyzed some performance characteristics of the tiered addressing scheme.

5.1.5 Routing Table Size Analysis of TRP

We now provide an example of applying the FCT model to a small network of 6 routers with 9 network segments. We provide the routing tables, when the network uses IP addresses and runs an IP routing protocol, and compare with the routing table sizes that can be expected if the network were running TRP.

Table 5.4: IP Routing Table of Router B in Figure 5.15

Destination Network	**Route Via**
10.1.1.0	connected
10.1.2.0	connected
10.1.3.0	10.1.4.0, 10.1.1.0
10.1.4.0	connected
10.1.5.0	10.1.2.0, 10.1.4.0
10.1.6.0	10.1.4.0
10.1.7.0	10.1.4.0
10.1.8.0	10.1.2.0
10.1.9.0	10.1.2.0, 10.1.4.0

In Figure 5.15, a stub network with 9 sub-networks is shown. We associate the sub-networks with IP addresses from 10.1.0.0/16 IP address space. The

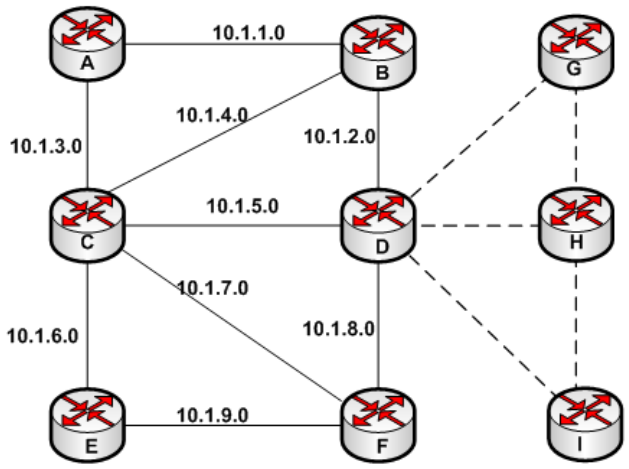


Figure 5.15: A Sample IP network Topology with 9 Subnets

routing table at any one of the routers will be similar to that shown for router B in Table 5.4. This table has 9 entries for the 9 segments, assuming that the tables are fully populated with all network segment addresses without depending on default forwarding addresses.

Without loss of generality, the routing tables for the above network was populated using Routing Information Protocol (RIP). Running OSPF or BGP, (as would be normally expected for inter-domain routing), would have resulted in routing tables of similar size. For the network as shown in Figure 5.15, and with routers having a minimal number of interfaces, requires a routing table with 9 entries.

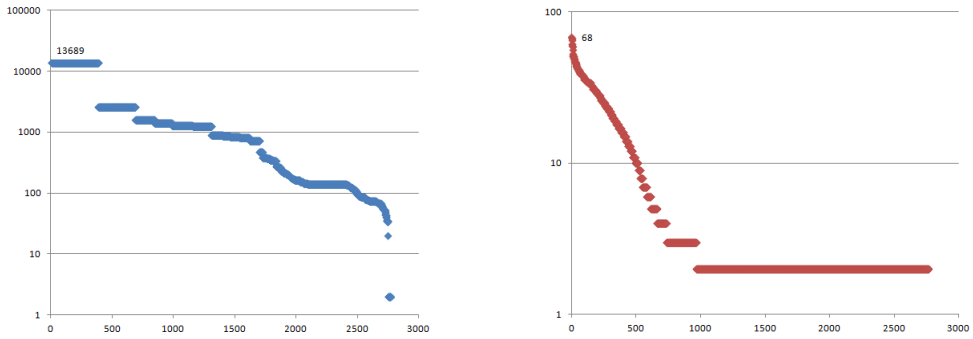
Let us now apply the FCT model to this network and investigate the routing entries required at a router for packet forwarding. Let us assume that routers A and B belong to two distinct backbone clouds, routers C and D belong to two distribution clouds and routers E and F to the access clouds connecting to access networks. The routing table at router B will have 3 entries each for the neighbors that it is connected to, while router A will have

only two entries. Routers E and F will have two and three entries respectively. That is irrespective of the number of segments, the routing entries in a router under the FCT model depends only on the number of directly connected neighbors or segments.

- Impact of Network Size: Let us now assume that to router D, we add 3 more routers (Routers G, H, and I) and 5 segments as shown by the dotted lines in Figure 5.15. The routing table sizes at every router will now increase to 14 entries. However, with TRP, the routing table sizes at routers A, B, C, E and F will remain the same. The routing table only at router D will increase from 3 to 6. The tiered based routing thus introduces independency of the routing table size from the network size.

IP addresses, their allocation and the meshed topology have all contributed to the complex growth of routing tables, which adversely impacts the scalability as the Internet grows in size. Route discovery process in the current Internet is essential to establish communication links and maintain information flow between millions of devices and networks. Routing problems are faced both in intra-domain and inter-domain routing. While intra-domain routing protocols like RIP and Open Shortest Path First (OSPF) continue to address loop avoidance and strive for faster convergence, concerns over inter-domain routing on the other hand are very high as the Border Gateway Protocol (BGP) routing table sizes escalate steadily. BGP routing table size at the core routers today has exceeded 490,000 entries [40]. This high load in the core routers is indicative of an imbalance in the routing information handling, which adversely impacts the advantages of the meshed structure, by making the routers a potential bottleneck.

Figure 5.16: Routing Table Size of OSPF and TRP in AT&T



OSPF TRP

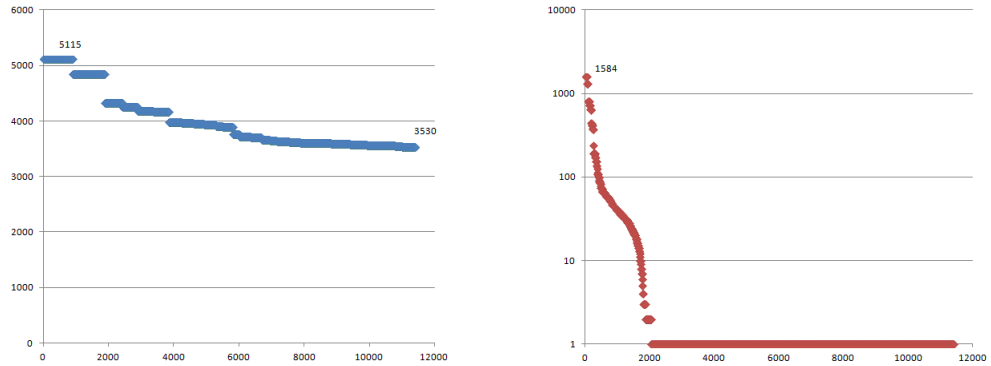
With the AT&T router-level topology, routing table size of OSPF and TRP are estimated. Figure 5.16 shows routing table sizes of OSPF and TRP in AT&T network. In OSPF, all backbone routers and links are recognised as an Area 0 in the OSPF routing domain, and routers and links in a POP recognized as an Area under the Area 0 domain. The largest routing table size in OSPF is 13,689 and around 300 routers have that size of routing table. On the other hand, the largest routing table size in TRP is only 68, and majority of routers in TRP are 2 because size of routing table in TRP is based on number of direct link at a router. Average routing table sizes of OSPF and TRP are 1,161.96 and 2.42 respectively. At any router in AT&T network, routing table size of TRP is significantly smaller than OSPF router.

5.1.6 Overhead Analysis of TRP

We also conducted simulation to count number of control packet to update a routing table after 1 link/node failure in AT&T network. Figure 5.17 shows distribution of the update packet at each router where the failure detected. In the case of OSPF, the maximum number of the update packets generated by

single failure is 5,115 and the minimum number is 3,530. On the other hand, the maximum number of the update packets in TRP is 1,584 and the minimum number is just 1. Average number of update packets of OSPF and TRP are 4,003.66 and 25.64 respectively. Number of updates packets are significantly small at an failure occurred any router location in AT&T network.

Figure 5.17: Number of Updates of OSPF and TRP in AT&T



OSPF TRP

5.1.7 Performance Statistics and Analysis of TRP on Testbed

In addition to the development of the local site testbed, we also conducted large scale studies to evaluate our research concepts on test facilities provided by Emulab [69]. After the TRP code was tested and evaluated over the local testbed it was then ported to the Emulab facilities of GENI. In this section, we show performance comparison with OSPF routing protocol to evaluate the TRP as an IGP.

Emulab is an experimentation facility which allows creation of networks with different topologies to provide a fully controllable and repeatable ex-

Table 5.5: Emulab Testbed Configurations

Topology	21 Nodes	45 Nodes
Type of processor	Pentium III	Quad Core Xeon Processor
Number of links	24	54
Link shaping nodes	12	20
Connection speed	100 Mbps	100 Mbps

perimental environment. Emulab uses different types of equipment for this purpose. D710 which is a 64 Bit Intel quad core Xeon based machine was used in a 45 nodes topology and PC850 type which is a Pentium 3 based machine was used in the 21 nodes topology; both topologies are shown in Figure 5.18. Different machines were used for the two topologies due to the allocation process at Emulab and systems availability at the time of request. Note that number of network interface of Emulab PC is limited to five where one is used for control, thus only four network interfaces are available (without virtual interface). The type of network topology is, thus limited by these constraints.

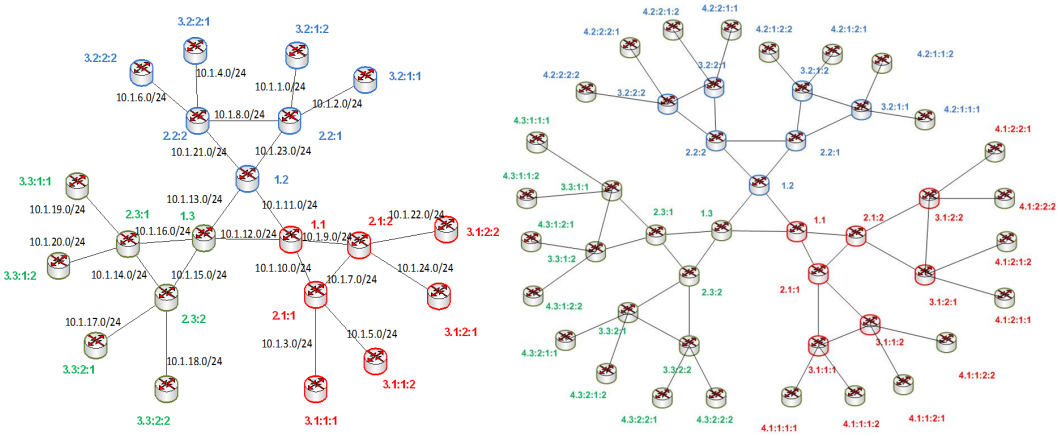


Figure 5.18: Testbed Topology with Tiered Routing Addresses

For the 21 nodes topology of Figure 5.18(a), the configuration details are provided in Table 5.5 along with the 45 nodes topology. In the 45 nodes topol-

ogy, the additional 24 nodes were added to the outer circle of routers utilizing a topological connection similar to that of the outer routers in Figure 5.18(b). In Figure 5.18(a), the IP addresses were allocated from address space 10.1.x.x /24 to the segments as shown. The TRA addresses to run TRP were allocated using the scheme described in Chapter 4.2. Link shaping nodes were used to emulate link failures. Emulab provides the use of such link shaping nodes that can be placed on the segments for this purpose. Emulab provided software commands were used to disable these nodes and thus emulate link failures.

Due to the limited numbers of machines, the limited durations of availability for use, and to provide a random environment for the test, which replicated topology using identical devices, they were conducted in two different sets of networks and the experiments were repeated five times in each case. The results were collected for convergence time of both OSPF and TRP in each topology. To collect convergence time and data on packet losses, each test topology required five hours and two hours of run time respectively.

Quagga 0.99.17 [70], a software routing suite for configuring OSPF was used for the comparison studies For the OSPF evaluations, only one area was defined, as the intention is to demonstrate the performance impacts to increase in the number of routers in a network or in an area. We also note that The TRP code which is in its research and development stage operates on the Linux kernel user space and hence the timings and dependent variables such as packet loss during convergence would project a higher value than if the code were run in kernel space. Comparatively the Quagga OSPF code, which was verified to run on Linux systems, runs in the kernel space.

Link Failure Detection Time

This is the same for OSPF and TRP as they detect a link failure on missing four hello messages. With a hello interval of 10 seconds, this was recorded to be 30 seconds with an additive time, which is the time between the first missing packet and the time when the link was actually brought down. This would be the same for OSPF and TRP due to the failure detection mechanism. However, TRP can fall back on an alternate path on missing a single hello packet, without affecting the forwarding operation. This was not implemented in the current version.

Time to Update the Routing Tables

This time is different for TRP and OSPF. The differences are explained below with the aid of Figures 5.19 and 5.20.

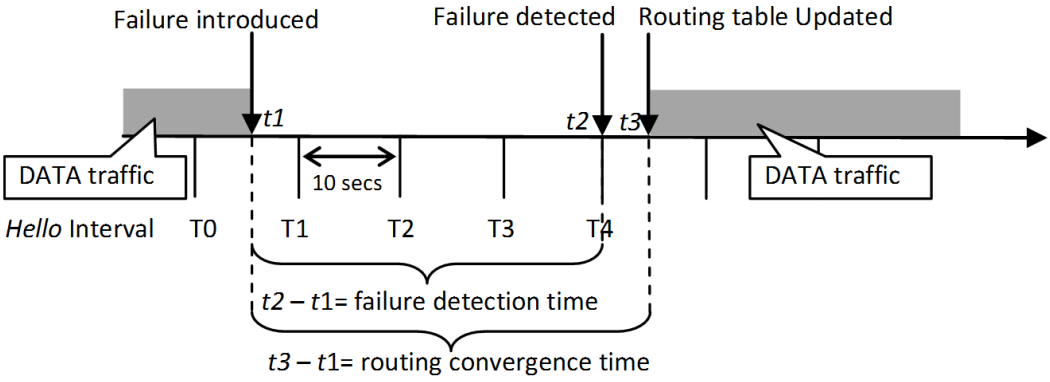


Figure 5.19: TRP Routing Convergence Time

- TRP Response to Link Failures: In Figure 5.19, the time t_1 when the link failed is noted along with time t_3, which is the time it took to remove the link from the routing table. Total time for convergence T_c is given

by

$$T_c = T_{ru} - T_{fd} \tag{5.6}$$

where T_{fd} is the failure detection time given by

$$T_{fd} = t_2 = t_1 \tag{5.7}$$

and T_{ru} is the routing table update time given by

$$T_{ru} = t_3 - t_2 \tag{5.8}$$

Thus,

$$T_c = t_3 - t_1 \tag{5.9}$$

T_{fd} will be the same for OSPF, but T_{ru} is negligible in the case of TRP as this is the time for the TRP code to access the routing tables and update its contents. In Figures 5.19 and 5.20, these times are identified based on the operations of TRP and OSPF respectively.

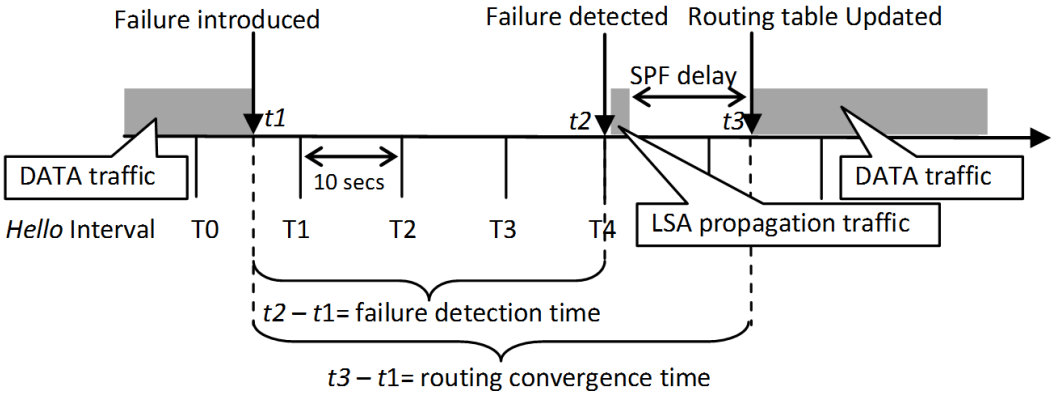


Figure 5.20: OSPF Routing Convergence Time

- OSPF Response to Link Failure: OSPF uses several timers on link failures, to rerun SPF algorithm and a few other hold times to avoid toggling. They are *hold_time*, which is the seperation time in milliseconds between consecutive SPF calculations. An *initial_hold_time* and *max_hold_time* are also specified. SPF starts with the *initial_hold_time*. If a new event occurs within the *hold_time* of any previous SPF calculation then the new SPF calculation is increased by *initial_hold_time* up to a maximum of *max_hold_time*.

  Let T_{LSA} be the LSA propagation delay, T_{SPF} be the time to run SPF on subsequent LSA messages and T_{TU} be the table update delay, then T_{ru} of OSPF is given by

$$T_{ru} = t_{LSA} + T_{SPF} + T_{TU} \tag{5.10}$$

  T_{SPF}, *initial_hold_time* and *max_hold_time* were set to 200 ms, 400 ms, and 5000 ms respectively for the test. Figure 5.20 captures the relationship between the delays for OSPF.

Performance Statistics and Analysis

The performance of OSPF and TRP, during the initial convergence phase and their response to subsequent link failures are presented in this section. The Network Time Protocol (NTP) is used to start OSPF and TRP protocols at the same time in all routers after the networks had stabilized. In the histograms, data collected for the two test sites are provided separately, to show the closeness of the two data sets collected under different environments which reflects the reliability of the experiments conducted.

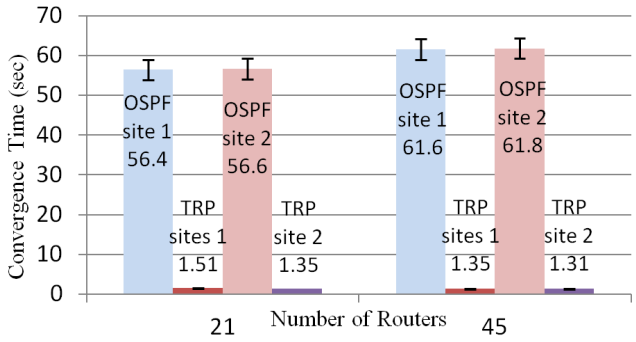


Figure 5.21: TRP vs. OSPF Initial Convergence Time (sec)

- Initial Convergence Times: Figure 5.21 records the average initial convergence times in seconds as collected from the two test sites and for the two different topologies, one with 21 routers and the other with 45 routers. While the convergence times recorded for OSPF range from 55 seconds in the case of the 21 router network to over 60 seconds in the case of the 45 router network, the convergence times for the network running TRP is around 1 second. While the convergence times are stable irrespective of the number of routers in networks running TRP, in the case of OSPF, the convergence times showed an increase by 5 to 6 seconds, indicating dependency of the convergence times to the network diameter. Thus, TRP under the FCT model offers an improvement in magnitude of 50-60 times as compared to OSPF.

- Control Overhead During Initial Convergence: Figure 5.22 shows the plot of the control overhead in Kbytes for OSPF and TRP. The control overhead in the case of OSPF varies from 250 Kbytes for the 21 router network to around 750 to 800 Kbytes for the 45 router network. The increase in overhead almost triples when the network size doubles. The

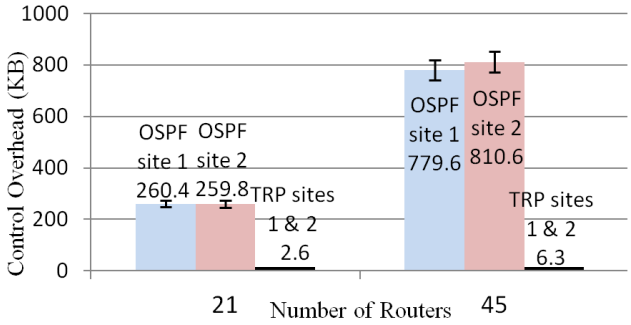


Figure 5.22: TRP vs. OSPF Routing Control Overhead Size (KB)

control overhead for TRP was 2.6 Kbytes for the 21 router network and around 6 Kbytes for the 45 router network. The improvement achieved with TRP in magnitude is 100 times in the case of the 21 router network and 130 times in the case of the 45 router network.

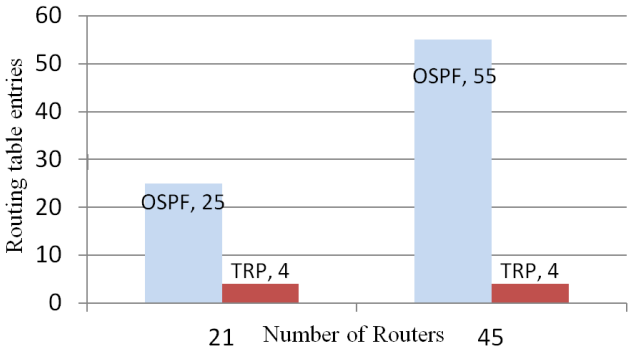


Figure 5.23: TRP vs. OSPF Routing Table Entry Size

- Routing Table Size: In Figure 5.23, the routing table sizes collected for the two test sites are the same in both case of TRP and OSPF, and hence the two site graphs have been merged into one. The figure records the maximum of the routing table entries noted in the routers. In the case of OSPF this value is 25 for the 21 router network (as there are

25 segments) and in the case of the 45 router network this value was recorded as 55. In the case of TRP, the routing table entry reflects the number of directly connected neighbors, so in both cases, i.e. the 21 router and 45 router networks the maximum routing table entry was 4. TRP routing table sizes do not depend on the network size, which provides proof to the scalability of TRP.

To summarize, Figures 5.21 and 5.22 provided the performance statistics relating to the initial convergence of the two routing protocols. These metrics were recorded as they reflect the difference of the underlying techniques in the two protocols. Their impact on routing performance during normal network operation especially when there are link or node failures is presented next.

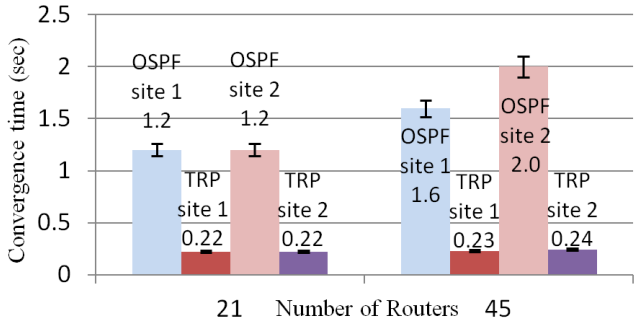


Figure 5.24: TRP vs. OSPF Convergence Time after Failure (sec)

- Convergence Time after Link Failure: Figure 5.24 is the routing table update time in seconds subsequent to a link failure detection. While OSPF shows an update time of 1.5 to 2 seconds for the 45 router network and just over a second for the 21 router network, TRP update times were as low as 200 ms to 240 ms; a magnitude of 6 improvement for the smaller network and a magnitude of 8 improvement for the larger network. The

routing table update time is invariant to the network size in the case of TRP and the increased improvement over OSPF as the number of routers increase is to be noted.

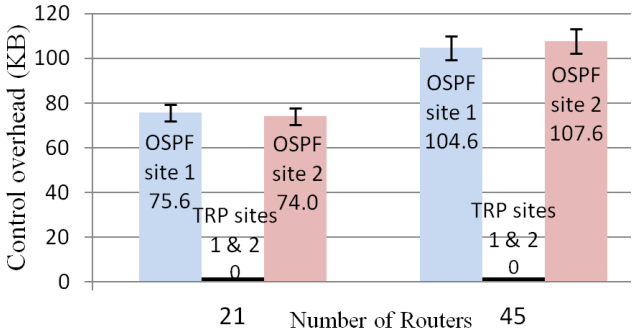


Figure 5.25: TRP vs. OSPF Control Packet Size after Failure (KB)

- Control Overhead after Link Failure: Figure 5.25 is the plot of the control overhead for TRP and OSPF collected during the convergence times after link failure, which includes the time to detect a failure and also the time to update the routing tables. For the given topologies no control overhead is incurred with TRP, i.e. there is no necessity to propagate any change messages to the network. OSPF required around 100 Kbytes and 70 Kbytes of control packets for the 45 router and 21 router networks respectively. Though for the given topology TRP does not incur any control overhead, for complex topologies, in TRP the change information may have to be propagated to networks downstream to block usage of the invalid address when a link goes down. Similarly upstream router may also have to be informed when a downstream link fails. These features are planned for testing using simulations as with the limited

interfaces supported per Emulab equipment it was not possible to have such configurations.

- Data Packets Lost: Since link failure detection mechanism of missing four hello messages is the same for both OSPF and TRP, the packets lost during failure detection is the same for both protocols and hence is not presented. The time to update the routing tables is recorded to be around 0.2 seconds for TRP and 1.2 seconds to 2.0 seconds for OSPF. Thus the packets lost during routing table update time was a maximum of 1 packet for TRP and a maximum of 10 packets with OSPF at a data rate of 5 packets per second generated by SIperf and Iperf respectively.

The TRA address in the FCT architecture is used by TRP for packet forwarding. Initial convergence time and convergence time after failure are significantly low because TRP does not require message flooding to all nodes in the network or any significant recalculations and re-computing on topology change. As a result of no message flooding, control overheads are also very low. Entries in routing tables used in TRP are satisfied by addresses of only the direct neighbors because of the inherent routing information in the TRA addresses. Thus, the routing table sizes in TRP are significantly small. From the results presented thus far it would be clear that TRP would be an ideal routing protocol to address scalability concerns as networks grow in number and in size. This is true as the routing table sizes and routing table update time is independent of network size. This in turn will positively impact the routing performance in the network.

5.2 Evaluation of TRP in Inter-domain Routing

In this section, we show performance comparison with BGP routing protocol to evaluate the TRP as an EGP. Inter-domain routing protocols due to their inherent nature of operation result in high churn rates leading to instability of routing information [71]. Churn rate is defined as the total number of routing updates generated by an event in the Internet. Currently it has been reported that 80% of events in BGP were globally visible [20]. These conditions are not conducive to a healthy and sustained growth of the Internet. To evaluate the TRP, we first analyzed the churn rate of TRP by using worldwide AS topology from the CAIDA dataset.

5.2.1 Analyzing Worldwide AS Network

In Chapter 4.1.2, the CAIDA dataset is used to identify the tiered structure among ISPs. A total 33,508 ASs and 75,002 AS links associated with provider-customer and peer-peer, and sibling relationships are recognized. There are 69,192 provider-customer links, 5,591 peer-peer links, and 219 sibling links. Figure 5.26 shows one example view of entire AS-level topology. As seen in the edge side of the topology in Figure 5.26, AS topology is tree-like topology because most of links are associated with provider-customer relationship and less number of peer-peer and sibling relationship. Furthermore, an AS tends to have more customer ASs than provider ASs because of their business model.

Based on a methodology explained in Chapter 4.1.2, total 7 tiers are identified in the worldwide AS topology. Figure 5.27 shows number of ASs at

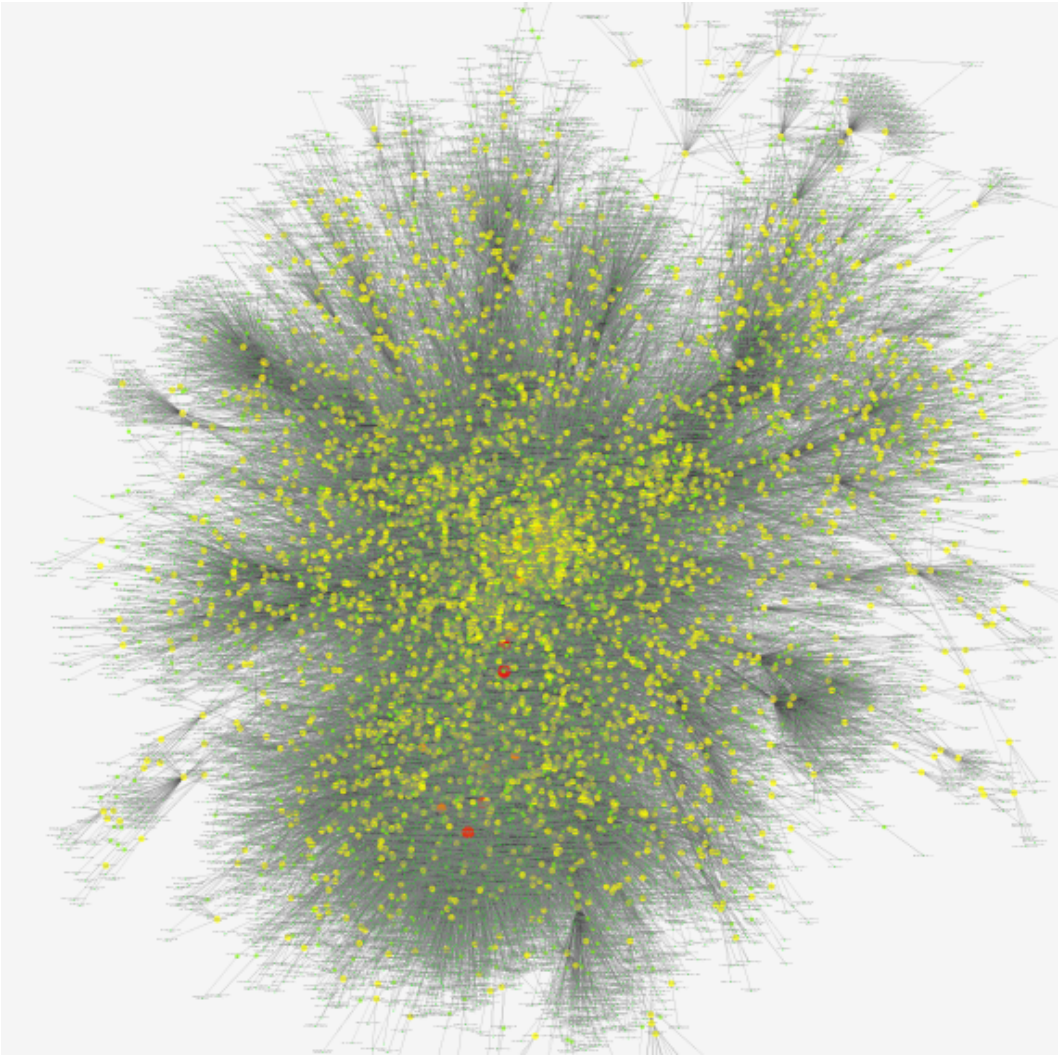


Figure 5.26: Visualized Worldwide AS Topology

each tier level. There are 121 tier 1 ASs, 11,199 tier 2 ASs, 16,635 tier 3 ASs, 5,005 tier 4 ASs, 524 tier 5 ASs, 23 tier 6 ASs, and 1 tier 7 AS, and Figure 4.5 shows types of AS at each tier level. An AS which has customer ASs called Provider AS and an AS does not have any customer AS is called Stub AS, which is edge of the AS topology.

Only .16 % of ASs are tier 1 AS and TRA address allocation for all ASs can start from these small number of tier 1 ASs. Unlike IPv4 and IPv6,

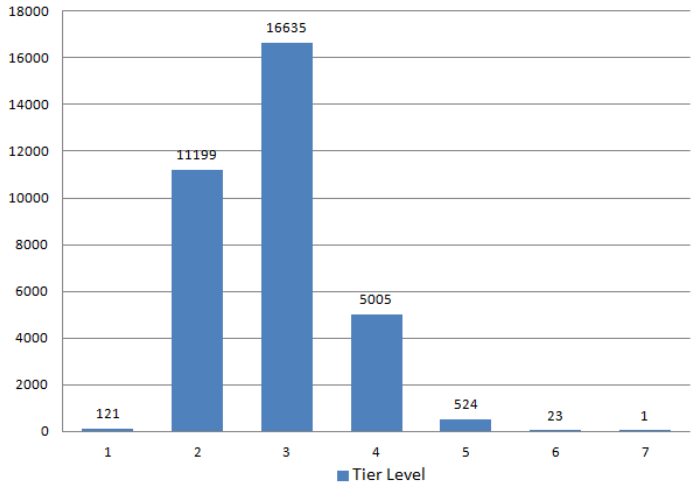


Figure 5.27: Identified Worldwide AS Tiers

central organizations like a reginal Internet registries (RIRs) are not needed. Figure 5.28 shows a partial topology of tier 1 ASs and Table 5.6 lists several tier 1 provider AS information. In Figure 5.28, ASs located center circle are well connected and having meshed topology because they are backbone of the Internet, and connected with peer links. Most of ASs located around the meshed topology are sibling ASs that have different ASN but the same organization. In Table 5.6, column ASN is AS number, and Customer, Peer, and Sibling are number of links.

5.2.2 AS Tiers and TRA Allocation

We allocated TRA addresses to all ASs. Simulation tool was implemented to allocate the TRA addresses to the different AS identified at the different tiers using the FCT model. Without loss of generality, in the case of a multi-homed AS, the address for customer AS was derived from the service provider with the lowest tier value. Figure5.29 shows example of TRA allocation at each

Table 5.6: List of Tier 1 Provider ASs

ASN	Name	Customer	Peer	Sibling
3356	Level 3 Communications	2611	19	1
174	COGENT Cogent/PSI	2480	20	2
7018	ATT-INTERNET4-WorldNet Services	2265	16	7
701	UUNET - MCI Verizon Business	2052	12	9
6939	HURRICANE - Hurricane Electric	1445	62	0
1239	SPRINTLINK - Sprint	1356	21	6
209	Qwest Communications Company	1355	30	2
3549	GBLX Global Crossing Ltd.	1332	26	0
4323	TWTC - tw telecom holdings, inc.	1255	45	0
10429	Telefonica Empresas SA	85	0	2
6298	Cox Communications	38	0	1
38022	REANNZ National Research Network	34	0	0
22927	Telefonica de Argentina	34	0	1
306	DoD Network Information Center	32	0	1
5006	ZAYOMN1 - Onvoy	28	0	2
20231	Road Runner HoldCo	23	0	1
7011	Frontier Communications	19	0	1
276	University of Texas System	16	0	3
7017	Road Runner HoldCo	16	0	1
10834	Telefonica de Argentina	16	0	2
30696	Texas Education Agency	10	0	0
100	FMC Central Engineering Lab	9	0	0
16905	NTG - North Texas GigaPOP	9	0	0
17373	MCI-COV - MCI WorldCom	6	0	1
22318	Cox Communications Inc.	4	0	1
27651	ENTEL CHILE S.A.	3	0	1
270	NASA	2	0	1
7845	ntegra Telecom	1	0	1
13398	K12LINK - WHRO	1	0	1
292	ESNET-WEST - ESnet	1	0	1
7315	Colombia Telecommunications	1	0	1

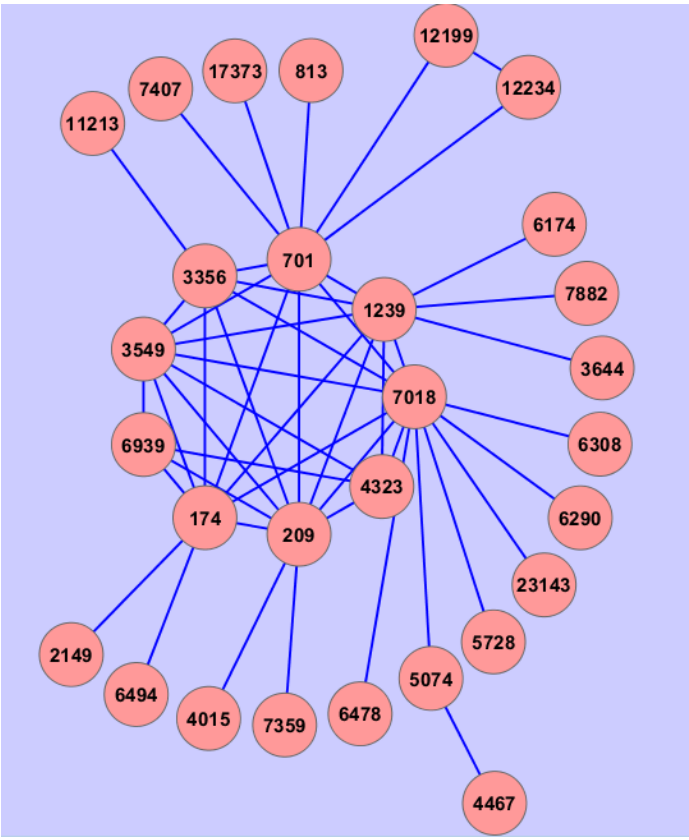


Figure 5.28: Topology of Tier 1 ASs

tier level. Instead of using any integer value for the tier 1 AS, the AS Number (ASN) which is allocated uniquely by Regional Internet Registry (RIR), was used. Thus the ASN of these ISP shows up in the TRA address at tier 1.

After TRAs were assigned, we also identified AS relationship trees which can imply churn rate of TRP. The concept of the TRA address tree is shown in Figure5.30. Based on a provider-customer relationship between ASs, an AS $X's$ relation tree includes AS X and all customers of the AS X and all customers of the customers, until reach the edge of AS relationship network. With TRA addressing, the AS relationship tree can be identified easily because

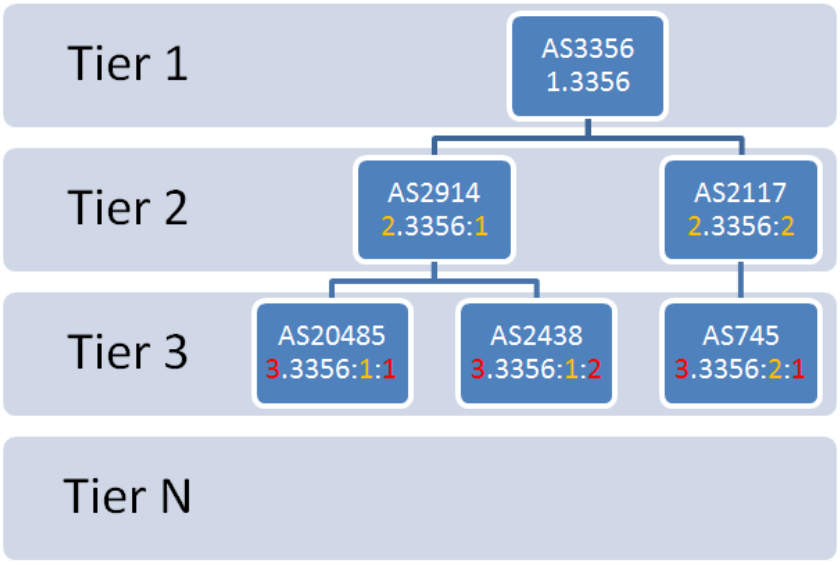


Figure 5.29: TRA allocation to ASs

TRA address allocation process is followed by AS relationships, and TRA address tree represents affected area of an event in the AS topology in FCT.

Figure 5.31 shows result of TRA allocation for Level3 ISP and its address tree. Level3 is identified as tier 1 AS and its TRA address is started from ASN of Level3, which is 3356. Customers of Level3 got TRA address from Level3, thus their address is also started from 3356 and appended unique number for tier 2 level TA, and continues to the lowest tier, which is 7. In each row we show the actual number of Provider ASs and Stub ASs that are supported in this tree. If we sum the values in the circles, we will get the value 13791 as shown in row 1 of Table 5.7.

In Figure 5.31, we also highlight the effect of nesting. The dotted lines are drawn from an AS at tier 4. The AS cloud which has adopted nesting is the 99th AS or child under the AS cloud 2.3356:248. As the cloud 3.3356:248:99 starts a new nesting of clouds, its address is now given in the 4th column

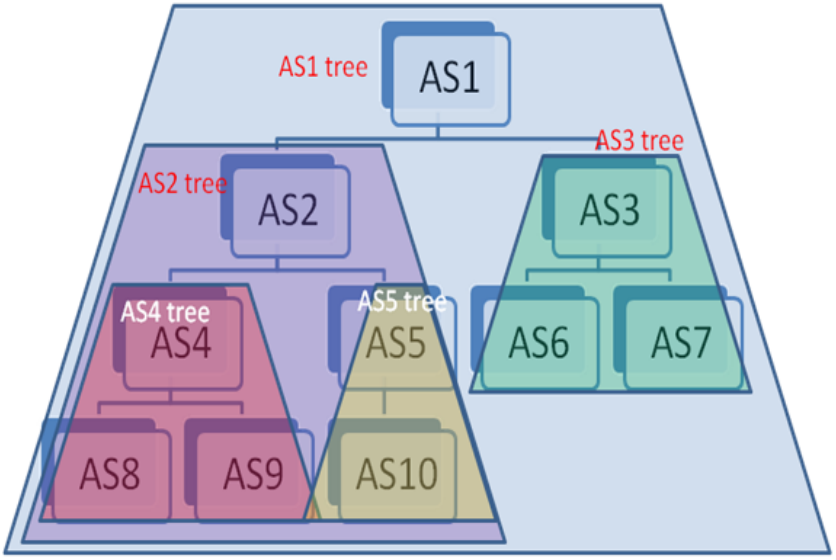


Figure 5.30: Concept of TRA Address Tree

	Provider AS	*Stub AS*	*Example TRA*	*Example Nested TRA*
Tier-1	Level3		1.3356	1.3356
Tier-2	955	1656	2.3356:248	2.3356:248
Tier-3	1074	8015	3.3356:248:20	3.3356:248:99{1.20}
Tier-4	134	1281	4.3356:248:20:72	3.3356:248:99{2.20:72}
Tier-5	3	129	5.3356:248:20:72:1	3.3356:248:99{3.20:72:1}
Tier-6	1	1	6.3356:248:20:72:1:1	3.3356:248:99{4.20:72:1:1}
Tier-7		1	7.3356:248:20:72:1:1:1	3.3356:248:99{5.20:72:1:1:1}

Figure 5.31: Level3 Address Tree with and without Nesting

as 3.3356:248:99{1.20}. A sample TRA for the nested addresses for all the clouds below this tier is thus given in the last column. Through this nesting approach, we have reduced the footprint of the largest Level3 address tree size from 13,791 by 1212. This may not seem significant as the nesting started at

tier 4. Nesting can be introduced at tier 2, in which case the impact on the footprint of the tree and the *churn rate* would have been very significant.

Table 5.7: Largest TRA Address Tree at Each Tier(Largest base)

Tier	TRA of Rooted AS	Root AS Name/ASN	Tree Size	Avg Size
1	1.3356	Level 3/3356	13791	1078.0
2	2.701:58	NTT Com/2914	894	10.4
3	3.701:58:91	Transtelecom/20485	251	3.1
4	4.701:58:91:21	ZSTTKAS/21127	30	1.9
5	5.4323:7:11:15:11	AFCONC/745	6	1.7
6	6.6939:45:36:31:1:1	FASTHIT/24381	2	n/a
7	7.3356:248:20:72:1:1:1	Selements/45594	1	n/a

Part of the tiered addresses so allocated is presented in Table 5.7 for the AS at the different tiers which showed the maximum number of ASs under them. Column 2 of Table 5.7 records the TRA addresses allocated to the *largest* AS at any tier. Column 3 provides the name of the AS and their ASN. Column 4 shows the number of ASs that have their addresses derived from the AS shown in column 3. At tier 1, Level 3 has the maximum number of AS who derive addresses from it. The number of AS that are supported by Level 3 was 13,791. At tier 2, NTT communications, serviced by Verizon Business/UUnet (ASN 701) is the largest tier 2 AS, supporting 894 ASs. Similar interpretation can be extended to other row entries in Table 5.7. If a change occurred to this ISP (tier 1, 2 etc.) then the number of AS that will be affected is given by its tree size. The value of tree size thus directly gives the churn rate. If there is a change in Level 3 at tier 1, then the number of AS that will be affected, is 13,791. Similarly the number of AS that will have to update their routing information if a change occurred at NTT Communications is 894. These values have been calculated without considering FCT nesting capability and hence the dependency extends to all ASs in all tiers under the AS noted in column

3. Table 5.8 records the TRA addresses allocated to the *smallest* AS at any tier, and tree sizes at tier 1, 2, and 3 shows smaller size than the largest base.

Table 5.8: Largest TRA Address Tree at Each Tier(Smallest base)

Tier	TRA of Rooted AS	Root AS Name/ASN	Tree Size	Avg Size
1	1.3549	Global Crossing/3549	9123	1078.0
2	2.3549:834	Rostelecom/12389	411	10.4
3	3.174:1280:2	Energis-Ireland/8760	174	3.1
4	4.174:1280:2:4	ZSTTKAS/21127	30	1.9
5	5.4323:7:11:15:1	AFCONC/745	6	1.7
6	6.6939:45:36:31:1:1	FASTHIT/24381	2	n/a
7	7.6939:45:36:31:1:1:1	Selements/45594	1	n/a

5.2.3 Churn Rate Analysis of TRP

- Average Churn Rate at Tier 1: The information provided in Figure 5.32 focuses only on tier 1 provider ISPs and records all tier 1 ISPs (31 of them) that support customer AS and their tree sizes. The information is to show the worst case scenario if any tier 1 ISP changed its TRA address under the FCT model. As stated earlier, changes at Level 3 can impact 13,791 out of 33,508 ASs, which is 41.15%. Given that currently 80% of the events in BGP are globally visible, the worst case situation in the non-optimized (without nesting) tiered model is twice as better [20]. Many of the tier 1 AS however affect less than 100 ASs. If one considers all tier 1 ASs and average their tree size, the average size will be 1,078. Hence a change at any tier 1 AS will impact on an average only 1,078 ASs, which is 3.21% of the AS in the Internet today.

- Average Churn Rate at Each Tier: Figure 5.33 records the average of all the tree sizes of all ASs at the different tiers. At tier 1 as stated earlier

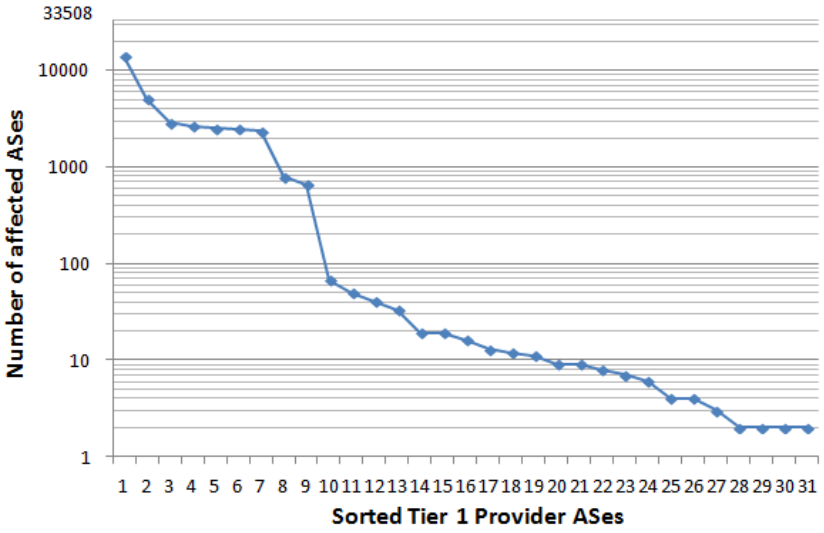


Figure 5.32: Number of AS impacted a Tier 1 Provider AS (sorted)

the average value was 1078.0. At tier 2 the average value is 10.6. That is if there is any change in a tier 2 AS on average only 10.6 ASs will be impacted. (These are also values recorded without considering nesting under the FCT model). A closer look reveals that the average churn rate for events between tiers 2 and 7 is less than 0.04% of the world ASs.

The analytical studies in [72] show that the average number of affected nodes (routers) is around 50% for edge nodes and around 3.5% for core nodes in BGP. Under the tiered model, the maximum of the average value recorded at each tier happens at tier 1 (3.2%). The average values recorded for the rest of the tiers is less than 0.04%. While with BGP 50% of the routers are effected by events at edge nodes, in the case of the tiered model, changes in edges affect only the routers that are in direct connection, thus the rest of the Internet is not affected at all i.e. close to 0% effect. Recent studies on BGP, have also indicated that network events occurring near edge (stub) AS

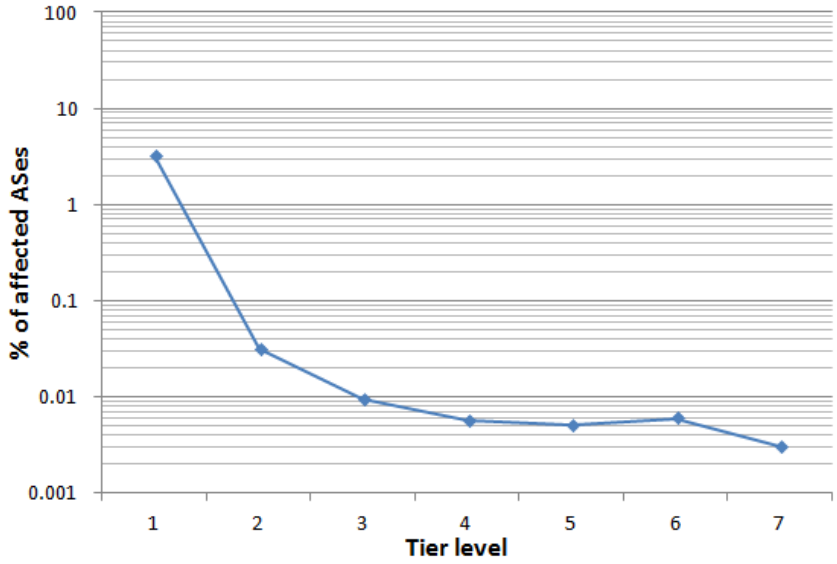


Figure 5.33: Average Number of Affected AS at Each Tier

are more than events in the core AS, since core AS tend to stable than edge AS [72]. Incorporating these probabilities into the statistics collected for churn rates with the FCT model would indicate the huge magnitudes in improvement that can be achieved with the proposed model.

5.2.4 Routing Table Size Analysis of TRP

In TRP, routing table sizes is based on number of links (node degree) connected to an AS. With the CAIDA dataset, Figure 5.34 shows routing table size distribution of TRP, more than 10 entries in a routing table is shown (1,662 ASs) which means 95% of ASs have less than 10 entries in TRP routing table. The Largest routing table size based on the TRP will be 2,631 at tier 1 AS (i.e. Level3) without nesting. This is 186 times smaller than current BGP table size which has 490,000 entries. An average size of TRP routing table size is 4.48 and 12,245 ASs have only 1 entry in TRP routing table that indicates

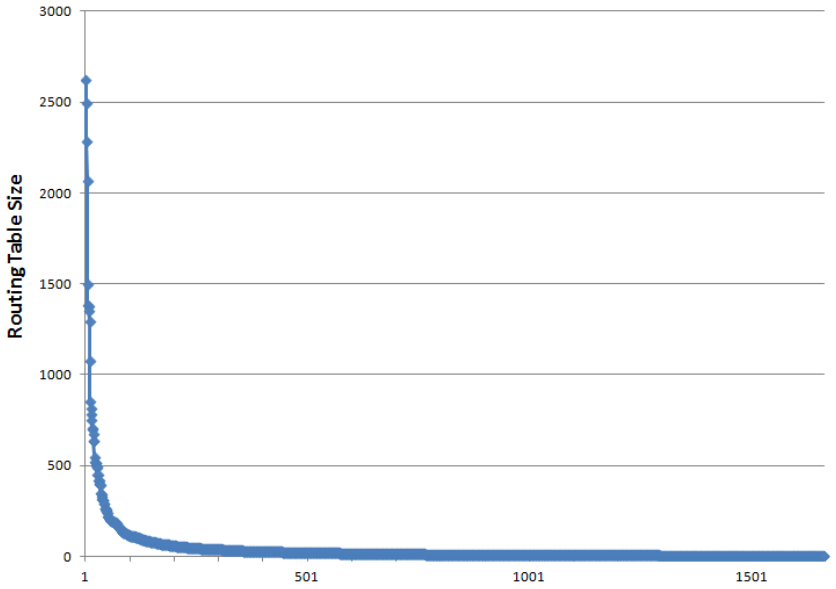


Figure 5.34: Routing Table Distribution of TRP

these AS are single homed and stub ASs. 13,028 ASs have 2 entries in TRP routing table, this is 40.1% of the AS topology.

Figure 5.35 shows distribution of TRP routing table size at each tier level. Larger routing table size tends to appear at higher tier level, especially at tier 1 because higher tier ASs have more customer ASs. While the routing tables with BGP and the current IP address format would increase exponentially with increasing number of networks, the routing table entries in the case of the TRP model would be affected only when there is change in the directly connected network segments/routers. The studies indicate the magnitudes of reduction in both churn rates and routing table sizes that can be obtained when compared with BGP.

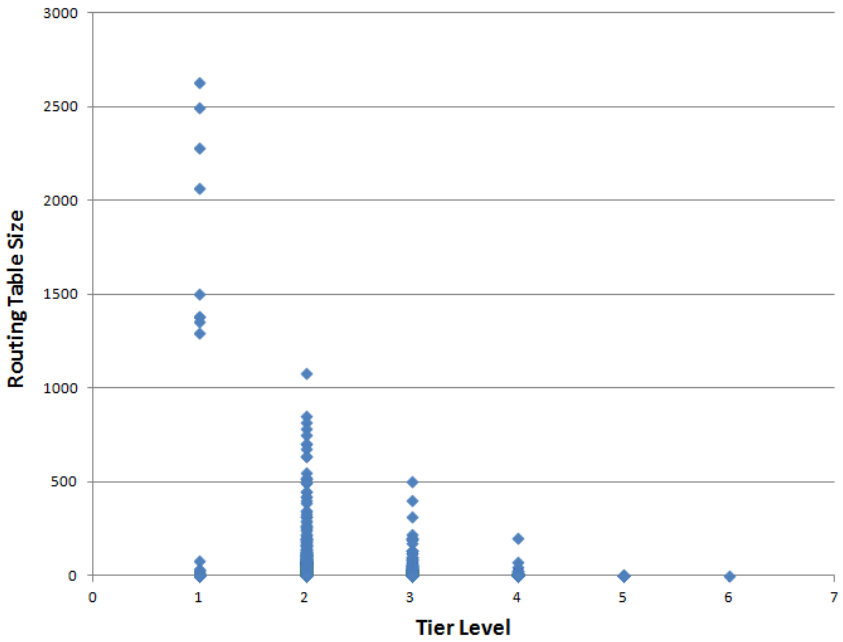
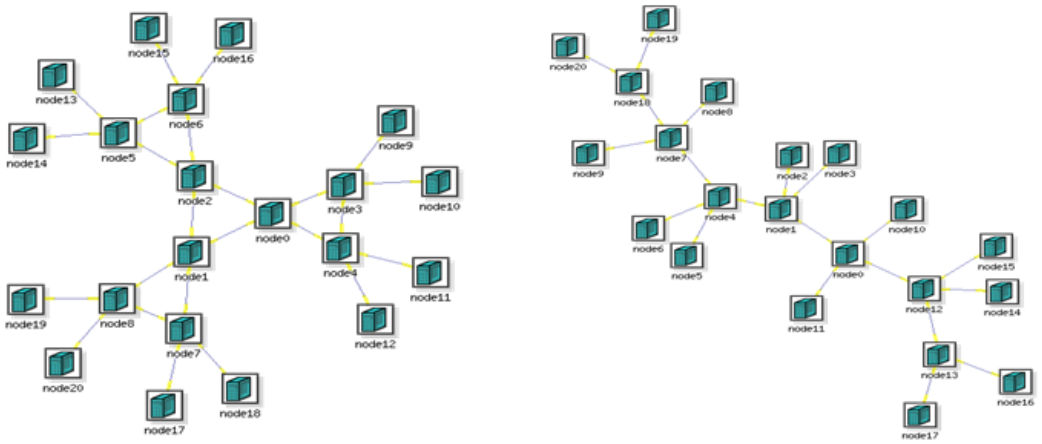


Figure 5.35: TRP Routing Table Size Distribution at Each Tier

5.2.5 Performance Statistics and Analysis of TRP on Testbed

The performances of BGP and TRP, during the convergence phase after link failure and churn rate are presented in this section. Quagga can also operate BGP and configure for BGP operation was set up in the Emulab PCs. Figure 5.36 shows two different type of topology used in the Emulab.

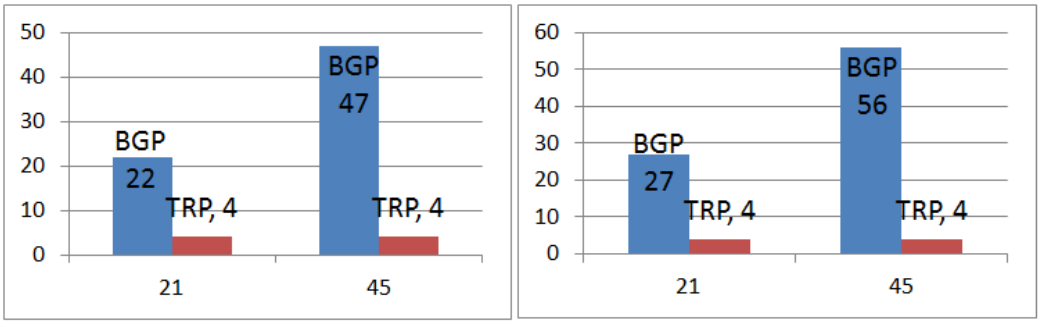
- Routing Table Size: In Figure 5.37, the routing table sizes collected for the two topologies in both case of TRP and BGP. The figure records the maximum of the routing table entries noted in the routers. In the case of BGP, these values are around the same numbers as number of nodes in the topology. In the case of TRP, the routing table entry reflects the number of directly connected neighbors, so in both cases, i.e. the 21



a) Tree like topology b) Caterpillar like topology

Figure 5.36: Testbed Topology used for BGP comparison

router and 45 router networks the maximum routing table entry was 4 which is the same with the case of OSPF comparison because number of network interfaces are limited to 5 in Emulab. TRP routing table sizes do not depend on the network size, which provides proof to the scalability of TRP.



a) Tree like topology b) Caterpillar like topology

Figure 5.37: Maximum Routing Table Entry Size

- Convergence Time after Link Failure: The same methodology used in OSPF comparison is also used for the BGP comparison. Figure 5.38 is the routing table update time in seconds subsequent to a link failure detection. While BGP shows an update time of 147 to 185 seconds for the 21 router topologies, TRP update times were as low as 220 ms to 230 ms.

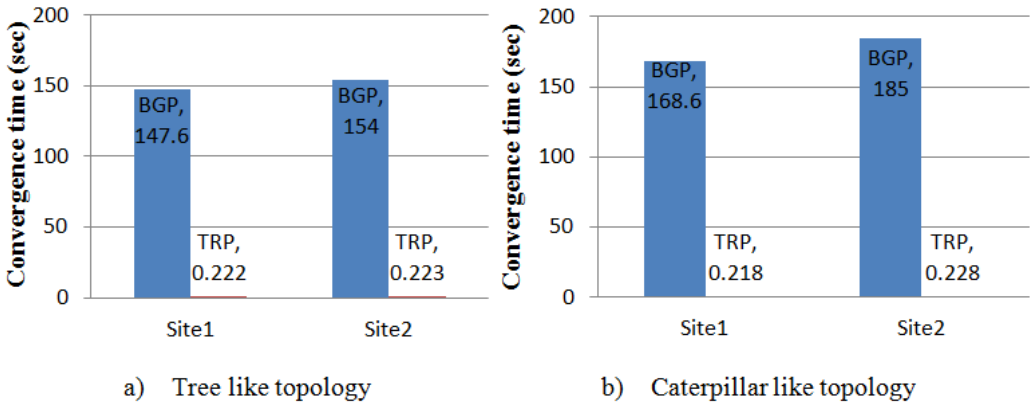


Figure 5.38: TRP vs. BGP Convergence Time after Failure (sec) of 21 nodes topology

- Churn Rate: Total number of nodes which updated their routing table subsequent to link failure detection was observed. In the case of BGP, all nodes in tree like topology (both 21 and 45 nodes) were updated which means that churn rate is 100%. 18/21 nodes and 45/45 nodes in caterpillar like topologies were updated their routing tables, so that churn rates are 86% and 100% respectively. In the case of TRP, only 5 nodes were updated in any topology, so that churn rates of 21 nodes and 45 nodes are 20% and 11% respectively.

5.3 Evaluation of Integrated TRP and MMT

To evaluate integrated TRP with MMT discussed in Chapter 4.4, TRP+MMT and IP+OSPF protocols are compared with Seattle POP of AT&T network used in Figure 5.11. We conducted OPNET [66] simulations for both TRP+MMT scheme and IP+OSPF as part of our research. We tested our scheme in various POPs of the AT&T topology. We are presenting the results of the simulation from Seattle POP which consisted of 393 routers with 437 links within the POP as per Rocketfuel [13] data. The clouds of backbone, distribution and access routers are shown distinctly in the Figure 5.39. We considered the routers that had links to other cities as Backbone routers (BB). The routers that were connected to customer networks were assumed to be Access (AR) routers and the rest of the routers were Distribution (DR) routers. 6 routers out of the 393 were Backbone routers, 94 routers were Distribution routers and the remaining 293 routers were Access Routers. These were the assumptions underlying the categorization.

5.3.1 IP+OSPF

Protocol implementation details of the AT&T internal network were not available thus we decided to run OSPF in the POP. In the absence of real network information from the POP like IP addresses of the various router interfaces and actual hierarchy we made a few assumptions to best run OSPF. OSPF inherently has a 2 level hierarchy. In order to get the best results out of the OSPF simulation we divided the POP into a 2 level hierarchy where we had the backbone OSPF area and other standard OSPF areas talking to each other through the backbone OSPF areas. The backbone routers that we identified

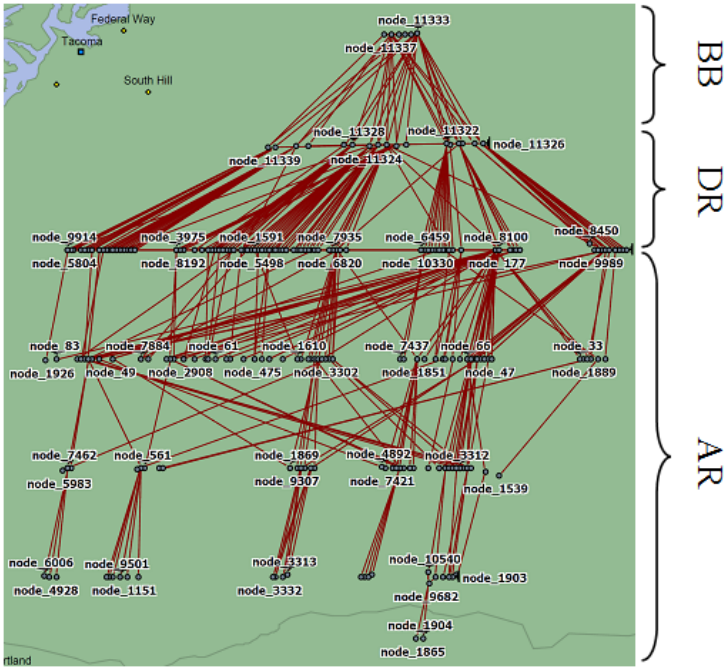


Figure 5.39: AT&T Seattle POP for OSPF simulation

from the Rocketfuel data were internal backbone OSPF routers. The routers that were 1 hop away from the identified backbone routers were Area border OSPF routers. Areas were assigned only if 1 hop routers had links to a 2 hop router. Once OSPF areas were assigned, there were a few links interlinking two different non backbone OSPF areas. Since OSPF does not allow adjacency between routers in different OSPF areas we disregarded these links for the simulation. All the areas in the POP were standard non stub areas. 14 non backbone areas were implemented in the POP. Rocketfuel data provided a single IP address for each router and so we were unable to allot IP addresses that the actual network was using. We allotted /24 networks for each different link in the POP.

5.3.2 TRP+MMT

Firstly, we identified tiered structure of Seattle POP in AT&T network as explained in Chapter 5.1.2, then applied MMT. The routing overheads using MMT is categorized into two; one as the number of VIDs that a node has which indicates its connectivity within the cloud; and the list of uplink, trunklink and downlink clouds that it is connected to. In the MMT based routing approach every router that is connected to the uplink or via a trunk-link to a neighboring cloud is a root for the creation of a meshed tree. In our topology, there are totally 313 clouds. We do not use MMT in cloud that have only one node as no routing is needed. In this simulation, we limited each node to accept only two VIDs from a particular CH which effectively meant that we have two paths to each CH. Details of address allocation of MMT is also explained in Appendix B.

5.3.3 Results

We compared the IP network running OSPF with our architecture running MMT. OSPF needed 30 seconds to converge where as MMT running in each of the clouds converged less than 2 seconds. OSPF routing overhead was approximately 18 Mbytes compared to 0.83 Mbytes for MMT in an hour long simulation. Apart from this the maximum routing table size in OSPF was 416. On the other hand for MMT the maximum was 47 which was a sum of the VIDs and the associated node table.

5.4 Transition Study with MPLS Approach

One major contribution of our work was the study of MPLS as a transition platform to introduce TRP and replace IP and its routing protocols. MPLS achieves similar goals in terms of replacing IP and the routing protocols, but uses the routes from IP routing tables to determine the MPLS paths. Once the paths are established MPLS bypasses the use of IP in the MPLS aware routers. Another feature of MPLS that aided the transition studies was the use of label and label stacking, where in the proposed transition the labels serve to carry the TRP addresses, and label stacking was used to achieve the tiered functionalities i.e. forwarding across tiers. The packet forwarding decision is the same as Algorithm 1. In this section, the implementation details are presented.

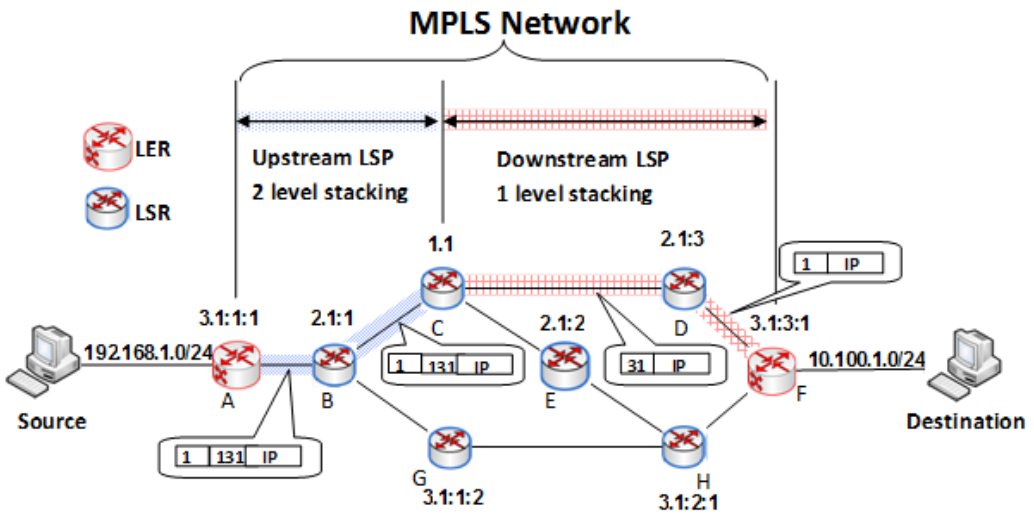


Figure 5.40: MPLS enabled network with TRP

In Figure 5.40, there are eight MPLS aware routers, Routers A to H. Of these Routers A and F are Label Edge Routers (LER) and the others are

Label Switch Routers (LSR). TRAs were assigned to all MPLS aware routers as shown in the figure. Based on the TRAs, it can be noted that Router C is a tier 1 router, Routers B, E, and D are tier 2 routers, while Routers A, G, H and F are tier 3 routers. To conduct the feasibility study, the MPLS tables were manually populated as shown in Tables 5.9 and 5.10 for LER Router A and F. For real implementations using MPLS, the operation of MPLS and its process of populating the tables have to be modified and are not included in this article.

Table 5.9: LER MPLS Table of Router A {3.1:1:1}

LER Table			
Destination Network	*Out Label*	*Action*	*Next Hop*
10.100.1.0/24	1(L1) 131(L2)	PUSHx2	Router B
L-1 Label Table			
In Label	*Out Label*	*Action*	*Next Hop*
1	IP header	POP	IP address

Table 5.10: LER MPLS Table of Router F {3.1:3:1}

LER Table			
Destination Network	*Out Label*	*Action*	*Next Hop*
192.168.1.0/24	1(L1) 131(L2)	PUSHx2	Router D
L-1 Label Table			
In Label	*Out Label*	*Action*	*Next Hop*
1	IP header	POP	IP address

We first explain the use of the tables. The first table in Table 5.9 is for Router A, which is a LER. This router is connected to the IP network 192.168.1.0/24. However, to forward a packet to the destination network 10.100.1.0/24, the forwarding table has an entry for the purpose. Interpreting this table; when a packet arrives with 10.100.1.0/24 as the destination address,

LER A will *push* two labels 1 and 131 where 1 is the outer label (L-1). This packet will then be sent on to the next hop which is Router B. If a packet arrives to be delivered to network 192.168.1.0/24 at LER A, Router A will *pop* the L-1 label and then forward the packet to the destination IP address in the packet. Similar entries can be noted for LER F in Table 5.11, which will also perform operations similar to Router A.

Table 5.11: LSR MPLS Table of Router B {2.1:1}

L-1 Label Table			
In Label	*Out Label*	*Action*	*Next Hop*
1	1	SWAP	Router C
2	N/A	POP	N/A
11	1	SWAP	Router A
12	2	SWAP	Router G
L-2 Label Table			
In Label	*Out Label*	*Action*	*Next Hop*
11	1	SWAP	Router A
12	2	SWAP	Router G

Table 5.12: LSR MPLS Table of Router C {1.1}

L-1 Label Table			
In Label	*Out Label*	*Action*	*Next Hop*
1	N/A	POP	N/A
L-2 Label Table			
In Label	*Out Label*	*Action*	*Next Hop*
111	11	SWAP	Router B
131	31	SWAP	Router D

At LSR B, it will check the outer label when a packet arrives from Router A and processes the packet forwarding based on the outer label (L-1, tier 1) table. As per this table, when the packet arrives from Router A, if it has a forwarding address where the tier value is 1 (L-1), then the packet will be sent uplink to Router C with a *swapped* label which will also have a value 1. If

Table 5.13: LSR MPLS Table of Router D {2.1:3}

L-1 Label Table			
In Label	*Out Label*	*Action*	*Next Hop*
1	1	SWAP	Router C
2	N/A	POP	N/A
31	1	SWAP	Router F
L-2 Label Table			
In Label	*Out Label*	*Action*	*Next Hop*
31	1	SWAP	Router F

the outer label (L-1) was 2, it indicates that the anchor tier level is 2 in the forwarding TRA, and Router B is the anchor router (at which time redirection will take place). Hence, Router B will *pop* the L-1 label and the packet will then be processed as per L-2 label table. In the L-2 label, when a packet is received, Router B will *swap* the incoming labels with new labels to deliver the packet to either Routers A or G. Similar entries can be noticed for Routers C and D and their operations will be similar to that explained for Router B and tables are shown in Tables 5.11, 5.12, and 5.13 for LSR Router B, C, and D.

Handling tier based forwarding with MPLS can be summarized as:

- For upstream forwarding, a L-1 label indicates that a MPLS packet is to be forwarded until the upper tier level specified in the label is reached. If L-1 label value is lower than router's tier value, it is forwarded to an upper tier.

- For downstream forwarding, if L-1 label value is the same as router's tier value, the router removes (*pop*) L-1 label and forwards the packet to a lower tier based on L-2 label.

We now work through an example of packet forwarding in the network scenario shown in Figure 5.40. Let the source node send a packet to a desti-

nation node with destination IP address 10.100.1.x, where x is the host identifier. LER has to be aware of the TRA allocated to network with IP address 10.100.1.0/24. This TRA is 3.1:3:1. Following are the steps.

1. Forwarding TRA calculation: Router A calculates the forwarding TRA to 3.1:3:1 by comparing with own TRA (3.1:1:1) with destination TRA 3.1:3:1. The forwarding TRA will be 1.1:3:1.

2. Adding MPLS header: Router A add two MPLS label to the packet using two *push* operations, where the L-1 label is 1, L-2 label is 131. The packet is then forwarded to the next hop Router B.

3. 1st hop: Router B checks the outer label i.e. L-1 label value of 1. This is less that Router B's tier value 2. Thus, the packet will be forwarded to an upper tier based on L-1 label table. In this case, the label will be *swapped* to 1 and then the packet will be forwarded to next hop Router C.

4. 2nd hop: Router C checks L-1 label value of 1. This equals Router Cs tier value of 1. Router C will remove the L-1 label through a *pop* operation and then packet should now be redirected. Router C will hence check the L-2 label value which 131 in the packet and compares it with its L-2 label table entry. Then, Router C forwards the packet to the next hop Router D after *swapping* the label from 131 to 31.

5. 3rd hop: Router D checks L-1 label value 31 and lookups its L-1 label table. It will *swap* 31 to 1 and then forward to the next hop Router F.

6. Removing MPLS header: Router F checks the L-1 label value of 1 and lookup its L-1 label table. It will then *pop* (removes the MPLS header

from the packet) and checks the IP header to forward to the final destination.

From our perspective, MPLS-based approach can offer a neat transition path to adopt the FCT architecture to the community.

5.5 Discussions

In this Section, we experimentally evaluate TRP's convergence times, control overheads, and routing table size. We further logically discuss TRP's scalability and portability here.

The tiered routing address in the FCT architecture is used by TRP for packet forwarding. Initial convergence time and convergence time after failure are significantly low because TRP does not require message flooding to all nodes in the network or any significant recalculations and recomputing on topology change. As a result of no message flooding, control overheads are also very low. Entries in routing tables used by TRP are satisfied by addresses of only the direct neighbors because of the inherent routing information in the TRA. Thus, the routing table sizes in TRP are significantly smaller.

The tiered address is a topology-independent hierarchical address and leverages the tiered structure existing in the network topologies in local area networks (LAN), ISP's router-level networks and Internet AS-level networks. An important property of TRP is that it must identify the top tier (tier 1) nodes or clouds in the network to assign tiered addresses. Tier 1 nodes or clouds can be a gateway router in the LAN, routers connected with different Point of Presences (POPs) within an ISP, and tier 1 ISPs that do not have any provider ISPs. TRP can run at any level once the top tier is identified.

Thus, TRP can be used as both an inter-domain routing protocol and an intra-domain routing protocol, and it can also cope with scalability due to the address nesting concept.

In TRP, the tiered address is assigned to a node, not a network interface. But, it is not limited to only one address per a node. If a (child) node is connected to more than one upper tier (parent) nodes, the child node can get more than one address, and use one of them as the primary address. With multiple tiered addresses, a child node can have local preferences or policies for forwarding. The child node has the ability to select which link/path/address to forward a packet on. This feature can be very useful not only for the end-user, but also for ASes that use policies in BGP routing.

Like the current Internet, the proposed FCT Internet Architecture also requires the use of a name resolution system like the Domain Name System (DNS) because the tiered address is not human friendly. However, the structure of the tiered addressing scheme is very similar to the structure of the domain name system, and transition from current DNS to tiered address name resolution system should be straightforward.

Chapter 6

Transition Study with Economic Model

A new routing protocol for inter- and intra-domain routing is being investigated. However, attractiveness for adopting new routing protocols is not quantitatively studied. Many of proposed protocols are considered incremental solution, which means that the protocol can be gradually adopted over a period of time. During this adoption period, the adopters of new routing protocol have full comparability. Although incremental solution helps in adoption, it is neither a necessary nor sufficient condition. This is because incremental solution is an inherent property of the routing protocol. Therefore, we believe that clean slate solution is the key for future and understanding adoption of non-incremental solution is important.

6.1 Assumptions and Cost Entities

A major study in the investigations will be the economic viability and sustainability when introducing the new routing protocol to replace existing routing protocols. This requires also replacing the current routing equipment running current routing protocols with new equipment that will run the new protocols. Towards this, the transition costs and the long-term benefits of adoption of the new routing protocol by the vendor community will be analyzed. In this study, the vendor communities are considered as the Internet Service Providers (ISPs), who will bear the costs of deployment of the new equipment to run the new routing protocol. A cost model to aid in the analysis is developed for the purpose and *our model is based on benefits offered by the new routing protocol to an ISP and is only considered standalone benefits*. In the cost model, transitions costs are assessed under two categories: 1) running costs of old routing protocol and 2) running costs of new routing protocol. The running costs are estimated by operating cost of the routing protocol includes three components that router maintenance (RM), human resource (HR), and power usage (PU) costs. In addition to the operating costs, an investment cost and a salvage value are considered for the running costs of new routing protocol. (Figure 6.1)

6.1.1 Transition Model

During the transition period, both routing protocols have to be running in an ISPs network because the transition has to be effected in a step-by-step manner over weeks or months as new equipment replaces the current ones to reduce risks. Thus during the transition period, routers running the two

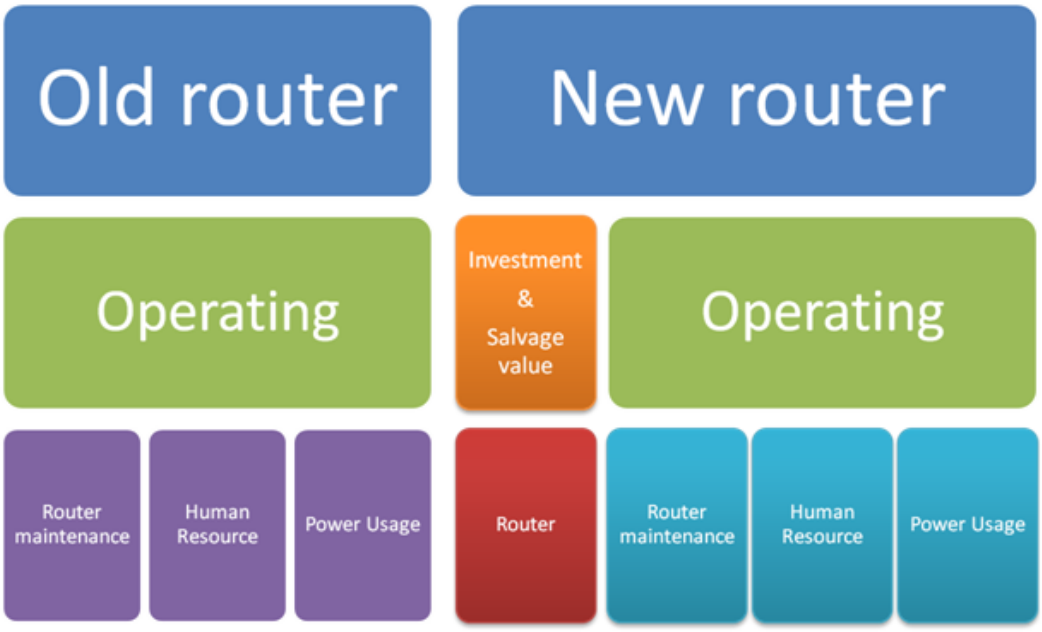


Figure 6.1: Cost components of old and new routing protocols in an ISP

protocols will coexist until the current routers are completely phased out. During the transition the total number of routers may exceed the current count as some current and new routers will have to operate in parallel and current routers will removed off the network as new routers take the operational load and start performing as expected. The operations costs during transition will thus consider, i) costs incurred to run the current routers with the current routing protocols and, ii) costs for running the new routers with the new routing protocol. The operations cost for the new implementation is considered separately as it is envisaged that this cost would be different and *significantly lower than that of the current implementation.* The total cost at any time during the transition thus will be the sum of operating costs of both the current and new implementations, to which has to be added the investment cost for the new implementation. The numbers of routers running the current

protocol and numbers of router running the new protocol and the total number of routers will vary with time as the new routers replace and take over the tasks of the current routers and will manifest as variable costs. In the model, the numbers and the corresponding costs will be represented as functions of time, and time will be measured in units of weeks.

6.2 Types and Number of Routers

Those cost components are closely related to the size of an ISP in terms of numbers of routers within the ISP. An ISP uses different types of routers. Hence the number of routers in each type will be counted and accounted for separately. Thus, in the study it is necessary to know the number of routers owned by an ISP in each type. To estimate number of routers and types of routers in an ISP, we use the Rocketfuel dataset which is also used to identify the tired structure within an ISP in Chapter 4.1. We use AT&T network as example to explain our methodology to identify types of router.

An ISP has several Points of Presence (POPs) which are located in some major cities for example in the US the POPs are located in New York, Seattle, Chicago and so on. The POP forms the backbone of that service provider. Each POP comprises of several routers, some of which primarily connect to other backbone (BB) routers in other POPs. The BB routers also connect to the distribution routers (DR). The distribution routers besides connecting the access routers (AR) to the backbone also provide redundancy and load-balancing between the backbone and access routers. The ARs in turn connect to customer or stub networks. The different types of routers have different specifications. In general, BB routers are high performance computing devices

that cost more. The second in complexity and cost are the distribution routers. To develop the cost model it is necessary to know the different types of routers and their numbers hosted by an ISP.

6.2.1 Identification of Router Types by Connectivity

Tier 1 ISPs provide Internet services worldwide and have POPs in major cities of the world. To scope the conducted study, the collected data was limited to ISP POPs in the United States (US). Since the topological data contains location of each router, it is possible to identify the POP and the ISP to which a router belongs. An examination of AT&T ISP network information from the Rocketfuel database, revealed a total 11,403 routers, 13,689 links, and 110 POP locations. From the POP information, the BB routers were identified by their links to other POPs. Within each POP, all edge routers were recognized as ARs. Routers connecting BB routers and AR routers were then identified as the DR routers. Based on these classification 389 BB routers, 6,395 DR routers, and 4,619 AR routers were identified within AT&T ISP network. The statistics so obtained are recorded in Table 5.2.

6.3 Modeling

To estimate the transition costs during the adoption period, cost components displayed in Figure 6.1 are used to develop our model. In this model, *we first assume that total number of routers* ***before*** *the transition and* ***after*** *the transition is the same and all of routers will be replaced with the new router.* However, number of each router type (BB, DR, and AR) may be changed based on routing complexity.

Our model consists of the total number of routers R_{total}. R_{total} is function of time t week and it includes number of both old routers and new routers. The number of new routers be R_{new}, and the number of current be R_{old}, where R_{new} will be function of the number of BB, DR and AR routers running the new routing protocol (TRP), and R_{old} will be a function of the number of BB, DR and AR routers running the old protocol such as OSPF and BGP. Hence at time t week, total number of routers R_{total} is:

$$R_{total}(t) = R_{new}(BB_{new}(t), DR_{new}(t), AR_{new}(t)) + R_{old}(BB_{old}(t), DR_{old}(t), AR_{old}(t)) \quad (6.1)$$

where BB is a backbone router, DR is a distribution router, and AR is an access router. The suffixes appended to each type of router gives their numbers. Suffix *new* is used to show the number of routers that implement the new solution, while suffix *old* denotes the number of routers that implement the current solution. Number of new routers, BB_{new}, DR_{new}, and AR_{new} are can be determined by transition scenarios.

6.3.1 Router Investment Cost (IC) and Salvage Value

In our model, investment cost for new routing protocol is cost of purchasing new routers. When purchasing new hardware, it plans to use the hardware for several years and at the end point of the plan, companies believe that it will be able to resell the hardware for some value. Therefore, when estimating the investment cost, it is important that the salvage value is also estimated.

Furthermore, the value is a part of the annual depreciation that must be recorder each year.

Investment Cost Assumptions:

- Router price is depended on the type of router
- When new routers are purchased, the cost included the Operation System (software) for the router
- Investment cost of old routers are excluded in this model since its already offset
- Salvage value for new routers is depended on how long the router will be used and to make it simple, we applied 5% of the price is considered as salvage value and it will be subtracted in each new router price when they are estimated.

Thus, total investment cost (IC_{total}) during a period between week t_1 and t_2 can be estimated by price of new routers and number of new routers has been increased (purchased).

$$IC_{total}\,(t_2 - t_1) = IC_{new\_bb}\,(t_2 - t_1) + IC_{new\_dr}\,(t_2 - t_1) + IC_{new\_ar}\,(t_2 - t_1) \quad (6.2)$$

$$IC_{new\_bb}\,(t_2 - t_1) = BB_{new}\,(t_2 - t_1) \times BB_{pr} \tag{6.3}$$

$$IC_{new\_dr}\,(t_2 - t_1) = DR_{new}\,(t_2 - t_1) \times DR_{pr} \tag{6.4}$$

$$IC_{new\_ar}\,(t_2 - t_1) = AR_{new}\,(t_2 - t_1) \times AR_{pr} \tag{6.5}$$

where IC_{new_bb}, IC_{new_dr}, and IC_{new_ar} are investment cost of BB, DR, AR routers which are purchased during $(t_2 - t_1)$ weeks. Since the price of router is different by type of router, we estimated investment cost separately by type of routers. The prices of these new routers are noted as BB_{pr}, DR_{pr}, and AR_{pr} that are estimated separately in the later section.

6.3.2 Router Maintenance Cost (RM)

Router maintenance includes parts failure replacement and software updates. These are supported by vendor service which is yearly base contract. The cost of this yearly based cost can be divided by 48 to be estimate one week cost of it. It is typical that the vender service is started when new router is purchased.

Thus, total router maintenance cost (RM_{total}) during a time period $(t_2 - t_1)$ weeks can be estimated by

$$RM_{total}\,(t_2 - t_1) = RM_{new}\,(t_2 - t_1) + RM_{old}\,(t_2 - t_1) \tag{6.6}$$

$$\begin{aligned} RM_{new}\,(t_2 - t_1) &= BB_{new}\,(t_2 - t_1) \times BB_{mt} \\ &+ DR_{new}\,(t_2 - t_1) \times DR_{mt} \\ &+ AR_{new}\,(t_2 - t_1) \times AR_{mt} \end{aligned} \tag{6.7}$$

$$\begin{aligned} RM_{old}\,(t_2 - t_1) &= BB_{old}\,(t_2 - t_1) \times BB_{mt} \\ &+ DR_{new}\,(t_2 - t_1) \times DR_{mt} \\ &+ AR_{old}\,(t_2 - t_1) \times AR_{mt} \\ &[t_2 > t_1, t_2 - t_1 > 0] \end{aligned} \tag{6.8}$$

where RM_{new} and RM_{old} are total router maintenance cost of old and new routers during $(t_2 - t_1)$ weeks. Since costs of the vender service are different by router types, *old* and *new* router are estimated separately. The yearly maintenance fee of each type of these new and old routers are BB_{mt}, DR_{mt}, and AR_{mt}.

6.3.3 Human Resource Cost (HR)

HR cost for routing in an ISP is the wage of the network administrators who maintain their networks. Since the task of network administrator is related to the number of network, number of network administrator can be estimated by total number of router in an ISP. During the transition, additional personnel may be required for installing and testing new routing protocol and the new routers. Also, wages of network administrator may be changed by experiences and level of tasks. In this study, we exclude network administrator who main-

tain servers and network services such as WWW, File share, Database, and so on.

Human Resource Cost Assumptions:

- Different level (entry, intermediate, senior)of network administrator maintains different types of router (AR, DR, BB).
- Number of routers maintained by one administrator is different by types of routers (AR, DR, BB) and estimated by ratio of RM cost between AR, DR, and BB routers because tasks of network administrator and maintenance support service are closely related.

Thus, total human resource cost HR_{total} during a time period $(t_2 - t_1)$ weeks can be estimated by

$$HR_{total}(t_2 - t_1) = HR_{BB}(t_2 - t_1) + HR_{DR}(t_2 - t_1) + HR_{AR}(t_2 - t_1) \quad (6.9)$$

$$HR_{BB}(t_2 - t_1) = \left\lceil \frac{BB_{new}(t_2 - t_1) + BB_{old}(t_2 - t_1)}{BB_{care}} \right\rceil \times BB_{wage} \quad (6.10)$$

$$HR_{DR}(t_2 - t_1) = \left\lceil \frac{DR_{new}(t_2 - t_1) + DR_{old}(t_2 - t_1)}{DR_{care}} \right\rceil \times DR_{wage} \quad (6.11)$$

$$HR_{AR}(t_2 - t_1) = \left\lceil \frac{AR_{new}(t_2 - t_1) + AR_{old}(t_2 - t_1)}{AR_{care}} \right\rceil \times AR_{wage} \quad (6.12)$$

$$[t_2 > t_1, t_2 - t_1 > 0]$$

where $HR_{BB}, HR_{DR}, HR_{AR}$ are a total wage of each router types and $BB_{care}, DR_{care}, AR_{care}$ are number of each types of router take cared by one administrator. $BB_{wage}, DR_{wage}, AR_{wage}$ are wage of network administrators.

6.3.4 Power Usage Cost (PU)

Power usage cost is the most important consideration in the cost estimation model. The studies of the energy consumption of the Information and Communication Technology (ICT) and the concept of energy-efficient networking have gained the attention of research community and also interested from ISPs and Telecommunication operators in terms of economic needs because their infrastructures are continuously growing. Networking devices such as routers are required to be power on 24/7 and a number of these devices owned by ISPs are huge, and energy expenses are becoming an increasingly important.

Table 6.1: Annual energy consumption of some of the major telecom operators and estimated electricity cost. Source: [73, 74]

ISP	*Energy Consumption (2009)*	*Cost ($60/MWh)*
AT&T	11.07×10^6MWh	$664.2M
Verizon	10.27×10^6MWh	$616.2M
NTT	2.75×10^6MWh	$165.0M
China Mobile	10.62×10^6MWh	$637.2M
France Telecom	4.38×10^6MWh	$262.8M
Deutsche Telecom	7.91×10^6MWh	$474.6M

Table 6.1 shows an annual energy consumption of some major ISPs and estimated electricity costs, and millions of dollars were spent and the increasing trend of these costs has been confirmed by report from the industry [74]. Furthermore, growth of customer population and the Internet traffic also contribute to the energy efficiency issues. Thus, reducing power consumption has become a high priority and ISPs are seeking efficient protocols, architectural solution, and innovative equipment that will perform a better power consumption.

The number of routers in an ISP is high and power consumption by the high performance BB routers is also high. PU in this case has to be estimated

by considering the power consumed by the routing operation complexity (software) and the hardware equipments separately. Hardware power usage is the wall-socket power used by a router and is constant by type of router. Software power usage is depends on the CPU and memory usage in a router, which depends on the complexity of the routing operations. The cost associated with the power usage is important in this study, since the new routing protocol would inherently require low processing and low memory usage, which would positively impact the economic justifications.

Power Usage Cost Assumptions:

- Power usage will be that same for the same type of router, which means that PU is not affected by data traffic in this study and we consider the PU by only routing maintenance by its protocol.
- Power usage of new router is estimated by comparing complexity ratio of old and new routing protocol (see Chapter 6.4), types of router, and memory usage.

Total PU cost (PU_{total}) during a time period ($t_2 - t_1$) weeks can be estimated by

$$PU_{total}\,(t_2 - t_1) = PU_{new}\,(t_2 - t_1) + PU_{old}\,(t_2 - t_1) \tag{6.13}$$

$$\begin{aligned} PU_{new}\,(t_2 - t_1) &= BB_{new}\,(t_2 - t_1) \times BB_{new\_pu} \\ &+ DR_{new}\,(t_2 - t_1) \times DR_{new\_pu} \\ &+ AR_{new}\,(t_2 - t_1) \times AR_{new\_pu} \end{aligned} \tag{6.14}$$

$$\begin{aligned} PU_{old}\,(t_2 - t_1) &= BB_{old}\,(t_2 - t_1) \times BB_{old\_pu} \\ &+ DR_{new}\,(t_2 - t_1) \times DR_{old\_pu} \\ &+ AR_{old}\,(t_2 - t_1) \times AR_{old\_pu} \\ &[t_2 > t_1, t_2 - t_1 > 0] \end{aligned} \tag{6.15}$$

where PU_{new} and PU_{old} are PU cost of old and new routers during $(t_2 - t_1)$ weeks. The PU of each new and old routers are BB_{new_pu}, DR_{new_pu}, AR_{new_pu}, BB_{old_pu}, DR_{old_pu}, and AR_{old_pu}.

6.3.5 Total Running Costs and Transition Scenarios

Total operating costs OC_{total} of old and new routers at time t week can be estimated from Equations 6.1 to 6.15 as follows:

$$OC_{total}\,(t) = RM_{total}\,(t) + HR_{total}\,(t) + PU_{total}\,(t) \tag{6.16}$$

As described in above, each cost entity is closely related to the number of routers at time t week. The number of routers replace with new routers at time t week is depended on the ISPs transition scenario. It is not a realistic scenario that all routers in the ISP are replaced at a time because of risk

management and there could be thousands of routers in the ISPs network. There are many possible ways for the transition scenarios. The ISP may replace their routers location base like POP by POP, or type of network base like backbone networks first and distribution networks second, and so on. In this study, we will apply several scenario considered by a size, topology, and number of routers of an example ISP. One of our goals is to show how much the operating cost is reduced before and after the transition. The operating cost before the transition is

$$OC_{total}\left(t_{start}\right) > OC_{total}\left(t_{end}\right) \tag{6.17}$$

where t_{start} is the start time of the transition. At the time, all routers in an ISP are running old routing protocols. Where t_{end} is the end time of the transition and all routers are replaced with new routers at the time. During the transition, any weeks when OC_{total} is exceeded the cost of $OC_{total}(t_{start})$ is considered as transition cost. Thus, transition period can also be determined by the budget of the total transition cost.

6.4 Complexity of Routing Protocol

We believe that complexity of new routing protocol is less than current routing protocols. Complexity of routing protocol can be referred by:

- Routing table size (big vs small)
- Method of populating routing table (SPF vs direct neighbor)
- Number of control packets (flooding vs one hop, churn rate)

- Convergence time of initial and after failure (long vs short)
- CPU usage during all the above operations

To estimate price, yearly maintenance fee, number of administrator required, and power usage of new router, comparison of routing complexity can be used. In this study, we use complexity of routing table update method and routing table size to compare between old and new routers.

6.4.1 Routing Table Size

Routing table size will affect to delay of packet forwarding, CPU usage for destination lookups, required memory size, and power usage. Therefore, reducing size of routing table is an important for cost reduction. Recent software router and old hardware router often use Static Random Access Memory (SRAM) or Dynamic Random Access Memory (DRAM) to store forwarding table entries, and use a longest prefix matching method for IP prefix lookup and time complexity for the look up is $O(\log(w))$, where w is the number of bits in the target IP address prefix [75]. On the other hand, modern hardware routers use Ternary Content Addressable Memory (TCAM), which allow parallel lookups and its time complexity is $O(1)$. However, TCAMs have several limitations:

- High power consumption: TCAM consumes $12 \sim 15$ Watts per chip (18MB) and also proportional to the number of bits enabled in the TCAM during the search operation [76]. 4 to 8 TCAM chips are often used in a router [77].
- Limited capacity: Low cell density compared to SRAM cell that consist of 6 transistors, but TCAM has 16 transistors on a cell [78].

- Cost: TCAMs are expensive.

Based on our TRP routing protocol analysis and evaluation, TRP requires significantly less number of routing table size. Therefore, number of TCAM chips used in new TRP router can be smaller than current routers. Moreover, TCAM can be replaced with SRAM or DRAM, which allow lower cost for TRP router. Since a routing table size is very small, TRP router can implement hash table lookup which has also $O(1)$ time complexity. In addition, TRA address requires less number of address length than IPv4 and v6, which can also reduce power consumption of TCAM chip. Thus, cost of memory and power usage by memory can be reduced and type of memory or number of TCAM chip will be determined by ISPs network and transition scenario.

6.4.2 Method of Populating Routing Table

The way of populating and updating routing table is a big factor of routing protocol complexity. In this cost estimation study, we use AT&T network for the transition scenario and intra-domain routing protocol complexity is compared with OSPF and TRP.

Table 6.2: Route Maintenance Complexity of OSPF and TRP

Routing Protocol	*Complexity*
OSPF	$O(r \times \log(r))$
TRP	$O(1)$ (with hash table) $O(n)$ (without hash table)

The OSPF routing protocol uses the Dijkstra algorithm to find the shortest path for each entry in the routing table and complexity of OSPF is $O(r \times \log(r))$, where r is number of router in the OSPF routing domain [34]. On the other hand, TRP does not required the shortest path calculation

because entries in the routing table is an address of direct neighbors (links) and complexity of TRP is $O(1)$ when hash table used. Without hash table, the complexity will be $O(n)$, where n is a number of direct neighbors.

With a router-level topology, we can estimate the routing protocol complexity at each node (router) based on number of node in the OSPF domain and number of links with the node. Then, router type can be down grade from BB to DR, DR to AR to reduce transition and operating costs. Therefore, number of BB, DR, and AR routers may be different between old and new routers. The threshold condition of router down grade will be referred to router price ratio.

6.5 Transition Scenarios and Estimation

To estimate total costs of the transition, AT&T network and cost ratios between BB, DR, and AR are used.

6.5.1 Number of Each Router Type

Number of new BB, DR and AR routers are determined by routing complexity at location of each router:

$$BB_{new} = BB_{old} - \Delta BB \tag{6.18}$$

$$DR_{new} = DR_{old} - \Delta DR + \Delta BB \tag{6.19}$$

$$AR_{new} = AR_{old} + \Delta DR \tag{6.20}$$

where ΔBB is number of routers down graded from BB to DR and ΔDR is number of routers down graded from DR to AR. Number of ΔBB and ΔDR are also used to estimate for the cost reduction.

6.5.2 Router Investment Cost and Salvage Value

Router price can vary by types and configuration of routers. In this study, we use three types of routers, core, distribution, and access routers and numbers of each type are identified based on AT&T network topology, and ISP uses different router model for different types. Rocketfuel dataset used for analyzing AT&T network also contains domain names assigned for routers. In [79], types and possible models of routers are guessed. Authors in [80, 81] try to estimate power consumption of routers in core, metro, and access network. We use similar router models identified by them in this study.

Table 6.3: Models and Prices of Router

Types	*Models*	*Prices*
Access Router (AR)	Cisco 7603	$10K
Distribution Router (DR)	Cisco 12816	$22K
Backbone Router (BB)	Cisco CSR-1	$100K

Router models and prices of AR, DR, and BB router used in this estimation study are displayed in Table 6.3. Prices of these routers are referred from [82]. However, the price of routers may not be accurate because of different discount system, router configurations, and so on. Thus, we use ratio of the prices to get idea of approximately price range. Price ratios of router types are estimated as follows:

$$[BB_{pr} : DR_{pr} : AR_{pr}] = [10.00 : 2.20 : 1.00] \tag{6.21}$$

Prices of new routers will be subtracted by 5% of its price when investment cost (IC) is estimated.

6.5.3 Router Maintenance Cost

Cisco offers SMARTnet service [83] that provides technical support services on 24 hours a day and 365 days a year. The price for the service is based on product model and support service types. In this study, price of this service is used for the cost of router maintenance (RM).

Table 6.4: Price of Router Maintenance Service

Types	*Models*	*Cost of RM / year*
Access Router (AR)	Cisco 7603	\$5K
Distribution Router (DR)	Cisco 12816	\$6K
Backbone Router (BB)	Cisco CSR-1	\$65K

Table 6.4 shows list of RM cost inferred from [84]. RM cost ratio of router types and maintenance fee of new routers are estimated as follows:

$$[BB_{mt} : DR_{mt} : AR_{mt}] = [13.00 : 1.20 : 1.00] \tag{6.22}$$

6.5.4 Human Resource Cost

Human resource (HR) cost for the transition is based on number of network administrator. The salary of network administrator can be categorized in entry,

intermediate, and senior level. In this study, levels of network administrator for access routers (AR), distribution routers (DR), and backbone routers (BB) are referred to entry, intermediate, and senior level.

Table 6.5: Salary of Network Administrator in the US

Types	*Level*	*Average Salary / year*
Access Router (AR)	Entry	\$56K
Distribution Router (DR)	Intermediate	\$67K
Backbone Router (BB)	Senior	\$82K

Table 6.5 shows average salary of network administrator in the US [85], and HR cost ratios of router types are estimated as follows:

$$[BB_{wage} : DR_{wage} : AR_{wage}] = [1.46 : 1.20 : 1.00] \tag{6.23}$$

Table 6.6: Employment Ratio of Occupations

Occupation (AT&T California)	*Employment Ratio* [86]
Managers	5.11%
Professionals	3.84%
Technicians	5.66%
Sales	3.28%
Admin&Support	38.59%
Craft Workers	43.52%
Occupation (Technicians in ISPs)	*Employee Ratio* [87]
Programmer	20.84%
Application Software Engineer	18.93%
System Software Engineer	13.80%
System Analyst	22.92%
Database Administrator	4.87%
Network System Administrator	11.49%
Network Data Analyst	7.15%

Employee ratios of occupations are presented in Table 6.6. In AT&T California, 5.66% of total employee are technicians in ISPs [86] and employee ratios

within technical occupation are shown in bottom of Table 6.6 and 11.49% of total technical employee are network administrator. Thus, an employee ratio of network administrator is estimated as 0.65%. Total number of employee of AT&T in 2010 is 282,720. Therefore, we estimated number of network administrator in AT&T is 1,838 in this study. Since total number of routers in AT&T is 11,403, an average number of routers taken cared by an administrator is 5.87 (routers/administrator).

To estimate how many routers are taken cared by a network administrator ($BB_{care}, DR_{care}, AR_{care}$), we use ratio of router maintenance cost between BB, DR, and AR. From Equation 6.22, number of routers in AT&T (BB_{old}: 389, DR_{old}: 6395, AR_{old}: 4619, Total 11,403 routers), and total number of estimated network administrator in AT&T which is 1,838, numbers of each level network administrators and number of taken cared by an administrator are estimated based in RM cost ratio as shown in Table 6.7.

Table 6.7: Estimated Network Administrators in AT&T

BB_{old}	DR_{old}	AR_{old}	*Total*
389	6,395	4,619	11,403
Senior	*Intermediate*	*Entry*	*Toral*
532.81	808.53	486.66	1,838.00
BB_{care}	DR_{care}	AR_{care}	*Average*
0.78	10.49	6.33	5.87

6.5.5 Power Usage Cost

In this study, power usage (PU) cost will be determined by types of router, usage of TCAM chips, and routing complexity.

Table 6.8 shows power consumption of each router type. Power consumption of cooling system and Uninterruptible Power Supply (UPS) are not in-

Table 6.8: Power Consumption of Router Types. Source: [80, 88]

Types	*Models*	*Power Consumption / hour*
Access Router (AR)	Cisco 7603	1.0KW
Distribution Router (DR)	Cisco 12816	6.0KW
Backbone Router (BB)	Cisco CSR-1	10.0KW

cluded in this data. Usage of TCAM chip and routing complexity are determined by a location and routing table size of a router in an ISP.

Table 6.9: Breakdown of Power Consumption by a Router. Source: [80]

Parts	*% of Total Power*
Supply loss and blowers	35.0
Forwarding engine	33.5
Switching Fabric	10.0
Control plane	11.0
I/O	7.0
Buffers	3.5
Total	100

Percentage of energy consumption in a router is displayed in Table 6.9. Based on the table, we assume that power consumption of "Forwarding engine" and "Control plane" are related to routing protocol, thus 44.5% of total power consumption can be estimated to PU of routing protocol, and power consumption of "I/O" is related to memory usage, which is 7.0%.

$$BB_{new\_pu} = BB_{old\_pu} - \Delta BB_{routing} - \Delta BB_{memory} \tag{6.24}$$

$$DR_{new\_pu} = DR_{old\_pu} - \Delta DR_{routing} - \Delta DR_{memory} \tag{6.25}$$

$$AR_{new\_pu} = AR_{old\_pu} - \Delta AR_{routing} - \Delta AR_{memory} \tag{6.26}$$

$$\Delta BB_{routing} = BB_{old\_pu} \times 0.445 \times C_{ratio} \tag{6.27}$$

$$\Delta BB_{memory} = BB_{old\_pu} \times 0.07 \times RT_{ratio} \tag{6.28}$$

$$[C_{ratio} < 1, RT_{ratio} < 1]$$

where $\Delta BB_{routing}$ and ΔBB_{memory} are reduced power usages by routing complexity and memory usage, and RT_{ratio} is a ratio of number of bits used in TCAM memory. Power consumption of TCAM chip is proportional to number of bits used in TCAM, therefore, a value of RT_{ratio} is used to estimate power usage of "I/O", which is 7.0% of total power consumption of a router. The same calculation is applied for $\Delta DR_{routing}$, ΔDR_{memory}, $\Delta AR_{routing}$ and ΔAR_{memory}.

6.5.6 Operating Cost Estimation

The first scenario compares operating cost of before and after transition. To compare the operating cost, all costs are estimated by ratio based on cost of AR.

Table 6.10 shows ratio of each cost based on cost of AR and ratio is estimated by Tables 6.3, 6.4, 6.5, and 6.8. Ratio of router price is used for threshold of degrading routers from BB to DR and DR to AR which are 4.55

Table 6.10: Ratio of Costs(AR based)

	BB/AR	*DR/AR*	*BB/DR*	*AR*
Price	10.00	2.20	4.55	1.00
RM	13.00	1.20	10.83	1.00
HR	1.46	1.20	1.22	1.00
PU	10.00	6.00	1.67	1.00

and 2.20. Thus, condition of regrading router (in terms of router hardware level) is as follows:

	Complexity Ratio(OSPF/TRP)	*Threshold*
BD: BB -> DR	$C_{ratio}(x) = \frac{O(R(x) \times \log(R(x)))}{O(N(x))} >$	4.55, $x \in BB$
DA: DR -> AR		2.20, $x \in DR$

where BD is a router degrade from BB to DR and DA is a router degraded from DR to AR. Routing complexity C_{ratio} is a complexity of updating routing table at router x in AT&T network. Where R is number of OSPF router in the routing domain of router x, and N is number of direct neighbor of TRP router x.

Table 6.11: Total Number of Routers before and after Transition

	BB	*DR*	*AR*	*Total*
Before	389	6395	4619	11403
After	1	399	11003	11403

Figure 6.2 show distribution of BB, DR, and AR router in each POP (total 110 POPs) of AT&T network. For example, Chicago POP has total 1010 routers that 26 BBs, 398 DRs, and 586 ARs. The details of other POP is presented in Appendix C. Based of the condition of degrading router, we identified degrading routers and the result is displayed in Table 6.11 and Figure 6.3. Ta-

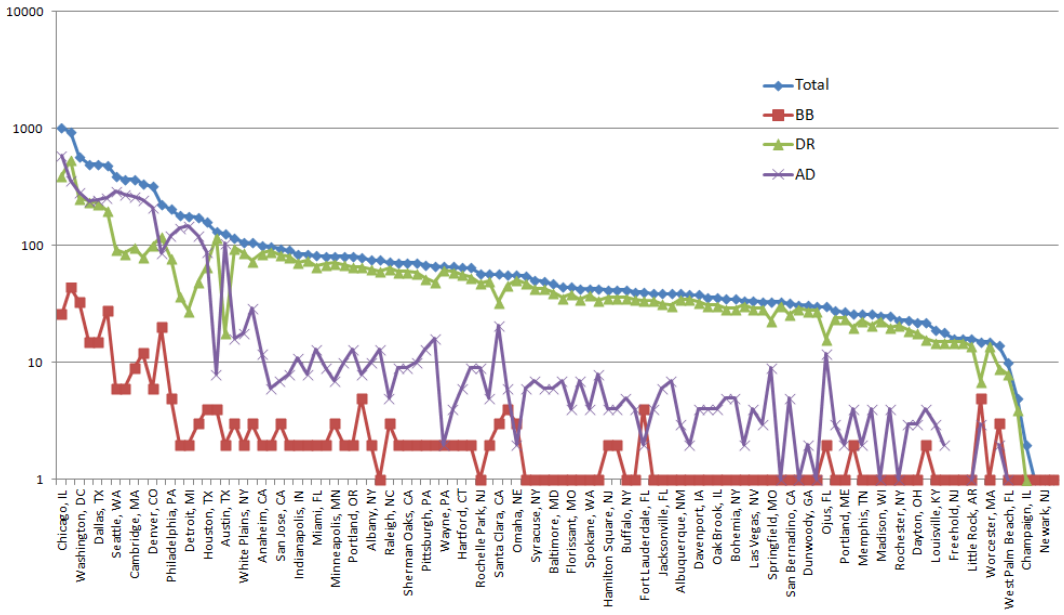


Figure 6.2: Distribution of BB, DR, and AR routers in Each POP (AT&T)

ble 6.11 shows total number of each router types before and after transition, and Figure 6.3 show number of each router types on each POP after transition.

Router Maintenance Cost Estimation

After identified number of degrade routers, router maintenance cost (RM) are estimated based on cost of AR type. Ratio of RM cost between BB, DR, and AR is 13.00 : 1.20 : 1.00, therefore, RM cost estimation is:

$$RM_{old} = BB_{old} \times 13.00 + DR_{old} \times 1.20 + AR_{old} \times 1.00 \quad (6.29)$$

$$RM_{new} = (BB_{old} - BD) \times 13.00 + (DR_{old} - DA + BD) \times 1.20 + (AR_{old} + DA) \times 1.00 \quad (6.30)$$

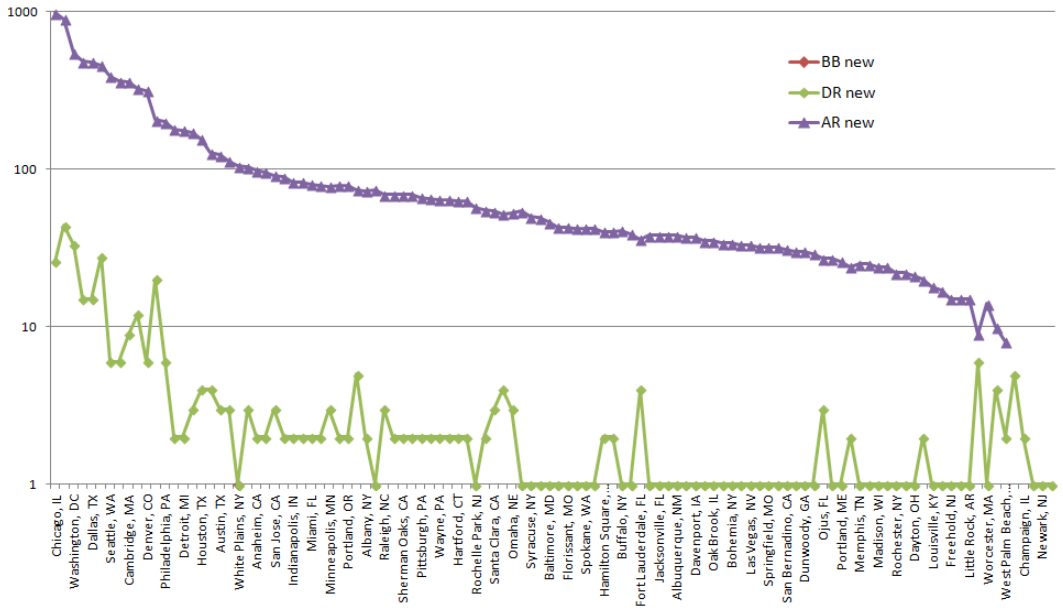


Figure 6.3: Distribution of BB, DR, and AR routers in Each POP after Transition (AT&T)

Figure 6.4 shows RM cost ratio of each POP after the transition. For example, RM cost of Chicago, IL POP is 72.43% of original RM cost (before transition). Total RM cost ratio in AT&T network is 66.25%.

Human Resource Cost Estimation

Ratio of HR cost between BB, DR, and AR is $1.46 : 1.20 : 1.00$, therefore, HR cost estimation is:

$$\begin{aligned} HR_{old} &= \lceil BB_{old}/BB_{care} \rceil \times 1.46 + \lceil DR_{old}/DR_{care} \rceil \times 1.20 \\ &\quad + \lceil AR_{old}/AR_{care} \rceil \times 1.00 \qquad (6.31) \\ HR_{new} &= \lceil (BB_{old} - BD)/BB_{care} \rceil \times 1.46 \\ &\quad + \lceil (DR_{old} - DA + BD)/DR_{care} \rceil \times 1.20 \\ &\quad + \lceil (AR_{old} + DA)/AR_{care} \rceil \times 1.00 \qquad (6.32) \end{aligned}$$

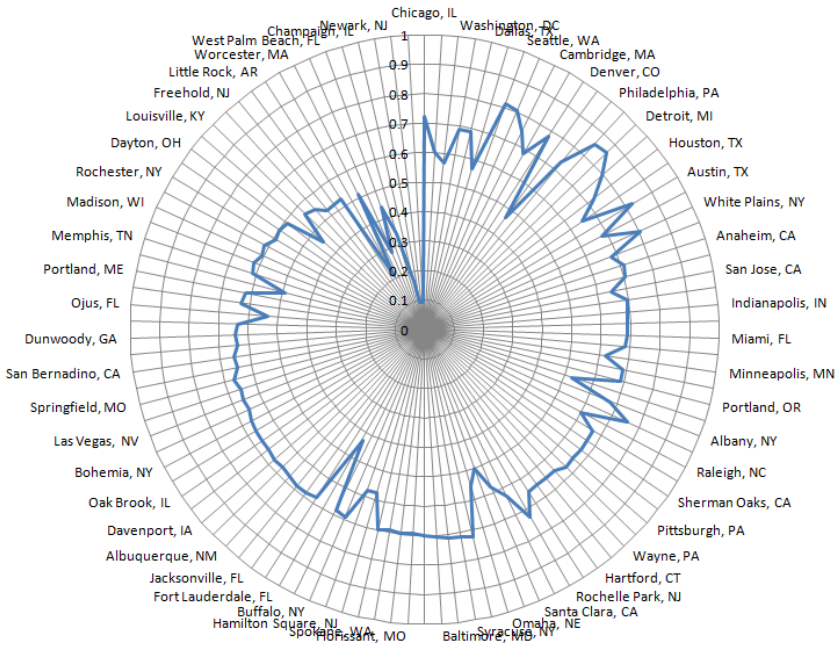


Figure 6.4: Cost Ratio of RM before and after Transition(POP)

Figure 6.5 shows HR cost ratio of each POP after the transition. For example, HR cost of Chicago, IL POP is 84.79% of original HR cost (before transition). Total HR cost ratio in AT&T network is 81.14%.

Power Usage Cost Estimation

Ratio of PU cost between BB, DR, and AR is 10.00 : 6.00 : 1.00, however, this PU cost is total PU of a router and routing protocol related power usage of a

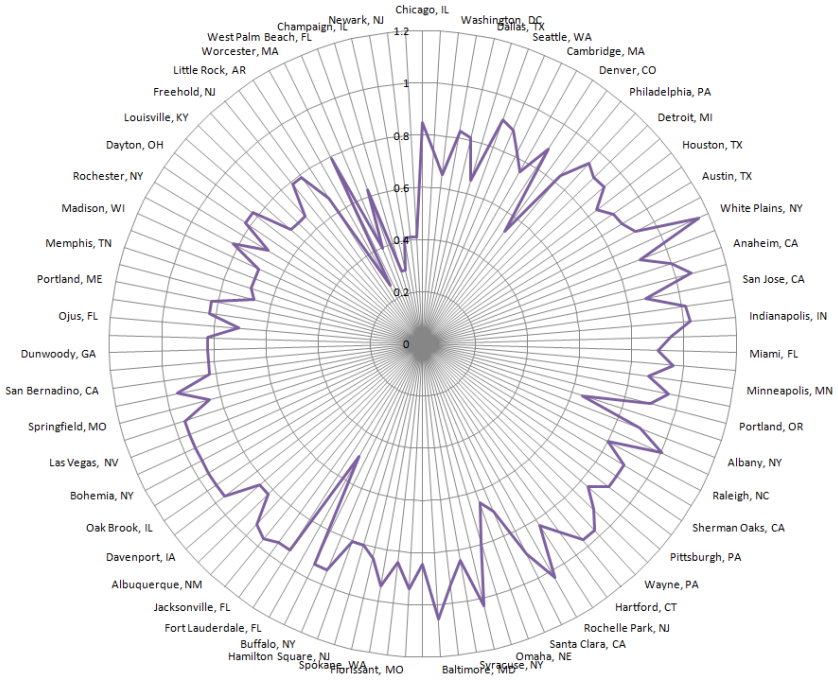


Figure 6.5: Cost Ratio of HR before and after Transition(POP)

router is 44.5% and 0.07 % for memory usage. Thus, PU cost estimation is:

$$PU_{old} = (BB_{old} \times 10.00) + (DR_{old} \times 6.00) + (AR_{old} \times 1.00) \tag{6.33}$$

$$\begin{aligned} PU_{new} = & \sum_{bb \in BB_{new}} (PU_{else} + PU_{rt} \times C_{ratio}(bb) + PU_{mem} \times RT_{ratio}(bb)) + \\ & \sum_{dr \in DR_{new}} (PU_{else} + PU_{rt} \times C_{ratio}(dr) + PU_{mem} \times RT_{ratio}(dr)) + \\ & \sum_{ar \in AR_{new}} (PU_{else} + PU_{rt} \times C_{ratio}(ar) + PU_{mem} \times RT_{ratio}(ar)) \end{aligned} \tag{6.34}$$

$$PU_{else}, PU_{rt}, PU_{mem} = \begin{cases} 10.0 \times 0.485, 10.0 \times 0.445, 10.0 \times 0.07 & BB \\ 6.0 \times 0.485, 6.0 \times 0.445, 6.0 \times 0.07 & DR \\ 1.0 \times 0.485, 1.0 \times 0.445, 1.0 \times 0.07 & AR \end{cases}$$

$$C_{ratio}(x) = \frac{N(x)}{R(x) \times \log R(x)} \tag{6.35}$$

$$RT_{ratio}(x) = \frac{N(x) \times 12bits}{L(x) \times 32bits} \tag{6.36}$$

$$L(x) = \begin{cases} all\ links\ in\ AT\&T & (x = BB : OSPFarea0) \\ all\ links\ in\ a\ POPX & (x = DR, AR : OSPFareaX) \end{cases}$$

where N is number of direct neighbor of router x which means routing table size of TRP router x. R is number of OSPF router and L is number of links in router x's OSPF routing domain, and L is also routing table size of OSPF router x. In this estimation, TCAM chip power usage is proportional to number of bits used in the TCAM, 12 bits and 32 bits length are used for an entry in routing table of TRP and OSPF.

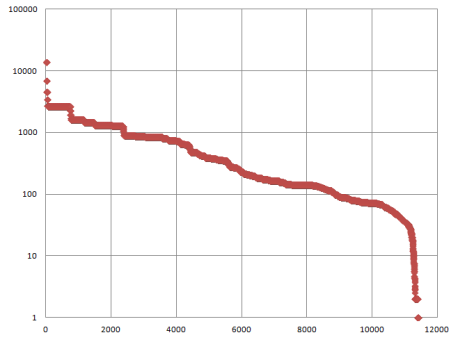


Figure 6.6: Routing Table Ratio

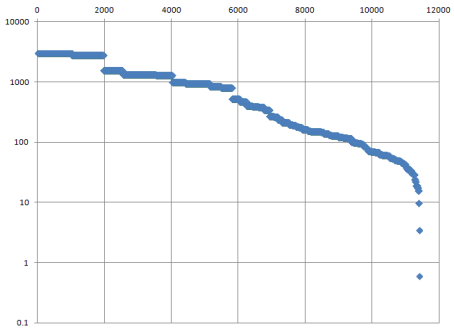


Figure 6.7: Complexity Ratio

Figures 6.6 and 6.7 show routing table size ratio and complexity ratio of between OSPF and TRP routers in AT&T network. TRP has significantly small size of routing table and less complexity compared to OSPF routing in AT&T.

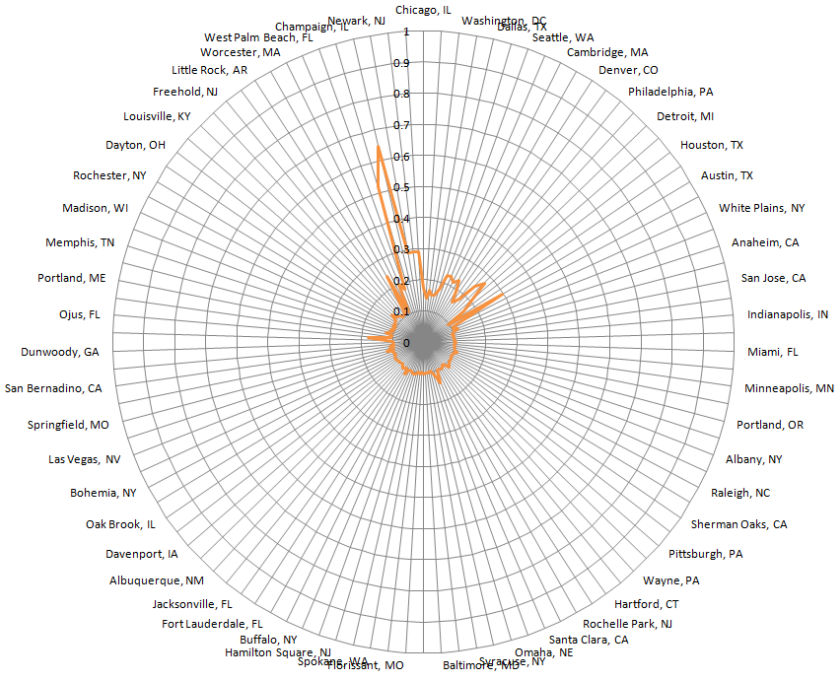


Figure 6.8: Cost Ratio of PU before and after Transition(POP)

Figure 6.8 shows PU cost ratio of each POP after the transition. For example, PU cost of Chicago, IL POP is 17.15% of original routing related PU cost (before transition). Total PU cost ratio in AT&T network is 14.02%.

Total Operating Cost Estimation

Table 6.12: Total Estimated Operating Cost and Reduction

Original Costs	RM cost	HR cost	PU cost	Total
BB	\$2,529K	\$6,069K	\$2,045K	
DR	\$38,370K	\$53,721K	\$20,167K	
AR	\$23,095K	\$53,882K	\$2,428K	
Total (before)	\$86,750K	\$113,672K	\$24,649K	\$225,062K
Cost Ratio	66.25%	81.14%	14.02%	
Total (after)	\$57,472K	\$92,234K	\$3,454K	\$153,160K

Operating cost after transition of RM, HR, and PU are 66.25%, 81.14%, and 14.02%. Table 6.12 shows estimated yearly cost of RM, HR, and PU($60/MWh) of before and after the transition. As a result, when $OC_{total}(t_{start})$ is 100%, operating cost after transition $OC_{total}(t_{end})$ is estimated to 68.05%, and total $71.9M is saved.

6.5.7 Transition Cost Estimation

To estimate operating cost during the transition, the following scenario is applied:

- Transition process unit: POP base (each POP works independently)
- Time unit: Week base
- Router replacement order: BB routers -> DR routers -> AR routers, and number of hops to BB router base
- HR allocation: POP base
- Number of replacement routers: Up to 50 routers @ POP / week
- Backup routers: Keeps old routers for 2 weeks after replaced

Since TRA allocation starts from the top tier in FCT architecture, transition scenario follows top-to-bottom steps in a network topology. Therefore, router replacement starts from BB routers in each POP. Figure 6.9 shows the concept of the transition in a POP. Each POP works independently, which means all POPs in AT&T network process the transition simultaneously. Operating cost during the transition is calculated a week by a week, and up to 50 routers can be processed per a week. For the risk management purpose,

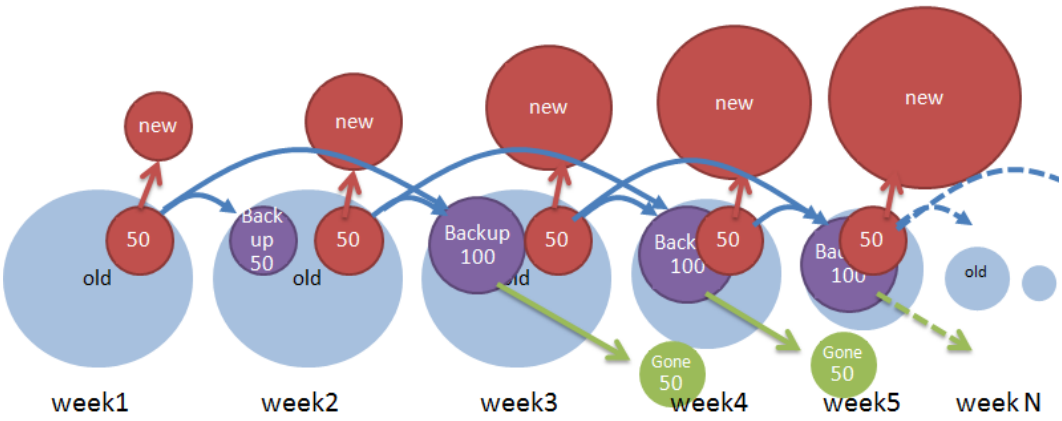


Figure 6.9: Transition Scenario in a POP

each replaced router will be kept for 2 weeks, then removed from a network. After all BB routers are replaced in a POP, DR routers will be replaced. The replace order of DR and AR routers in POP is based on the topology in each POP. The shortest path to BB router is identified on each DR and AR routers, and routers have shorter hops to BB are replaced first. So, DR and AR router replacement is processed a hop by a hop base. Number and location of BB, DR, and AR routers in a POP of AT&T network is already determined by previous study and will be used it for this estimation scenario.

Number of Routers during Transition

Figure 6.10 shows result of router type change during the transition. All old routers are replaced to new routers at **24th week**. Number of old routers are do not change in the first 3 weeks because old routers will be kept for 2 weeks after replaced. Thus, total number of routers are increased significantly in the first 3 weeks. After the first 3 weeks, number of old routers are started to

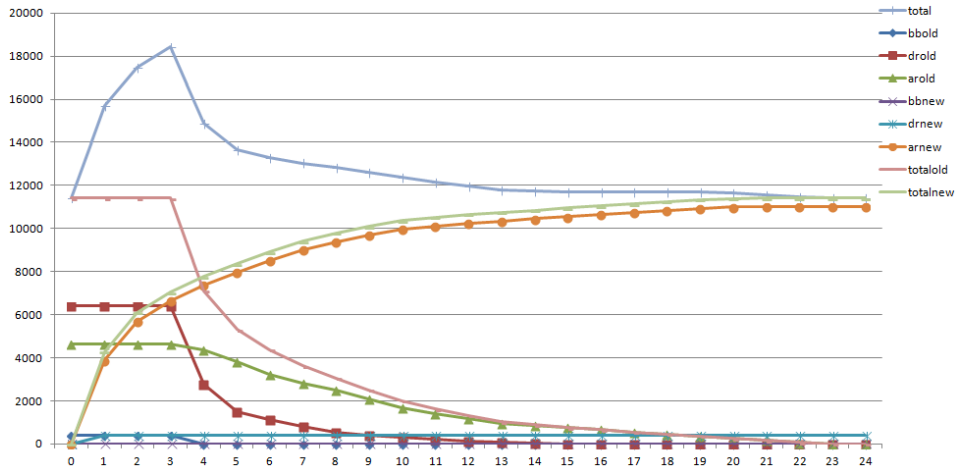


Figure 6.10: Number of Each Types of Routers during Transition

decrease. Due to the router type degrading, most of old routers are replaced with AR router.

Router Maintenance (RM) Cost during Transition

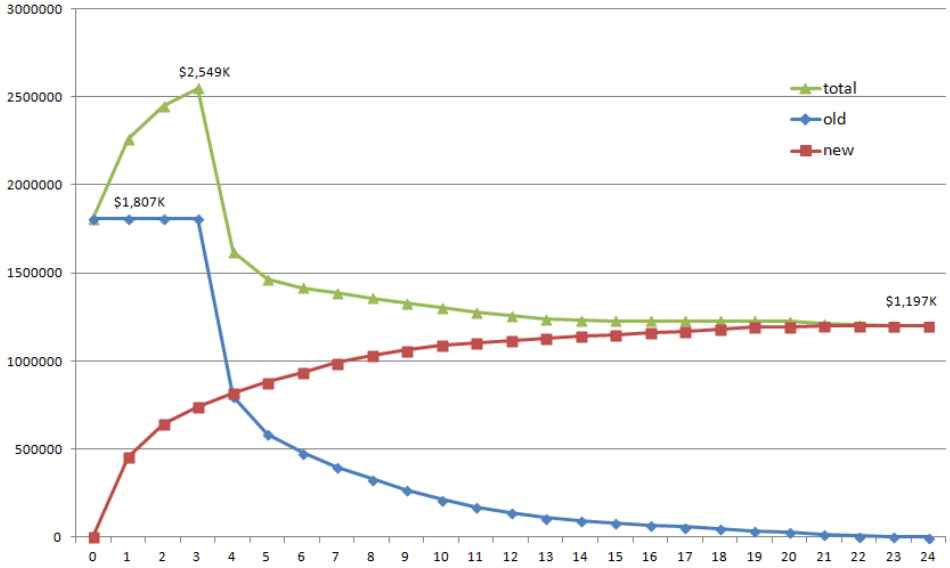


Figure 6.11: RM Cost ($) during Transition

RM cost during the transition is shown in Figure 6.11. Weekly RM cost before the transition (week 0) is $1,807K and RM cost after the transition (week 24) is $1,197K, and the peak RM cost is $2,549K on week 3 because number of router is peaked at week 3.

Human Resource (HR) Cost during Transition

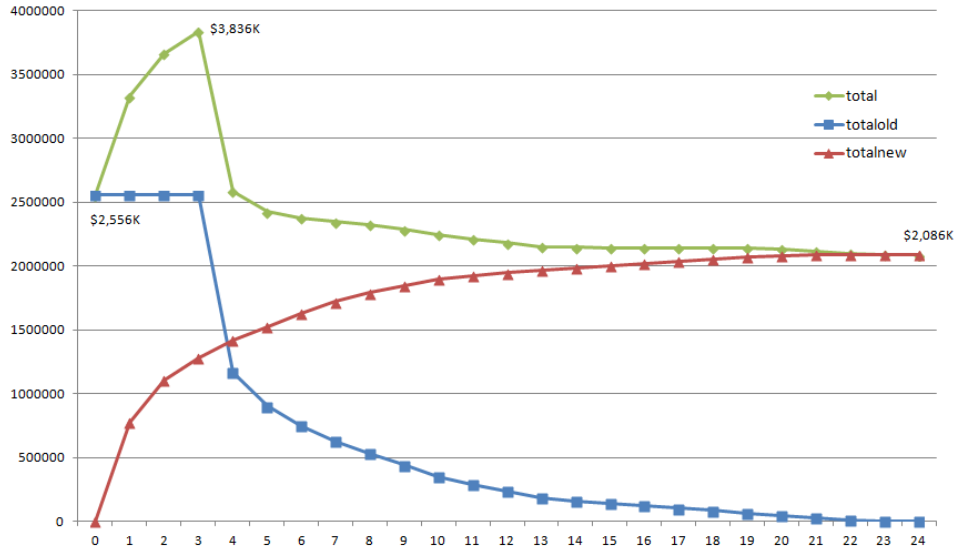


Figure 6.12: HR Cost ($) during Transition

Figure 6.12 shows HR cost of the transition period. Weekly HR cost before the transition (week 0) is $2,556K and HR cost after the transition (week 24) is $2,086K, and the peak HR cost is $3,836K on week 3. The cost trend is very similar to RM cost both cost is based on number of routers.

Power Usage (PU) Cost during Transition

PU cost during the transition is shown in Figure 6.13. Weekly PU cost before the transition (week 0) is $473K and PU cost after the transition (week 24)

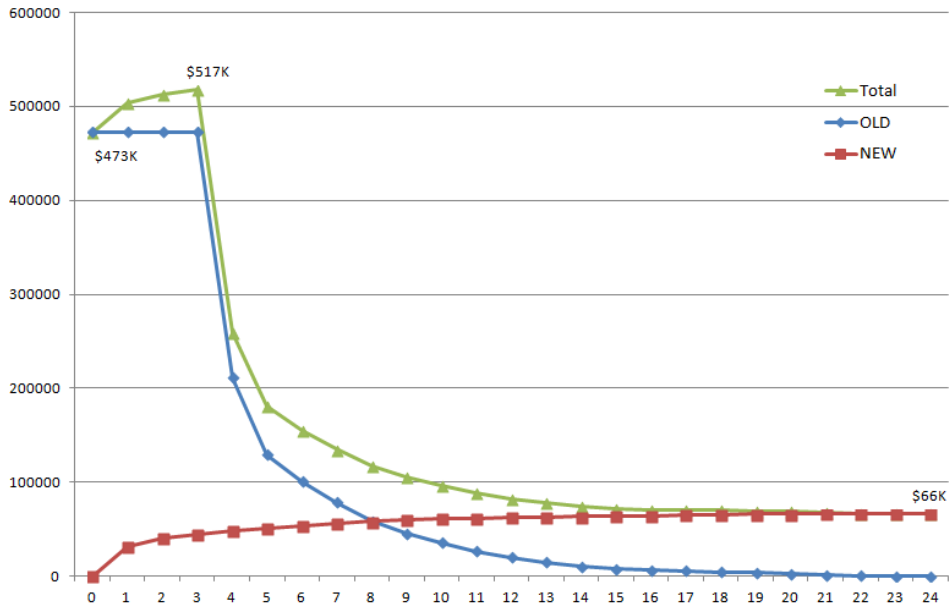


Figure 6.13: PU Cost ($) during Transition

is $66K, and the peak PU cost is $517K on week 3. Most of old BB and DR routers are replaced to new AR routers in the first few weeks, PU cost is reduced a lot by week 5.

Total Operating Cost during Transition

Total OC cost and cost of RM, HR, and PU during the transition period is presented in Figure 6.14. Total weekly operating cost before the transition is $4,835K and total weekly operating cost after the transition is estimated as $3,350K, which means $1,475K can be saved every week.

Investment Cost during Transition

Total prices of all new router is estimated as $118,611K and 5% savage value is applied to estimate total investment cost that is $112,963K. Investment cost during the transition period is shown in Figure 6.15.

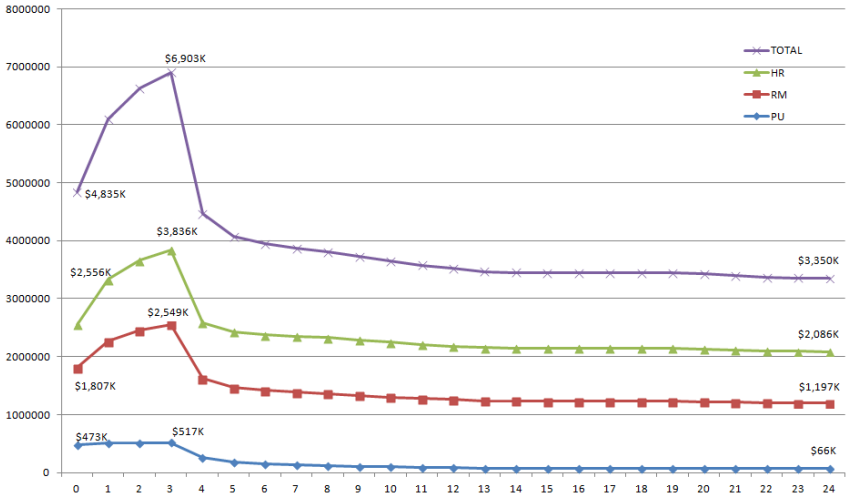


Figure 6.14: Total OC Cost ($) during Transition

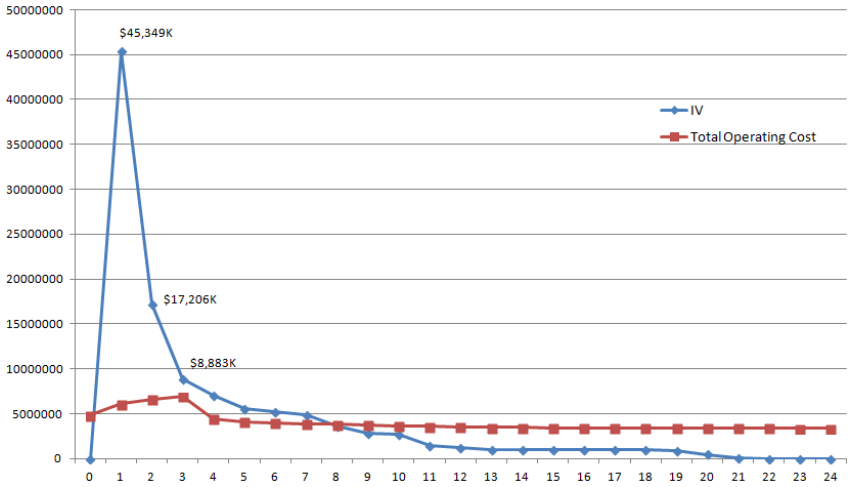


Figure 6.15: Investment Cost ($) during Transition

Total Payout Period

Finally, total payout period is estimated based on payback of operating cost. Figure 6.16 shows cost cash flow until all transition costs are returned. At 12th week, operating cost cash flow is turned to positive and after the week, $1,467K is returned every week. At 87th week, all transition costs are returned.

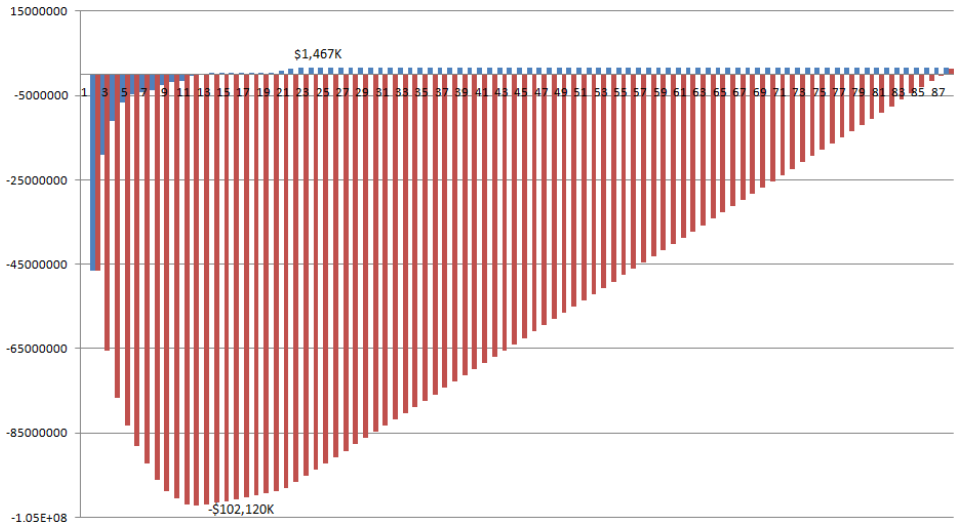


Figure 6.16: Estimated Total Payout Period

6.6 Limitations

This work is an initial step in studying the adoption of new routing protocol from an economic and ISPs standpoint. The parameter values used in the analytical model do not directly map on real world numbers because it is very difficult to accurately quantify the benefits provided by different routing protocols. We attempt only to draw general conclusion of relative importance of the model parameters, thus the model of cost estimation is quite basic.

Chapter 7

Conclusion

A Tiered Routing protocol (TRP) was developed under a new tiered Internet architecture called Floating Cloud Internet (FCT) architecture. The tiered routing addresses (TRA) in this architecture are used by TRP for packet forwarding. The TRP based forwarding can be used both for inter and intra-cloud forwarding. The network sizes do not constrain the use of the TRP model; In RIP the maximum network diameter can be 16, while OSPF handles the network size issue through the use of OSPF areas. With BGP the problem is complex and difficult to address due to the path vector routing and the use of policies. In this study, TRP is evaluated as both an IGP and an EGP. Initial convergence time and control overhead with networks running TRP is very low as the protocol does not require message flooding or any calculations subsequent to a link status change. Due to the inherent routing information in the tiered addresses, the routing table sizes in TRP are significantly low. Stability in the routing entries and their invariance to network size also indicates the strengths of such new approaches. Comparison with OSPF and BGP validates this.

With the ongoing future Internet initiatives sponsored and funded by research organizations all over the world, it is equally important to have a test and evaluation platform that mimics real world situations and operational conditions as closely as possible. Especially when new architectures are being investigated to replace completely the current architectures, experimentation testbeds play a significant role. This is because the current Internet is composed of ISPs who may not be willing to expose their network to testing and validating a new architecture before it is deployed. In the real world such evaluations could not be done without these emulation facilities. In the US, the National Science Foundation thus initiated the GENI project, which could be used for large scale emulation tests, one of them being the future Internet architectures. The project presented in this article proposed a novel Floating Cloud Tiered architecture, which required the explicit recognition and use of tiers existing in the AS worldwide topology and also within an AS, to design a tier-based addressing scheme and a tiered routing protocol, with the ultimate goal of overcoming current Internet address limitations and routing scalability. The tiered routing protocol replaces IP and all its routing protocols both for inter- and intra-domain purposes.

The Emulab testbed was used as the platform for evaluating this new protocol and comparing its performance with the more accepted intra-domain routing protocol OSPF running on IP. The testbed allowed incorporation of a number of accessory softwares, such as IPerf, modified version of SIPerf, a command line version of Wireshark called Tshark, Quagga software to run OSPF, and various other scripts to setup and run the different tests. The presented work has been limited to a comparison with OSPF only and for intra-domain routing, as the goal is to show the capabilities of GENI to evaluate

new protocols that can completely replace the existing protocol stacks, which is especially useful to test new Internet architectures.

In addition to the evaluation of TRA and TRP, economical impact on the transition from the current IP routing to the TRA/TRP routing is investigated by defining cost estimation model and estimating the transition cost. Showing economic viability and sustainability is more important for our study because our solution is taking revolutional approach and the economical study is one of the key component for the clean-slate solution of the future Internet architecture.

Based on the evaluation study of TRA and TRP, we validated that TRP is very simple protocol compared to the current routing protocols (OSPF and BGP) and TRP does not require high specs in router hardware. As the result, new routers, which support TRP, can be much cheaper, easier to maintain, and less power consumption than old routers that running OSPF and BGP. Especially, power consumption in TRP is remarkably small and it is important for ISPs because the energy efficiency issue has became a high-priority objective. Through the estimation of the transition cost, we show huge potential to reduce not only operating costs, and to contribute for energy-efficient network.

Perspectives

Novelty of the FCT architecture includes that TRP can be applied to any size of network which means single routing protocol can be used in both intra- and inter- domain routing, utilization of existing logical and topological tiers in the current network, and free from IP address by adopting new addressing TRA that is assigned to a cloud, not to an network interface. Significance of the FCT architecture is that high scalability, fast convergence, simple and loop

free packet forwarding, and less power consumption achieved by small routing table size and less routing complexity. Contributions of this study include definition of TRA addressing scheme and TRP routing protocol, implementation of software FCT router, evaluation of TRA and TRP, proposing transition scenarios, and developed cost estimation model for the transition.

Finally, a possible future work would be to cooperate with security. The proposed architecture is base of the future Internet architecture and it can easily work with many different security mechanism because of simplicity and scalability of the FCT. Another path for the future work would be to cooperate with network applications such as QoS, VoIP, and energy-aware network by adding properties to TRAs. Since a cloud can be multihomed and having different paths by having multiple TRAs, each TRA can have different priority for the applications.

Appendix A

Internet Topology

In this appendix, we show additional POP level topology of several ISPs and router level topology of AT&T network discussed in Chapter 4.1.1.

A.1 POP Level Topology of ISPs

Based on Rocketfuel datasets, several ISP topologies on the US map are visualized.

Steps of topology visualization:

1. Read and parse the dataset to get information of city name of each router location and connectivity
2. Obtain latitude and longitude coordinate of each city
3. Calculate relative XY coordinate on the US map size
4. Find connectivity between each city based on the router connectivity
5. Draw the lined based on the connectivity on the US map

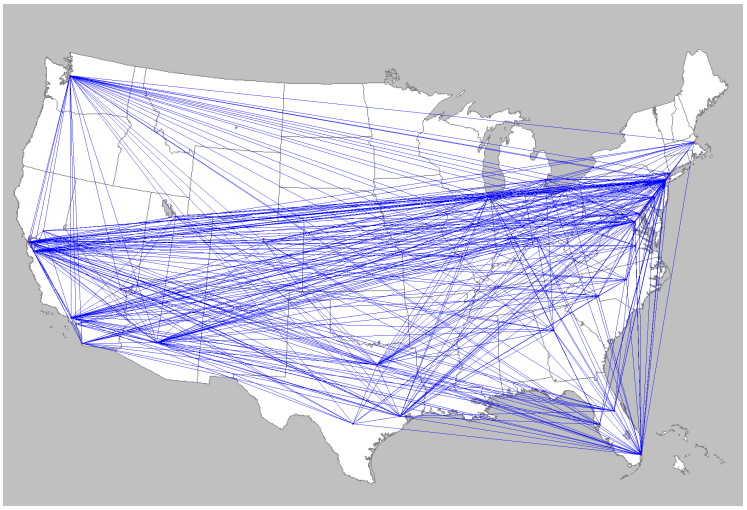


Figure A.1: Level3 POP Level Topology in the US

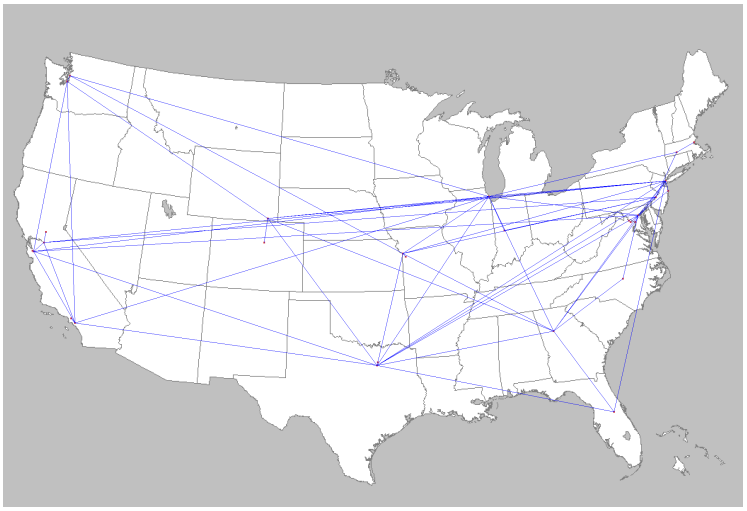


Figure A.2: Sprint POP Level Topology in the US

A.2 Router Level Topology of AT&T

To visualize router-level topology, Cytoscape tool is used. The layout of routers is set based on:

- Calculate relative XY coordinate on the US map size

- Calculate POP size based on number of routers in a POP
- Locate routers as circle that a center XY position from POP location and radius by size of a POP

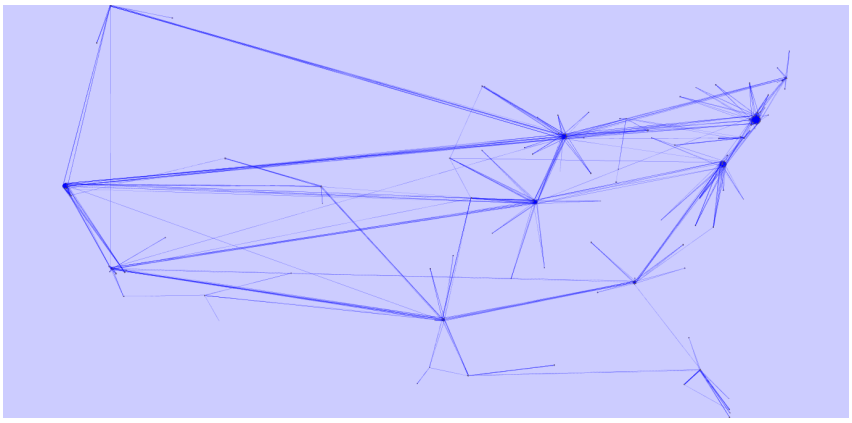


Figure A.3: Router Level topology of AT&T

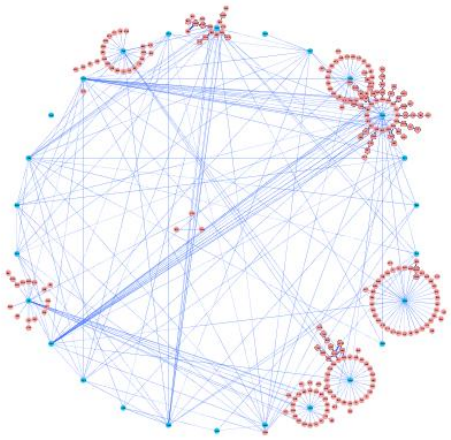


Figure A.4: Router Level topology of Chicago POP

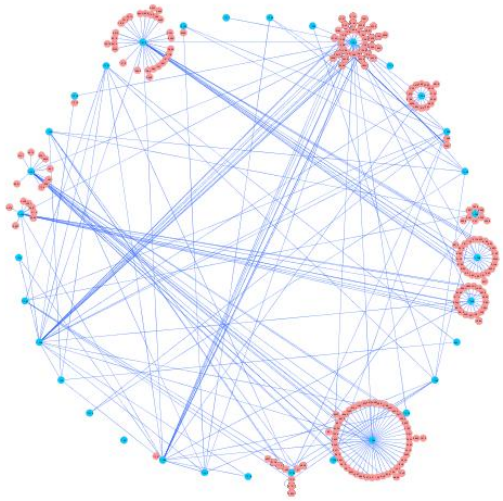


Figure A.5: Router Level topology of Washington D.C. POP

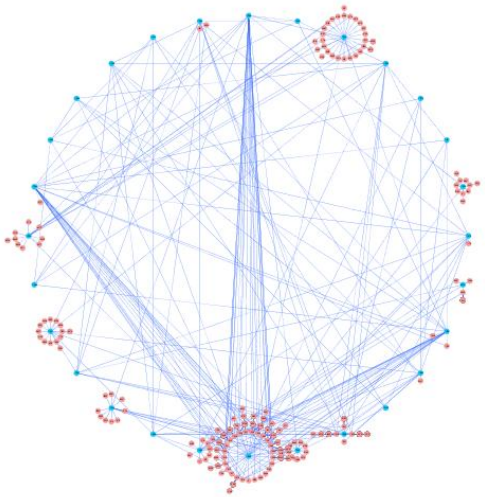
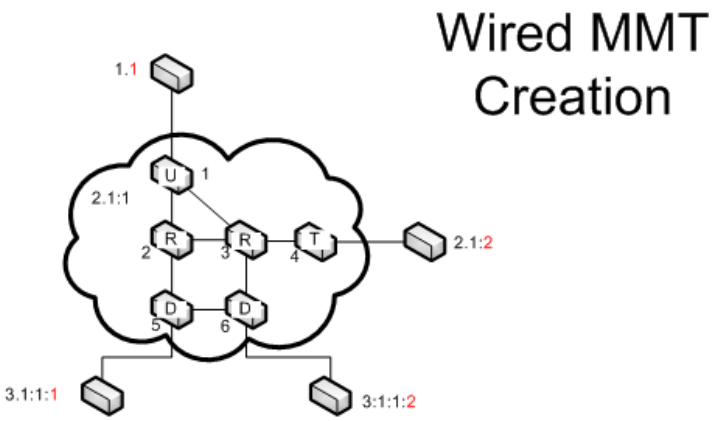


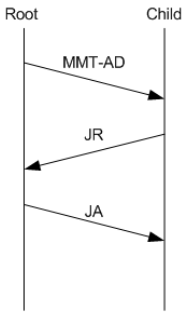
Figure A.6: Router Level topology of San Francisco POP

Appendix B

Multi-Meshed Tree (MMT)

MMT routing protocol uses similar addressing scheme with TRA. To transit between intra domain and inter domain routing, MMT can be used in the FCT. The followings explain address allocation of MMT.





MMT creation starts from root nodes which are Node 1(up) and Node 4(trunk)

1. Root advertises MMT-AD including own unique id as root vid
   ex) Up Root : root vid(1), Root Type(1), Neighbor SI(1)
   Trunk Root : root vid(4), Root Type(2), Neighbor SI(2)

2. Child receives MMT_AD and sends JR
   ex) Node 2 : requested vid(1), Node Type(0), UID(2), Neighbor SI(n/a)
   Node 3 : requested vid(1), Node Type(0), UID(3), Neighbor SI(n/a)
   Node 3 : requested vid(4), Node Type(0), UID(3), Neighbor SI(n/a)

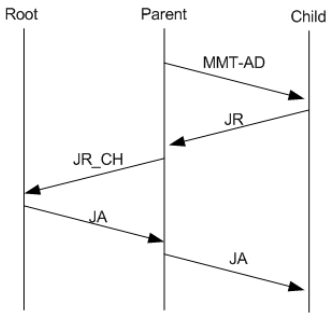
3. Root receives JR and replies JA with allocated new vid
   ex) Up Root : new vid(1:1) to Node 2
   : new vid(1:2) to Node 3
   Trunk Root : new vid(4:1) to Node 3

Root (1) maintains tree table

UID	VID	TYPE	NB
2	1:1	R	
3	1:2	R	

Root (4) maintains table

UID	VID	TYPE	NB
3	4:1	R	



Once node got vid, it starts sending advertisement packet

1. Parent advertises MMT-AD including all own vids
   ex) Node 2 : vid(1:1), Root Type(1), Neighbor SI(1)

2. Child receives MMT_AD and sends JR to Parent
   ex) Node 3 : requested vid(1:1), Node Type(0), UID(3), Neighbor SI(N/A)
   Node 5 : requested vid(1:1), Node Type(4), UID(5), Neighbor SI(1)

3. Parent receives JA and sends JA_CH with allocated new vid to root (can be multi-hops)
   ex) Node2 : new vid(1:1:1), Node Type(0), UID(3), Neighbor SI(N/A)
   : new vid(1:1:2), Node Type(4), UID(5), Neighbor SI(1)

4. Root receives JR_CH and accepts/updates tree table, then sends back JA to Parent
   ex) Up Root : new vid(1:1:1)
   : new vid(1:1:2)
5. Parent sends JA to child
   ex) Node 2 : new vid(1:1:1) to Node 3
   : new vid(1:1:2) to Node 5

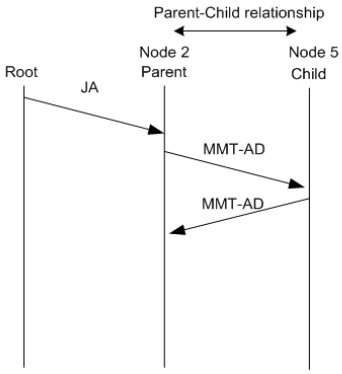
Root maintains tree table

UID	VID	TYPE	NB
2	1:1	R	
3	1:2 1:1:1	R	
5	1:1:2	D	1

* always loops in vid are avoided by child before sending JA packet
** parent always use the same last digit to the same child for new vid allocation

Once Parent-Child relationship has created, Child can join new vid without JR

Case1) When Node 2 got new vid(1:2:1) from Node 3,



0: Up Root update its tree table

UID	VID	TYPE	NB
2	1:1 1:2:1	R	
3	1:2 1:1:1	R	
5	1:1:2	D	1

→

UID	VID	TYPE	NB
2	1:1 1:2:1	R	
3	1:2 1:1:1	R	
5	1:1:2 1:2:1:2	D	1

Up Root gets JA, then insert Node2's new vid (1:2:1) into own tree table. Then, checks Parent-Child relationship which is Node 2 as Parent. Find out that Node 3 and 5 are child of Node2. For the new vid (1:2:1) can be parent of Node 5, not Node 3 because of loop in vid.
Therefore, new vid(1:2:1:2) is added to Node 5 based on last digit of Parent-Child vid which is "2"

1. Parent receives JA and got new vid, then advertise updated MMT-AD to all neighbors
   ex) Node 2 : vid(1:1), Root Type(1), Neighbor SI(1)
   : new vid(1:2:1), Root Type(1), Neighbor SI(1)

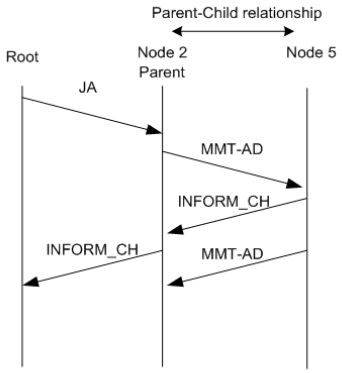
2. Child receives updated MMT_AD from parent and founds new vid
   ex) Node 5 : found new parent vid(1:2:1). Node 5 already has vid(1:1:2) from the parent. The last digit of the vid is "2". Then, child joins to the parent new vid(1:2:1) without JoinRequest process, the new vid is 1:2:1:2 which is the last digit "2" is appended.

3. Child added new vid and advertises updated MMT-AD
   ex) Node 5 : vid(1:1:2), Root Type(1), Neighbor SI(1)
   : new vid(1:2:1:2), Root Type(1), Neighbor SI(1)

4. Parent receives update MMT-AD and confirm that the child got new vid

Once Parent-Child relationship has created, Child can join new vid without JR

Case2) When Node 2 got new vid(4:1:1) which is different Root from Node 3,



0: Trunk Root update its tree table

UID	VID	TYPE	NB
2	4:1:1	R	
3	4:1	R	

Although Node 2 and 5 have already made Parent-Child relationship at the time, Node 5 cannot join the new vid(4:1:1) because Node 5 does not get have Trunk Root vid from the parent yet. Therefore, Trunk Root does not know the Parent-Child relationship between Node2 and 5.

1. Parent receives JA and got new vid, then advertise updated MMT-AD to all neighbors
   ex) Node 2 : vid(1:1), Root Type(1), Neighbor SI(1)
   : vid(1:2:1), Root Type(1), Neighbor SI(1)
   : new vid(4:1:1), Root Type(2), Neighbor SI(2)

2. Child receives updated MMT_AD from parent and founds new vid from different Root
   ex) Node 5 : found new parent vid(4:1:1). Node 5 already has vid(1:1:2) from the parent. The last digit of the vid is "2". Then, child joins to the parent new vid(4:1:1) without JoinRequest process, the new vid is 4:1:1:2 which is the last digit "2" is appended.

3. Child added new vid and Inform it to new Root by sending INFORM_CH
   ex) Node 5 : new vid(4:1:1:2), Node Type(4), UID(5), Neighbor SI(1)

4. Root receives INFORM_CH and accepts/updates tree table

UID	VID	TYPE	NB
2	4:1:1	R	
3	4:1	R	
5	4:1:1:2	D	1

5. Child sends updated MMT_AD to neighbors
   ex) Node 5 : vid(1:1:2), Root Type(1), Neighbor SI(1)
   : vid(1:2:1:2), Root Type(1), Neighbor SI(1)
   : new vid(4:1:1:2), Root Type(2), Neighbor SI(2)

Appendix C

Router Statistics of AT&T

POP	Total	BB	DR	AR	POP	Total	BB	DR	AR
Chicago, IL	1010	26	398	586	Birmingham, AL	44	1	36	7
New York, NY	946	44	542	360	Florissant, MO	44	1	39	4
Washington, DC	576	33	257	286	Richmond, VA	43	1	34	8
Atlanta, GA	499	15	241	243	Spokane, WA	43	1	38	4
Dallas, TX	495	15	231	249	Tulsa, OK	43	1	35	7
San Francisco, CA	485	28	200	257	Buffalo, NY	42	1	36	5
Seattle, WA	393	6	93	294	Greensboro, NC	42	2	36	4
Cambridge, MA	368	9	97	262	Hamilton Square, NJ	42	2	36	4
Orlando, FL	368	6	86	276	Fort Lauderdale, FL	40	4	34	2
Los Angeles, CA	337	12	80	245	Plymouth, MI	40	1	35	4
Denver, CO	321	6	101	214	Albuquerque, NM	39	1	35	3
St Louis, MO	226	20	120	86	Jacksonville, FL	39	1	32	6
Philadelphia, PA	205	5	78	122	Oakland, CA	39	1	34	4
Phoenix, AZ	181	2	37	142	Providence, RI	39	1	31	7
Detroit, MI	178	2	28	148	Columbia, SC	38	1	35	2
San Diego, CA	174	3	49	122	Davenport, IA	38	1	33	4

Houston, TX	159	4	67	88	Oak Brook, IL	36	1	31	4
Cleveland, OH	131	4	119	8	Stamford, CT	36	1	31	4
Austin, TX	126	2	18	106	Bohemia, NY	35	1	29	5
New Brunswick, NJ	115	3	96	16	South Bend, IN	35	1	29	5
White Plains, NY	107	2	87	18	Grand Rapids, MI	34	1	31	2
Salt Lake City, UT	106	3	74	29	Las Vegas, NV	34	1	29	4
Anaheim, CA	100	2	86	12	Gardena, CA	33	1	29	3
Arlington, VA	98	2	90	6	Springfield, MO	33	1	23	9
San Jose, CA	94	3	84	7	St. Paul, MN	33	1	31	1
Charlotte, NC	91	2	81	8	San Bernadino, CA	32	1	26	5
Cedar Knolls, NJ	85	2	75	8	Des Moines, IA	31	1	29	1
Indianapolis, IN	85	2	72	11	Dunwoody, GA	31	1	28	2
Miami, FL	82	2	67	13	Ojus, FL	30	2	16	12
Milwaukee, WI	81	2	69	10	San Antonio, TX	30	1	28	1
Minneapolis, MN	81	3	71	7	Bridgeport, CT	28	1	24	3
Portland, OR	81	2	66	13	Portland, ME	27	1	24	2
Riverside, CA	81	2	70	9	Camden, NJ	26	1	21	4
Kansas City, MO	79	5	66	8	Fort Worth, TX	26	2	20	4
Albany, NY	75	2	63	10	Memphis, TN	26	1	23	2
Framingham, MA	75	1	61	13	Madison, WI	25	1	23	1
Raleigh, NC	72	3	64	5	Manchester, NH	25	1	20	4
New Orleans, LA	71	2	60	9	Norfolk, VA	23	1	19	3
Rolling Meadows, IL	71	2	59	10	Rochester, NY	23	1	21	1
Sherman Oaks, CA	71	2	60	9	Colorado Springs, CO	22	2	16	4
Pittsburgh, PA	68	2	53	13	Dayton, OH	22	1	18	3
Harrisburg, PA	67	2	49	16	Louisville, KY	19	1	15	3
Nashville, TN	66	2	60	4	Brookhaven, MI	18	1	15	2

Wayne, PA	66	2	62	2	Akron, OH	16	1	15	0
Hartford, CT	65	2	57	6	Freehold, NJ	16	1	15	0
Oklahoma City, OK	65	2	54	9	Little Rock, AR	16	1	14	1
Rochelle Park, NJ	58	1	48	9	Madison Heights, VA	15	5	7	3
Galva, IL	57	2	50	5	Worcester, MA	15	1	14	0
Santa Clara, CA	57	3	33	21	Bridgeton, MO	14	3	9	2
Omaha, NE	56	3	51	2	West Palm Beach, FL	10	1	8	1
Tampa, FL	56	4	46	6	Abingdon, VA	5	1	4	0
Silver Springs, MD	55	1	48	6	Champaign, IL	2	1	1	0
Syracuse, NY	51	1	43	7	Newark, NJ	1	1	0	0
Cincinnati, OH	50	1	43	6	Palo Alto, CA	1	1	0	0
Baltimore, MD	47	1	40	6	Tucson, AZ	1	1	0	0

Bibliography

[1] T. Anderson, D. Blumenthal, D. Casey, D. Clark, D. Estrin, L. Peterson, D. Raychaudhuri, J. Rexford, S. Shenker, and J. Wroclawski, "GENI: Global environment for network innovations," 2007.

[2] "NSF NeTS FIND Initiative." `http://www.nets-find.net`. Accessed: 2014-05-29.

[3] "National Science Foundation Future Internet Architecture Project." `http://www.nets-fia.net/`. Accessed: 2014-07-01.

[4] "The Network of the Future Projects of EU FP7." `http://cordis.europa.eu/fp7/ict/future-networks/home_en.html`. Accessed: 2014-05-29.

[5] "AKARI Architecture Design Project." `http://akari-project.nict.go.jp`. Accessed: 2014-05-29.

[6] J. Pan, S. Paul, and R. Jain, "A survey of the research on future internet architectures," *Communications Magazine, IEEE*, vol. 49, no. 7, pp. 26–36, 2011.

[7] "Global Environment for Network Innovations (GENI) project." `http://www.geni.net/`. Accessed: 2014-05-29.

[8] "Future Internet Research and Experimentation." `http://cordis.europa.eu/fp7/ict/fire/`. Accessed: 2014-05-29.

[9] "JGN2plus - Advanced Testbed Network for R&D." `http://www.ign.nict.go.jp/`. Accessed: 2014-05-29.

[10] "CERNET2 project." `http://www.cernet2.edu.cn/`. Accessed: 2014-05-29.

[11] P. Oppenheimer, *Top-Down Network Design*. Cisco Press, 2010.

[12] G. Huston, "Analyzing the internet bgp routing table," *The Internet Protocol Journal*, vol. 4, pp. 2–15, March 2001.

[13] N. Spring, R. Mahajan, D. Wetherall, and T. Anderson, "Measuring isp topologies with rocketfuel," *Networking, IEEE/ACM Transactions on*, vol. 12, no. 1, pp. 2–16, 2004.

[14] D. Meyer, L. Zhang, K. Fall, *et al.*, "Report from the IAB Workshop on Routing and Addressing," *IETF Internet Standard, RFC*, vol. 4984, 2007.

[15] Y. Nozaki, H. Tuncer, and N. Shenoy, "A tiered addressing scheme based on a floating cloud internetworking model," in *Proceedings of the 12th international conference on Distributed computing and networking*, ICDCN'11, (Berlin, Heidelberg), pp. 382–393, Springer-Verlag, 2011.

[16] Y. Nozaki, P. Bakshi, H. Tuncer, and N. Shenoy, "Evaluation of tiered routing protocol in floating cloud tiered internet architecture," *Computer Networks*, vol. 63, pp. 33–47, 2014.

[17] A. Feldmann, L. Cittadini, W. Mühlbauer, R. Bush, and O. Maennel, "Hair: hierarchical architecture for internet routing," in *Proceedings of the 2009 workshop on Re-architecting the internet*, ReArch '09, (New York, NY, USA), pp. 43–48, ACM, 2009.

[18] X. Yang, "NIRA: a new Internet routing architecture," in *Proceedings of the ACM SIGCOMM workshop on Future directions in network architecture*, FDNA '03, (New York, NY, USA), pp. 301–312, ACM, 2003.

[19] X. Xu, "Routing Architecture for the Next Generation Internet (RANGI)." Internet draft (Informational), august 2010.

[20] L. Subramanian, M. Caesar, C. T. Ee, M. Handley, M. Mao, S. Shenker, and I. Stoica, "HLP: a next generation inter-domain routing protocol," *SIGCOMM Comput. Commun. Rev.*, vol. 35, pp. 13–24, Aug. 2005.

[21] "Internet Research Task Force Routing Research Group." `http://trac.tools.ietf.org/group/irtf/trac/wiki/RoutingResearchGroup`. Accessed: 2014-07-09.

[22] M. Caesar, T. Condie, J. Kannan, K. Lakshminarayanan, and I. Stoica, "ROFL: routing on flat labels," in *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*, SIGCOMM '06, (New York, NY, USA), pp. 363–374, ACM, 2006.

[23] J. Pan, R. Jain, S. Paul, M. Bowman, X. Xu, and S. Chen, "Enhanced MILSA architecture for naming, addressing, routing and security issues in the next generation internet," in *Communications, 2009. ICC '09. IEEE International Conference on*, pp. 1 –6, june 2009.

[24] 3GPP, *General Packet Radio Service (GPRS) Service Description: Stage 2*, 3gpp ts 23.060 v7.1.0 ed., June 2006.

[25] S. Zhuang, K. Lai, I. Stoica, R. Katz, and S. Shenker, "Host mobility using an internet indirection infrastructure," *Wirel. Netw.*, vol. 11, pp. 741–756, November 2005.

[26] A. Anand, F. Dogar, D. Han, B. Li, H. Lim, M. Machadoy, W. Wu, A. Akella, D. Andersen, J. Byersy, S. Seshan, and P. Steenkiste, "XIA: An Architecture for an Evolvable and Trustworthy Internet," Tech. Rep. Technical Report CMU-CS-11-100, Department of Computer Science, Carnegie Mellon- University, Feb 2011.

[27] "MobilityFirst Future Internet Architecture Project." `http://mobilityfirst.winlab.rutgers.edu/`. Accessed: 2014-05-29.

[28] "The FP7 4WARD Project." `http://www.4ward-project.eu/index.php`. Accessed: 2014-05-29.

[29] "DAIDALOS, Designing Advanced network Interfaces for the Delivery and Administration of Location independent, Optimised personal Services project." `http://www.ist-daidalos.org/`. Accessed: 2014-05-29.

[30] National Institute of Information and Communication Technology (NICT) of Japan, "New generation network architecture akari conceptual design v 1.1," tech. rep., October 2008.

[31] V. Cerf, Y. Dalal, and C. Sunshine, "Specification of Internet transmission control program." RFC 675, 1974.

[32] "American Registry for Internet Numbers (ARIN), Number Resource Policy Manual (NRPM)," Tech. Rep. 2010.1, Department of Computer Science, Carnegie Mellon- University, April 2010.

[33] M. Yannuzzi, X. Masip-Bruin, and O. Bonaventure, "Open issues in interdomain routing: a survey," *Network, IEEE*, vol. 19, pp. 49–56, Nov 2005.

[34] J. Moy, "OSPF Protocol Analysis," *IETF Internet Standard, RFC*, vol. 1245, 1991.

[35] C. Alaettinoglu, V. Jacobson, and H. Yu, "Toward millisecond IGP convergence," in *Proceedings of the NANOG*, 2000.

[36] P. Pan, G. Swallow, and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels." Standards Track, May 2005.

[37] A. Kvalbein, A. F. Hansen, T. Čičic, S. Gjessing, and O. Lysne, "Multiple routing configurations for fast ip network recovery," *IEEE/ACM Trans. Netw.*, vol. 17, pp. 473–486, Apr 2009.

[38] Y. Liu and A. Reddy, "A fast rerouting scheme for OSPF/IS-IS networks," in *International Conference on Computer Communications and Networks, ICCCN 2004*, pp. 47 – 52, 2004.

[39] P. Narvaez, "Routing reconfiguration in ip networks," *Ph.D. dissertation, MIT*, Jun 2000.

[40] "BGP analysis." http://bgp.potaroo.net/index-bgp.html. Accessed: 2014-05-29.

[41] L. Subramanian, S. Agarwal, J. Rexford, and R. H. Katz, "Characterizing the internet hierarchy from multiple vantage points," in *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 2, pp. 618–627, IEEE, 2002.

[42] Y. Rekhter and T. Li, "A border gateway protocol 4 (bgp-4). Request for Comment RFC-1771," *Network Information Center*, 1995.

[43] S. Agarwal, C.-N. Chuah, and R. H. Katz, "Opca: Robust interdomain policy routing and traffic control," in *Open Architectures and Network Programming, 2003 IEEE Conference on*, pp. 55–64, IEEE, 2003.

[44] "IPv4depletion, IPv4/IPv6 and TCAM memory." http://www.ipv4depletion.com/?p=672. Accessed: 2014-05-29.

[45] "APNIC Labs, BGP Routing Growth in 2011." http://labs.apnic.net/blabs/?p=25. Accessed: 2014-05-29.

[46] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed internet routing convergence," *ACM SIGCOMM Computer Communication Review*, vol. 30, no. 4, pp. 175–187, 2000.

[47] T. G. Griffin and B. J. Premore, "An experimental analysis of BGP convergence time," in *Network Protocols, 2001. Ninth International Conference on*, pp. 53–61, IEEE, 2001.

[48] Z. M. Mao, R. Govindan, G. Varghese, and R. H. Katz, "Route flap damping exacerbates internet routing convergence," in *ACM SIGCOMM Computer Communication Review*, vol. 32, pp. 221–233, ACM, 2002.

[49] C. Villamizar, R. Chandra, and R. Govindan, "RFC 2439: Bgp route flap damping," *Internet Engineering Task Force*, 1998.

[50] D. Pei, M. Azuma, D. Massey, and L. Zhang, "BGP-RCN: Improving BGP convergence through root cause notification," *Computer Networks*, vol. 48, no. 2, pp. 175–194, 2005.

[51] H. Zhang, A. Arora, and Z. Liu, "G-BGP: Stable and fast convergence in the Border Gateway Protocol," *Ohio State University, Tech. Rep. OSU-CISRC-6/03-TR36*, 2003.

[52] A. Kivi, T. Smura, and J. Töyli, "Technology product evolution and the diffusion of new product features," *Technological Forecasting and Social Change*, vol. 79, no. 1, pp. 107–126, 2012.

[53] J. P. Choi, "The Provision of (Two-way) Converters in the Transition Process to a New Incompatible Technology," *The Journal of Industrial Economics*, vol. 45, no. 2, pp. 139–153, 1997.

[54] J. Farrell and G. Saloner, "Converters, compatibility, and the control of interfaces," *The journal of industrial economics*, pp. 9–35, 1992.

[55] T. Smura, A. Kiiski, and H. Hämmäinen, "Virtual operators in the mobile industry: a techno-economic analysis," *NETNOMICS: Economic Research and Electronic Networking*, vol. 8, no. 1-2, pp. 25–48, 2007.

[56] W. B. Arthur, "Increasing Returns and the New World of Business," *Harvard business review*, vol. 74, no. 4, pp. 100–109, 1996.

[57] K. J. Arrow, "The economic implications of learning by doing," *The review of economic studies*, pp. 155–173, 1962.

[58] N. Rosenberg, *Inside the black box: Technology and economics*. Cambridge University Press, 1982.

[59] M. L. Katz and C. Shapiro, "Technology adoption in the presence of network externalities," *The journal of political economy*, pp. 822–841, 1986.

[60] A. Hovav, R. Patnayakuni, and D. Schuff, "A model of Internet standards adoption: the case of IPv6," *Information Systems Journal*, vol. 14, no. 3, pp. 265–294, 2004.

[61] D. Joseph, N. Shetty, J. Chuang, and I. Stoica, "Modeling the adoption of new network architectures," in *Proceedings of the 2007 ACM CoNEXT conference*, p. 5, ACM, 2007.

[62] S. Killcoyne, G. W. Carter, J. Smith, and J. Boyle, "Cytoscape: a community-based framework for network modeling," in *Protein Networks and Pathway Analysis*, pp. 219–239, Springer, 2009.

[63] "The CAIDA AS Relationships Dataset." `http://www.caida.org/data/active/as-relationships/`. Accessed: 2013-01-20.

[64] N. Shenoy and Y. Pan, "Multi-meshed tree routing for internet manets," in *Wireless Communication Systems, 2005. 2nd International Symposium on*, pp. 145–149, IEEE, 2005.

[65] "Iperf: The TCP/UDP bandwidth measurement tool." `http://sourceforge.net/projects/iperf/`. Accessed: 2014-06-20.

[66] "OPNET Modelr: Making Networks and Applications Perform." `http://www.opnet.com/`. Accessed: 2012-03-20.

[67] C. Huitema, "The H Ratio for Address Assignment EfJiciency (RFC 1715)," *Retrieved May*, vol. 15, p. 2005, 1994.

[68] A. Apnic and N. RIPE, "Ipv6 address allocation and assignment policy." `http://www.ripe.net/ripe/docs/ipv6policy.html`. Accessed: 2014-05-29.

[69] "Emulab - Network Emulation Testbed." `http://www.emulab.net/`. Accessed: 2014-05-29.

[70] "Quagga Software Routing Suit." `http://www.quagga.net/`. Accessed: 2014-05-29.

[71] D. Pei, X. Zhao, D. Massey, and L. Zhang, "A study of BGP path vector route looping behavior," in *Distributed Computing Systems, 2004. Proceedings. 24th International Conference on*, pp. 720–729, IEEE, 2004.

[72] X. Zhao, B. Zhang, A. Terzis, D. Massey, and L. Zhang, "The impact of link failure location on routing dynamics: A formal analysis," in *ACM SIGCOMM Asia Workshop*, 2005.

[73] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for internet-scale systems," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 4, pp. 123–134, 2009.

[74] R. Bolla, R. Bruschi, A. Carrega, F. Davoli, D. Suino, C. Vassilakis, and A. Zafeiropoulos, "Cutting the energy bills of Internet Service Providers and telecoms through power management: An impact analysis," *Computer Networks*, vol. 56, no. 10, pp. 2320–2342, 2012.

[75] K. Argyraki, S. Baset, B.-G. Chun, K. Fall, G. Iannaccone, A. Knies, E. Kohler, M. Manesh, S. Nedevschi, and S. Ratnasamy, "Can software routers scale?," in *Proceedings of the ACM workshop on Programmable routers for extensible services of tomorrow*, pp. 21–26, ACM, 2008.

[76] V. Ravikumar and R. N. Mahapatra, "TCAM architecture for IP lookup using prefix properties," *Micro, IEEE*, vol. 24, no. 2, pp. 60–69, 2004.

[77] R. Panigrahy and S. Sharma, "Reducing tcam power consumption and increasing throughput," in *High Performance Interconnects, 2002. Proceedings. 10th Symposium on*, pp. 107–112, IEEE, 2002.

[78] H. Yu, "A memory-and time-efficient on-chip TCAM minimizer for IP lookup," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2010*, pp. 926–931, IEEE, 2010.

[79] D. Alderson, L. Li, W. Willinger, and J. C. Doyle, "Understanding internet topology: principles, models, and validation," *Networking, IEEE/ACM Transactions on*, vol. 13, no. 6, pp. 1205–1218, 2005.

[80] J. Baliga, R. Ayre, K. Hinton, and R. Tucker, "Photonic switching and the energy bottleneck," in *Photonics in Switching*, vol. 2007, pp. 125–126, 2007.

[81] L. G. GIANOLI, "Models and algorithms for energy saving in ip networks," 2010.

[82] "Global Price List." https://www-304.ibm.com/easyaccess3/fileserve?contentid=107260. Accessed: 2014-02-03.

[83] "Service Description: SMARTnet and SMARTnet Onsite Services." http://www.cisco.com/web/about/doing_business/legal/service_descriptions/docs/Smartnet_Onsite_Exhibit.pdf. Accessed: 2014-02-03.

[84] "Cisco U.S. Global Price Sheet." http://www.centurylink.com/business/asset/channel/cisco-us-global-price-sheet-pricing2-pr120537.pdf. Accessed: 2014-02-03.

[85] "Salary.com." http://www.salary.com/category/salary/. Accessed: 2014-07-03.

[86] "Third State Video Franchise Holder Employment Report." http://www.cpuc.ca.gov/NR/rdonlyres/2E0A3428-1D45-45D2-86CD-1E582D06FAFD/0/ThirdEmploymentReportforDIVCAFranchiseHoldersApril8_2011.pdf. Accessed: 2014-05-29.

[87] M. Gallaher and B. Rowe, "Ipv6 economic impact assessment (final report)," *National Institute of Standards and Teehnology, US Department of Commerce*, 2005.

[88] R. Bolla, R. Bruschi, F. Davoli, and F. Cucchietti, "Energy efficiency in the future internet: a survey of existing approaches and trends in energy-aware fixed network infrastructures," *Communications Surveys & Tutorials, IEEE*, vol. 13, no. 2, pp. 223–244, 2011.