Smart-Table Technology — Enabling Very Large Server, Storage Nodes, and Virtual Machines to Scale Using Flexible Network Infrastructure Topologies

Sujal Das
Product Marketing Director
Network Switching

July 2012

Introduction

Private and public cloud applications, usage models, and scale requirements are significantly influencing network infrastructure design. Broadcom's StrataXGS® architecture-based Ethernet switches support the SmartSwitch series of technologies to ensure that such network infrastructure design requirements can be implemented comprehensively, cost-effectively, and in volume scale. This set of innovative and unique technologies, available in current and future StrataXGS Ethernet switch processors, serves as the cornerstone of Ethernet switch systems from leading equipment manufacturers worldwide.

A critical element of cloud network scalability is the size of the forwarding tables in network switches deployed in the data center. This factor impacts many elements of data center scalability—the number of servers and virtual machines per server, and the ability to load-balance and provide full cross-sectional bandwidth across switch links. In turn, these scalability elements directly impact application performance and mobility. Virtual machines (VMs) and server sprawl, along with the increasing use of tunneling or overlay technologies in the data center, further exacerbate scaling challenges. Traditionally, the design approach for scaling forwarding table sizes has been to add more memory resources into the switch silicon or allow use of external memory resources. However, today's increasing density and bandwidth needs of data center switches, combined with the need for cost and power optimization, demand new innovations in how switch-forwarding tables are best integrated, utilized, and scaled.

This white paper explores the forwarding table size requirements in private and public cloud data center networks, and considers how such requirements affect the design of data center network switches. It also describes features that are enabled by Broadcom's Smart-Table technology, part of Broadcom's SmartSwitch series of technologies, engineered specifically to meet feature and scale requirements of private and public cloud networks. Smart-Table technology encompasses comprehensive best practices for today's high-performance data center switches, and illustrates an optimal solution for addressing the evolving needs of next-generation cloud implementations.

The Role of Switch-Forwarding Tables

A forwarding table or forwarding information base (FIB) implemented in a network switch is used in network bridging, routing, and similar functions to find the proper interface to which the input interface should send a packet for transmission. A layer 2 (L2) forwarding table contains Media Access Control (MAC) addresses, a layer 3 (L3) forwarding or routing table contains IP (Internet Protocol) addresses, a Multi Protocol Label Switching (MPLS) table contains labels, and so on. Within the context of the data center, certain forwarding tables are most relevant; these include the L2 MAC address table, the L3 Host and IP Multicast Entries table, the Longest Prefix Match (LPM)[1] routes table, and the Address Resolution Protocol (ARP) with Next-Hop Entries table[2]. The

1. LPM refers to an algorithm used by routers in IP networking to select an entry from a routing table. Because each entry in a routing table may specify a network, one destination address may match more than one routing table entry. The most specific table entry — the one with the highest subnet mask — is called the longest prefix match. It is called this because it is also the entry where the largest number of leading address bits in the table entry match those of the destination address.
2. ARP is used to associate an L3 address (such as an IP address) with an L2 address (MAC address). Next-hop is a common routing term that indicates the IP address of the next hop to which packets for the entry should be forwarded.

size of each of these tables in network switches has a bearing on how cloud networks can scale. When these tables reach capacity — because the forwarding tables in switches are small — scaling problems occur. One example is MAC address learning. If the working set of active MAC addresses in the network (affected by the number of servers or VMs in the network) is larger than the forwarding table in switches, some MAC address entries will be lost. Subsequent packets delivered to those MAC address destinations will cause flooding and severely degrade network performance. Similar performance implications affect other types of forwarding tables as well. Optimal network performance can be ensured only by deploying switches that incorporate table sizes larger than the active addresses in the network.

Sizing Needs of Switch-Forwarding Tables

The number and types of active addresses in the data center network (L2 MAC, L3 host and IP multicast addresses, LPM and ARP/next-hop entries) are impacted by multiple data center server, VM, and network deployment scenarios. These scenarios may include a broad range of various network topologies and network virtualization technologies.

Proliferation of Standalone Switches

Mega-scale data centers are being built to satisfy the needs of public cloud service and business models. These data center networks are being designed from the ground up for commodity-level scaling, using cost effective, off-the-shelf, and easily replaceable switches. The widespread evolution of this type of implementation has resulted in increased use of standalone switch form factors, in both the access and aggregation layers of the switch. In private cloud data centers, standalone top-of-rack switches have become the norm in the access layer, enabling cost-effective scaling and flexible connectivity to the server racks. This causes increasing pressure on switch silicon designs because standalone switches are typically designed with single switch silicon, incorporating no external memories. Yet single high-density switch silicon must be able to support forwarding table scale requirements for access- and aggregation-layer deployments. Unlike chassis-based switches that rely on multiple switch and fabric processor silicon with external memories for forwarding table scaling, this new breed of switch silicon must accomplish the necessary forwarding table scale with only internal memory, while maintaining minimum cost and power consumption.

Server and Virtual Machine Sprawl

The number of servers and VMs per server in private and public cloud data centers is increasing exponentially, and by all appearances, without limitation. Mega data centers have tens of thousands of servers; each server is capable of hosting ten to twenty VMs and is expected to support fifty or more in the coming years. This dramatic expansion in capacity has significant implications for data center networking and represents a sea change in data center architecture and design. Traditionally, data center networks were designed with the basic premise that a server has a single identity; the sum total of one MAC address, one IP address, and each application requiring its own server. Today, VMs increase the density of server identities in terms of MAC and IP addresses as well as applications. The rate of growth of such deployments places stress on the data forwarding capacity of the network and, specifically, the forwarding table sizes in switches.

Varied Network Topologies

Today's data center and cloud networks incorporate varied network topologies, impacting the types and sizes of forwarding tables needed in network switches.

L2 Networks

Some data center clustered applications require L2 adjacency for best performance. The clustered database is an example of such an application; in this scenario, data warehousing and business analytics operations are scaled by adding more compute and storage nodes to the cluster. High-performance trading and other latency-sensitive applications may also achieve maximum performance though the use of such 'flat' L2 networks, or architectures with fewer layers than a traditional three-tier network. In other instances, networks are provisioned for L2-based forwarding only to ensure network management simplicity. Network switches that connect servers running such applications must support L2 MAC address tables only, at required scales. An example of such a network configuration is shown in Figure 1. In these cases, the size of switch-forwarding tables for L3 host and IP multicast addresses, LPM routes and ARP/next hop entries are less relevant.

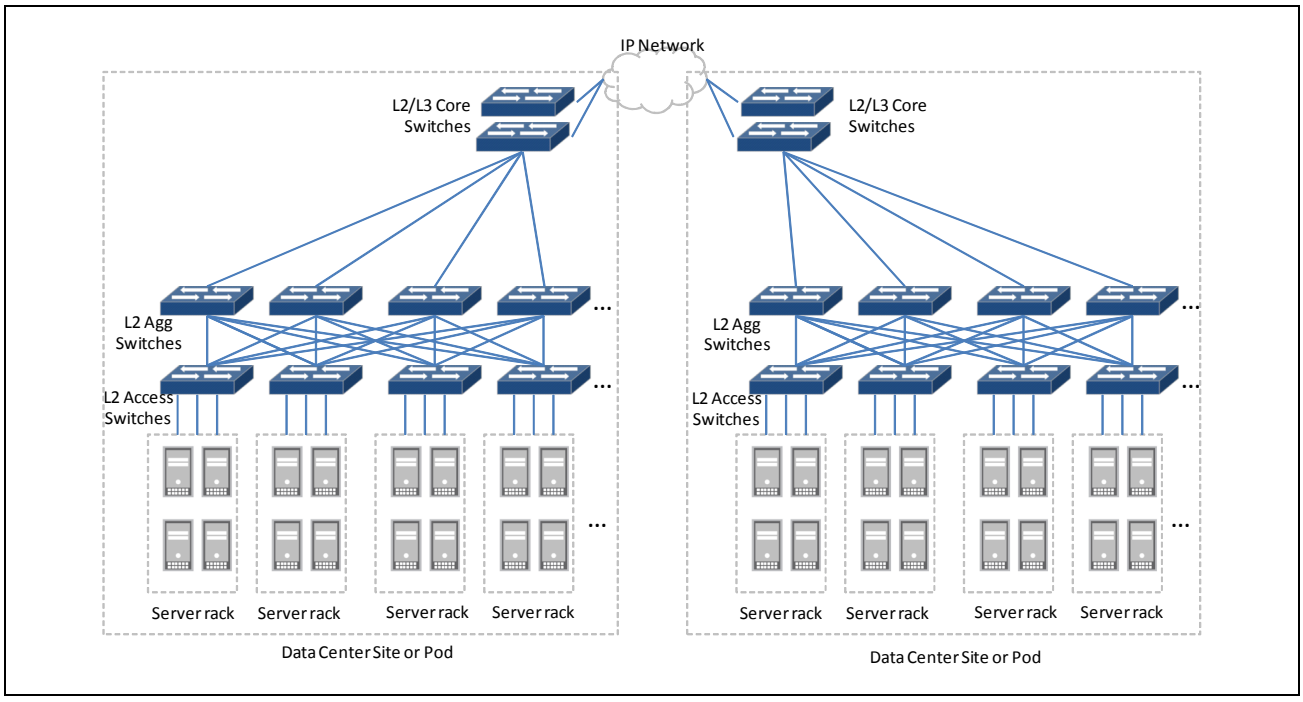

Figure 1: L2 Network Configuration Example

While database and storage applications in the data center may not always be virtualized (to eliminate networking performance overheads induced by the hypervisor software layer), business logic and web front-end applications are almost always virtualized, resulting in the presence of a large number of VMs in such networks. As the density of servers and the number of VMs per server increases, the number of active MAC addresses that must be forwarded by switches increases. Considering a data center network of 10,000 servers with eight VMs per server, switch-forwarding tables can easily need to support 80,000 to 100,000 MAC addresses.

Virtual machine mobility requires L2 adjacency; this can be achieved in different ways, the simplest being an L2-only network. Illustrated in Figure 1, L2 adjacencies can be maintained in a pod, or even a site, by configuring the access and aggregation switches as L2 switches and the core switches as L2/L3 routers. To enable such flat L2 networks with multipathing for full cross-sectional bandwidth, technologies such as Transparent Interconnection of Lots of Links (TRILL) or Shortest Path Bridging (SPB) can be deployed. To ensure scalability, these L2 technologies require large L2 MAC forwarding tables in switches.

To enable VM mobility across network segments, a flat virtual L2 domain can be formed using tunneling technologies. These tunneling technologies utilize L2-only schemes such as MAC-in-MAC, and again, such deployments require large L2 MAC forwarding tables in switches to ensure scalability.

L3 Networks

When data centers are designed for mega-scale, as in public clouds, the proven scalability and reliability of L3 networking is used. In this type of network design, access layer and aggregation layer switches are configured as L3 switches, shown in Figure 2. Multipathing is achieved using routing protocols such as Open Shortest Path First (OSPF) and Equal Cost Multipathing (ECMP). To enable L3-based scaling, network switches must support a large number of L3 forwarding table entries. In this scenario, a small L2 MAC table is adequate, but some L3 host entries and a very large number of LPM routes entries are desirable. The situation changes if the servers are virtualized, in which case MAC addresses assigned to VMs become active in the network, and switches must be provisioned for larger L2 MAC tables.

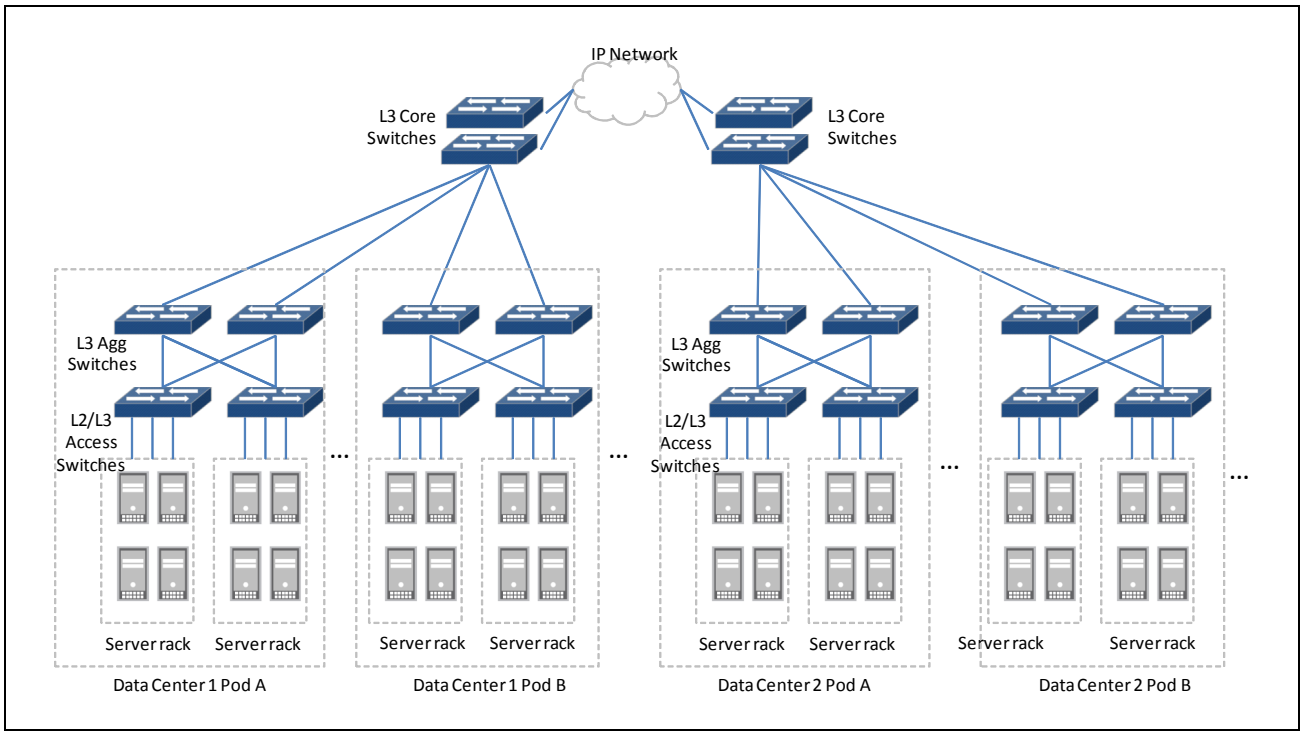

Figure 2: L3 Network Configuration Example

L3 Networks with L2oL3 Overlays

In the L3 networks-based scenario depicted in Figure 2, L2 adjacencies and multi-tenancy scale are achieved using Layer 2 over Layer 3 (L2oL3) overlay network virtualization technologies such as Virtual Extended LAN (VxLAN), Network Virtualization using Generic Route Encapsulation (NVGRE) or Layer 2 over Generic Route Encapsulation (L2GRE). In Figure 3, virtual L2 domains are created by the hypervisor virtual switches that serve as overlay end points. The L2 MAC address table forwarding requirements on a per-VM basis are limited to the hypervisor virtual switches. Switches carrying L2oL3 tunneled packets have smaller L2 forwarding requirements. Some L3 Host entries are required, for example those associated with each virtual switch. To address the server downlinks and multi-way ECMP links on access and aggregation switch layers, a large number of LPM routes entries is desirable.

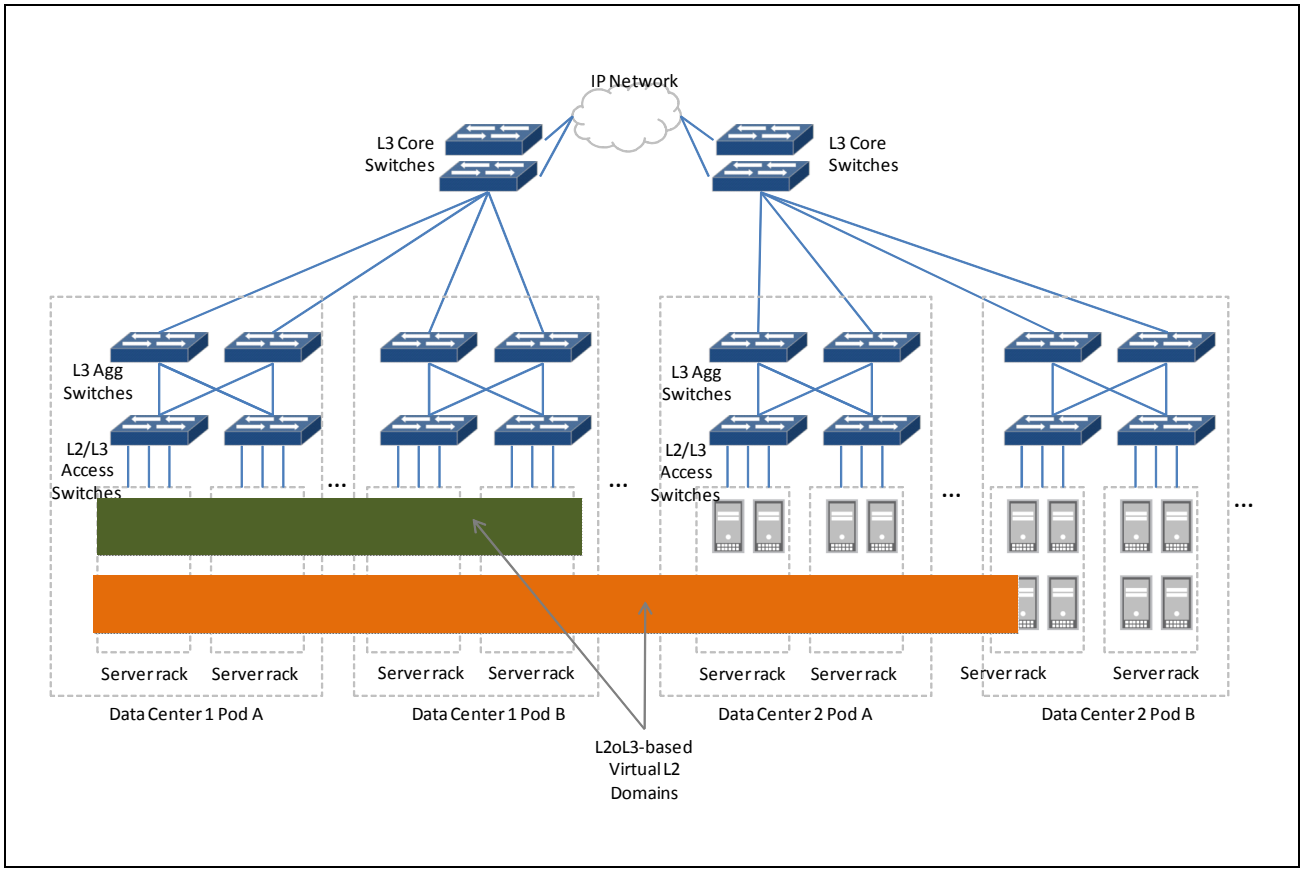

Figure 3: Example of L3 Network With L2oL3 Overlays, Showing Two Virtual L2 Domains

Summary of Sizing Needs for Switch-Forwarding Tables

The preceding material illustrates the range of different network topology needs and practices that are evolving in today's data center. It is by no means an exhaustive discussion, but serves the purpose of identifying three key network switch silicon design requirements for switch-forwarding tables:

- Switch silicon must meet the necessary forwarding table scale requirements with internal memory only, while maintaining minimum cost and power consumption.
- Increasing and varied network traffic equates to increased stress on the data forwarding capacity of the network, requiring larger forwarding table sizes in switches.
- Choices in which data center applications and network topologies are deployed affects the types and sizes of forwarding tables needed in switches.

Impact on Switch Silicon Design

Increased bandwidth and port densities, and larger forwarding table sizes, translate into large on-chip memories and complex combinational logic stages that must run at very high speeds. While the largest forwarding table scale can be guaranteed most simply by increasing the size of memories, it is generally prohibitive in terms of cost and power requirements to include very large integrated forwarding table memories on a single switch chip that operates at such elevated performance levels. Conversely, relying on external forwarding table memories to maximize table scale places a ceiling on performance, as external memory access times cannot feasibly match the single-chip switching throughputs demanded of today's data center access layer switches. The optimal solution is a fully integrated forwarding table architecture that enables maximum sizing of table resources.

Data center switch chip architectures now are facing aggregate processing bandwidth requirements that favor a multi-pipeline approach to meet performance, cost, and power requirements. While multi-pipeline switch designs allow for bandwidth scalability by localizing the packet processing, forwarding plane decisions are most optimal when made globally across all switch ports; this option avoids synchronization delays and overheads. Further, adopting a multi-pipeline design creates partitioning challenges and chip tradeoffs that demand careful consideration. The global scope and multi-pipeline approach mandate an optimum shared forwarding plane architecture.

Finally, while sizes of the switch-forwarding tables matter, the type and size of forwarding tables in the switch silicon cannot be a fixed measurement. Depending on where the switch is deployed with respect to the network topology and the data center applications it serves, the sizes of the forwarding tables must ideally be configurable, preferably using forwarding table profiles.

Introducing Smart-Table Technology

The Broadcom StrataXGS switch architecture is optimized for cloud networking and features Smart-Table technology, addressing the switch-forwarding table sizing needs of high-performance cloud network designs today and as these sophisticated networks are poised to evolve.

The StrataXGS switch architecture for high-density data center switches features a multi-pipeline design for performance and port density scaling. This architecture is differentiated by its centralized processing — namely the shared forwarding plane that forms the heart of Smart-Table technology, a global multistage classifier, and a centralized dynamic buffer memory management unit (MMU) that enables global admission control, queuing, policing, and shaping engines. This unique centralized design enables all data path and quality of service system resources to be configured on a system-wide basis, regardless of the number of pipelines used in the switch for port and bandwidth scaling purposes. This architecture allows global rules and flows to be replicated across all ports in the system and system-wide synchronous updates. It enables forwarding table coherency and replication, efficient link aggregation (LAG) resolution and failover, and effective load balancing across port groups.

At a high level, a packet switching pipeline has two components including an ingress/egress packet processing pipeline, and a packet MMU. The efficient shared forwarding plane architecture supported by Smart-Table technology enables these components to support large and integrated memory instances for L2 MAC entry tables, L3 IP unicast and multicast forwarding tables, LPM routes, next-hop, and other forwarding tables. A flexible lookup partitioning further provides superior efficiency and scale.

In addition, Smart-Table technology includes a significant innovation that enables dramatic improvements in utilizing available forwarding table space implemented in on-chip integrated memory. For example, instead of having four fixed-size tables for L2 MAC entry tables, L3 IP unicast and multicast forwarding tables, LPM routes and next-hop tables — as seen in switches available in the market today — the tables can be unified into a single shareable forwarding table (see Figure 4).

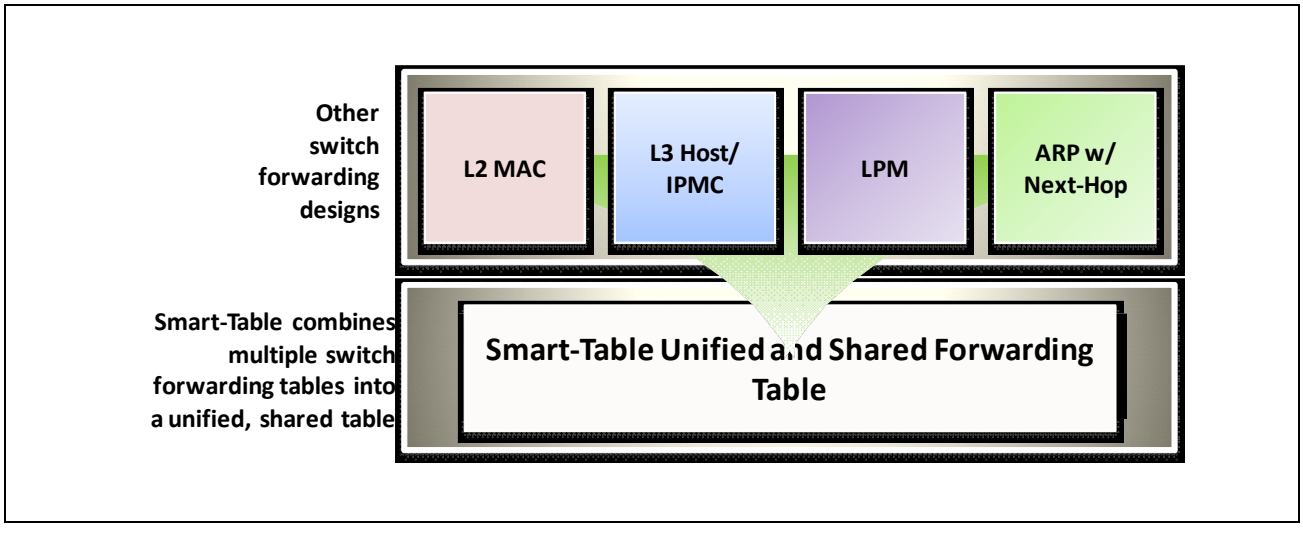

Figure 4: Smart-Table Technology Enables a Unified and Shared Forwarding Table with High Utilization

Since switch-forwarding table type and size requirements vary, Smart-Table technology allows configuration of the unified forwarding table capacity into unique Smart-Table Profiles, optimized for the specific type of network deployment. For example, Smart-Table Profiles can be used to configure the same network switch, built using Smart-Table technology, but catering to various network topology requirements (see Figure 5):

- Smart-Table Profile A is a balanced L2 and L3 profile. For example, 25 percent of the total table size can be allocated to each of the system's L2 MAC, host, LPM routes and ARP/next-hop entry tables.
- Smart-Table Profile B is an L2-heavy profile. For example, 90 percent of the total table size can be allocated to L2 MAC entries, with the reminder allocated to host, LPM routes and ARP/next-hop entry tables.
- Smart-Table Profile C is an L3 LPM routes-heavy profile with an adequate number of IP host entries. For example, 75 percent of the total table size can be allocated to LPM routes entries, 10 percent allocated to IP host and next-hop entry tables, with the remainder allocated to the L2 MAC entry table.

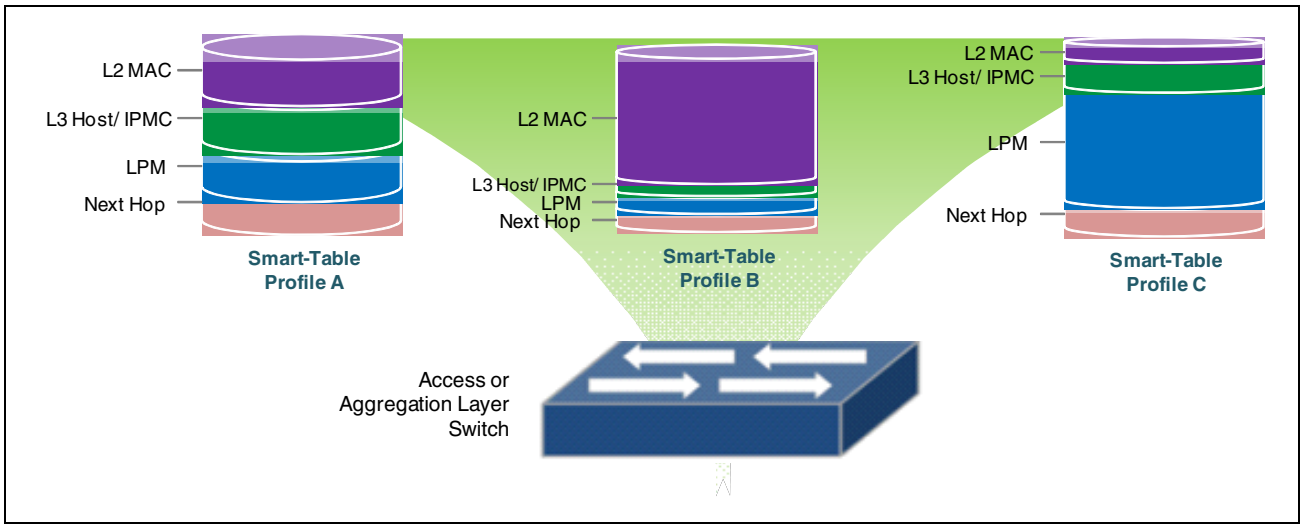

Figure 5: Examples of Smart-Table Profiles

Additional profiles can be defined and deployed in Smart-Table enabled network switches. Software development kits (SDKs), available with Broadcom Smart-Table enabled switches, allow easy configuration of the desired Smart-Table Profiles. This delivers excellent flexibility in switch deployment, and fuels a higher return on investment by allowing the same network switch system to be repurposed for different roles within the data center network.

Application of Profiles and Table Sizing

With its large table scale and profiling abilities, Smart-Table enabled switches can facilitate large cloud networks, supporting the range of cloud network topologies. The following example Smart-Table Profiles can be applied to specific topologies, allowing network designers to estimate the number of servers and VMs that can be deployed in the network. Determining the actual number of servers and VMs that can be supported depends on the specific Broadcom silicon used in the network switch.

L2 Networks

In this scenario, access and aggregation layer switches are L2 with L3 deployed at the core. Multiple VLANs are configured all the way from access to core ports. All access and aggregation switches must learn and switch dependent upon the MAC addresses of the connected servers, possibly with multiple VLAN instances per server. When servers are virtualized, the MAC addresses to be learned include all physical server addresses on each VLAN as well as the virtual machine's MAC addresses. Smart-Table Profile B, an L2-heavy profile, can be applied to the access and aggregation layer switches. By implementing this profile, up to 150,000 physical servers, or 200,000 VMs on a smaller number of physical servers, may be supported.

L3 Networks

In this scenario, access and aggregation layer switches operate at L3 (i.e., as routers). Access and aggregation switches must store routes to all server subnets as LPM entries, and may also need store routes to interior links as LPM entries. Exact-match IP host route entries may be used for hosts themselves or for routers. In this case, Smart-Table Profile C, an L3 LPM-heavy profile with an adequate number of IP host entries, is suitable for the access and aggregation layer. The actual number of entries used is proportional to the number of server subnets, which is usually an order of magnitude smaller than the number of physical servers. The number of entries also may depend, however, on topology — or the number of transit links and whether they are numbered by the deployed IP routing protocol. This profile allows up to 40,000 physical servers to be serviced by the L3 network in this mode. If the servers are virtualized, the additional MAC addresses belonging to VMs are not material to the profile, as only one access switch/router needs to store the VM addresses behind a given physical server. Virtualization, however, may dramatically push the LPM table usage, to the extent that additional subnets are created to accommodate additional VMs.

L3 Networks with L2oL3 Overlays

Virtualization overlays are designed to perform over L3, L2, or combination networks, although they are a particularly good fit for L3 down to the access layer. In either case, the overlays reduce the number of table entries consumed in transit switches and routers because only the underlay network addresses (outer headers) must be accommodated. When access and aggregation layer switches are L3, only the underlay network routes must be stored. Unlike the previous scenario, the addition of VMs has no impact on the number of routes stored. Smart-Table Profile C, an L3 LPM routes-heavy profile with an adequate number of IP host entries for use with physical servers, is suitable for this scenario. Based on the example of 40,000 physical servers deployed as an L3 network (using L2oL3 technologies where the hypervisor manages the overlay addresses or inner header), the same total number of servers can still be supported. Assuming about 20 VMs per server, the number of VMs in such a network can approach 1M.

Summary

Data center applications and network topologies affect the types and sizes of forwarding tables needed in switches. Inadequate forwarding table sizes can severely degrade network performance. The use of VMs and large-scale cloud network build-outs exacerbate the problem, demanding not only larger table sizes but also variations in table sizes based on the chosen network topology. The need for high-bandwidth and high-density switches — and the economies of scale essential to cloud networking — require efficiencies in networking switch designs, including how the switch-forwarding plane is architected. Broadcom's Smart-Table technology, available in its StrataXGS® architecture-based data center switches, delivers larger table sizes on single-silicon solutions with integrated memory. Smart-Table Profiles are incorporated to significantly enhance forwarding table utilization. Application of these profiles enables network switches to be flexibly deployed in various network topologies by optimizing forwarding table sizes. Return on investment is significantly improved, as the same network switches can be repurposed if network topologies change and a different profile of forwarding table sizes is required. Network and IT managers building high-performance cloud networks need maximum flexibility in building network topologies to service their business needs today and tomorrow. Smart-Table technology from Broadcom future-proofs network designs, enabling changes induced by new scaling or application needs, and providing peace of mind to network and IT managers by delivering on these critical long-term design needs.

Connecting everything®

BROADCOM

BROADCOM CORPORATION
5300 California Avenue
Irvine, CA 92617

Phone: 949-926-5000
Fax: 949-926-5203
E-mail: info@broadcom.com
Web: www.broadcom.com

StrataXGS_SmartSwitch-WP101-R July 2012