Innovations in Switch Forwarding Tables Boost Cloud Deployments

Oct 12 2012 Sujal Das (Brodcom Director Product Marketing)

> Trident2 was introduced in August 2012. It appears that is the switch ASIC that Sujal is talking about. Both Trident2 and Trident2+ mention Smart-table in their introductory slide deck.

Private and public cloud applications, usage models, and scale requirements are significantly influencing network infrastructure design, with a growing focus on scalability. 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 VMs (virtual machines) per server, as well as 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. 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.

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 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.

Key Considerations in Sizing Needs for 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 – which in turn may include a broad range of various network topologies and network virtualization technologies.

For example, some data center clustered applications require L2 adjacency for best performance. The clustered database illustrates 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. 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. 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.

In other L3 network-based scenarios, 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). 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.

When the network switch sits at the edge of the enterprise, facing the WAN (wide area network), the switch may need to support protocols such as Multi-Protocol Label Switching (MPLS) in addition to L2 and L3 forwarding. In this scenario, the switch forwarding table should be able to support an adequate number of MPLS table entries.

By examining this range of different network topology needs and practices that are evolving in today's data center, three key network switch silicon design requirements emerge as essential 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. And lastly, choices in which data center applications and network topologies are deployed affect the types and sizes of forwarding tables needed in switches.

Impact on Switch Silicon Design

In turn, these requirements have specific impact on ideal switch silicon design. For example, 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 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.

Driven by these requirements, network operators are turning to switch architectures optimized for cloud networking such as the Broadcom StrataXGS architecture. StrataXGS features Smart-Table technology, part of Broadcom's SmartSwitch series of technologies, engineered specifically to addresses the switch forwarding table sizing needs of high-performance cloud network designs today and as these sophisticated networks are poised to evolve.

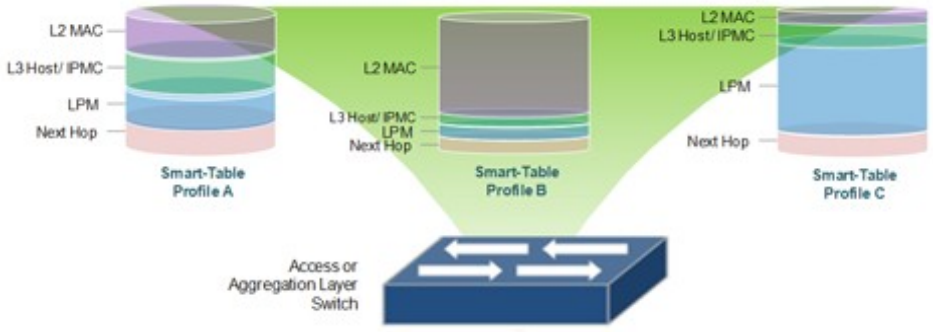
Defining Optimal Switch Architectures

Intelligent table sizing technology such as Smart-Table 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.

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. These profiles can be used to configure the same network switch, but cater to various network topology requirements. For example, in a balanced L2 and L3 profile, 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. In an L2-heavy profile, 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. A third option may be an L3 LPM routes-heavy profile with an adequate number of IP host entries. In this 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. In an L3 and MPLS profile, 90 percent of the total table size can be allocated to host, LPM routes, next-hop, and MPLS entries, with the remaining 10 percent allocated to the L2 MAC entry table.

Examples of Smart-Table Profiles



With extensive table scale and profiling abilities, Smart-Table enabled switches can facilitate large cloud networks, supporting the range of cloud network topologies. This in turn allows network designers to estimate the number of servers and VMs that can be deployed in the network, whether it is based on L2, L3 or L3 with L2oL3 overlay technologies. Click to enlarge graphic.

Applying Profiles and Table Sizing Moving Forward

Today, innovative and proven forwarding table technology delivers larger table sizes on single-silicon solutions with integrated memory. Flexible table profiles 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. Intelligent forwarding table technology 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.

---

[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-op is a common routing term that indicates the IP address of the next hop to which packets for the entry should be forwarded.

Editor's note: Sujal Das's co-author on this column is Rochan Sankar, Associate Product Line Director, Infrastructure and Networking Group, Broadcom Corp. He serves as Associate Director of Product Line Management for Broadcom Corporation's Infrastructure and Networking Group (ING) and has 13 years of experience in defining and managing leading-edge semiconductor products for the networking and communications industries.

Industry Perspectives is a content channel at Data Center Knowledge highlighting thought leadership in the data center arena. See our guidelines and submission process for

information on participating. View previously published Industry Perspectives in our Knowledge Library.