Avoiding Network Polarization and Increasing Visibility in Cloud Networks Using Broadcom Smart-Hash Technology

Sujal Das
Product Marketing Director
Network Switching

Karthik Mandakolathur
Sr Product Line Manager
Infrastructure and Networking/XGS

August 2012

Introduction

Today's massive scale data centers offer high aggregate network bandwidth for large compute clusters. Fast, fat, and flat networks are widely deployed with the more ubiquitous adoption of computer virtualization and the proliferation of clustered applications. 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.

Data centers must be purpose-built to handle current and future workloads — evolving rapidly and driven by high volumes of end users, application types, cluster nodes, and overall data movement in the cloud. In turn, the cross-sectional bandwidth of these cloud-scale data center networks is increasing rapidly, outpacing the increase in physical link speeds. ECMP (Equal Cost Multipathing) and port-channeling are common implementations that construct point-to-point higher-capacity logical paths using multiple redundant parallel physical paths. Traditionally, both ECMP and port-channel implementations attempt to distribute flows uniformly across the physical links that form the logical path. The decision as to which flows use which physical links is based on a static hash of a fixed set of fields from the packet header. This static hashing scheme is sub-optimal and gives rise to network polarization whereby multiple traffic flows may traverse and burden the same link and leave other links underutilized. This limitation no longer suits the scalability needs of today's data centers. Broadcom's Smart-Hash technology, part of Broadcom's SmartSwitch series of technologies, implements several new hashing enhancements that improve network performance and overcome limitations imposed by traditional schemes. This white paper discusses current data center traffic trends, the implications of such, and related enhancements featured in Smart-Hash technology.

Traditional Hashing Mechanism

As shown in Figure 1, traditional load-balancing systems split traffic bound through a logical fat link to multiple outgoing physical links. Typically, the physical link corresponding to a flow is ascertained by calculating a hash based on packet header fields and a subsequent modulo operation based on the number of physical links. A good load-balancing system should be able to split evenly the traffic to the multiple outgoing links. In addition, packets belonging to the same flow should flow out in order to the end destination.

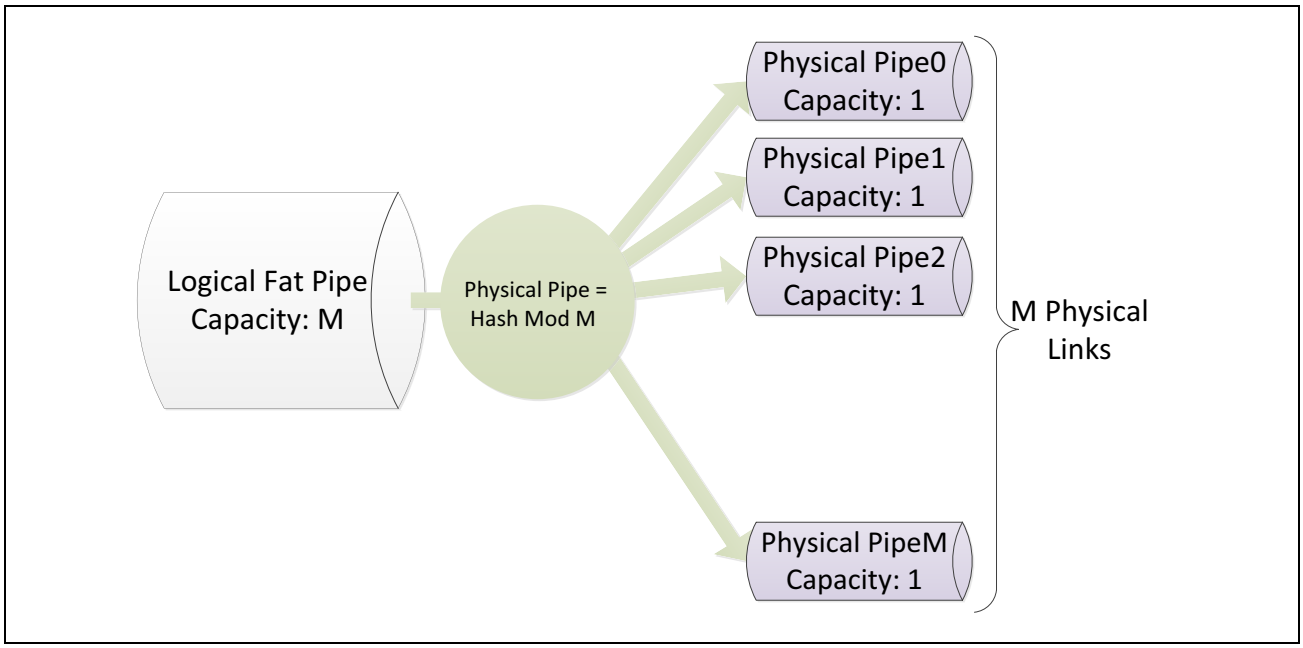

Figure 1: Traditional Hashing Used for Load Balancing

This static hashing scheme, which has worked well in carrier and enterprise networks, was later borrowed for use in the data center. Considering current trends in the data center traffic patterns, however, this scheme is no longer effective for data centers and cloud networks.

Web, application, and database server applications running as VMs (virtual machines) that can reside in any server in any rack — coupled with the increased use of clustered applications (such as Hadoop) in modern data centers — results in increased east-west traffic patterns in data center networks. Such east-west traffic includes server-to-server, server-to-storage, and server rack-to-server rack traffic. This trend is changing the inherent design of network topologies, from oversubscribed and tiered networks to fast, fat, and flat networks, which require new features in network switches.

Data Center Traffic Trends and Implications

Data center networks are a hotbed for innovation and, in recent years, are seeing unprecedented growth. Large enterprises are building enormous data centers with massive scale. Others are choosing to host their data centers in the cloud. Driven by the latest silicon advances, increased bandwidth and port densities of switch-on-a-chip systems, traditional three-tier network designs are being replaced by fast, fat, and flat networks that consist of resilient and flexible CLOS topologies with very high cross-sectional bandwidth.

Following are some essential implications of this trend:

- Data centers of such massive scale, for example, those using several thousand links, will experience frequent link failures resulting in network polarization. It is critical that networks perform normally and deliver packets in order under such failure conditions.
- Newer protocols and encapsulations are introduced year after year to improve data center automation and network management. Newer packet header fields redefine flows and how packets need to be treated.
- With the introduction of new network features, the general ability to debug and trace packets becomes a necessity.
- With the adoption of cloud hosting services, security has become ever more essential. Stateful packet inspection and intrusion detection systems will continue to gain importance.

To effectively manage these new implications, Broadcom's Smart-Hash technology features the following enhancements to the traditional hashing scheme:

- *Resilient Hashing* addresses link failure and live topology changes.
- *Flexible Hashing* handles newer encapsulations and protocols.
- *Symmetric Hashing* incorporates packet traceability and stateful protocol debug.

Resilient Hashing

Consider the example of traditional static hashing shown in Figure 2.

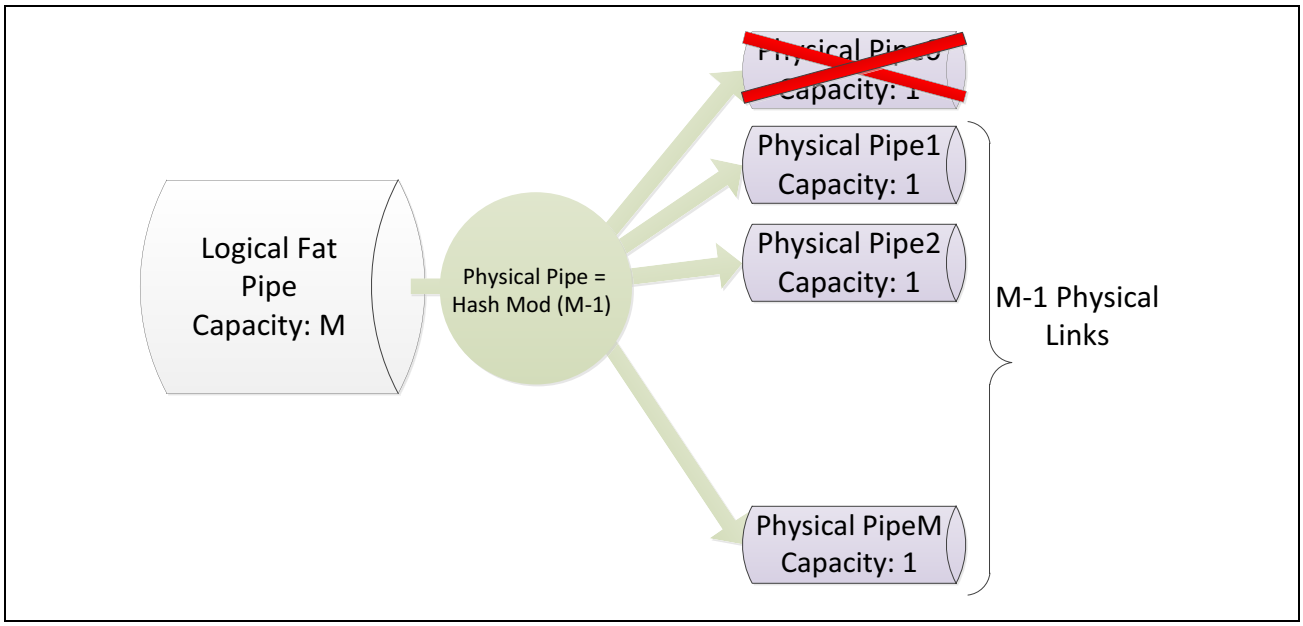

Figure 2: Traditional Static Hashing During Link Failure

M physical links are used to form a logical fat pipe. The static hash scheme uses a modulo-M operation to associate a flow with a physical link. In case of a link failure, this modulo operation will change to a modulo-(M-1) operation. In this scenario, even the flows that did not originally flow through the failed link may be assigned a new link. This reassignment may temporarily result in out-of-order packet delivery even for the flows that were not using the failed link.

In contrast, the Smart-Hash Resilient Hashing scheme, as shown in Figure 3 on page 6, incorporates a resilient hashing engine to associate flows with physical ports. In case of link failure, only the affected flows are redistributed uniformly across the remaining good physical links. Flows originally using the good links remain unaffected and are not reassigned to a new link.

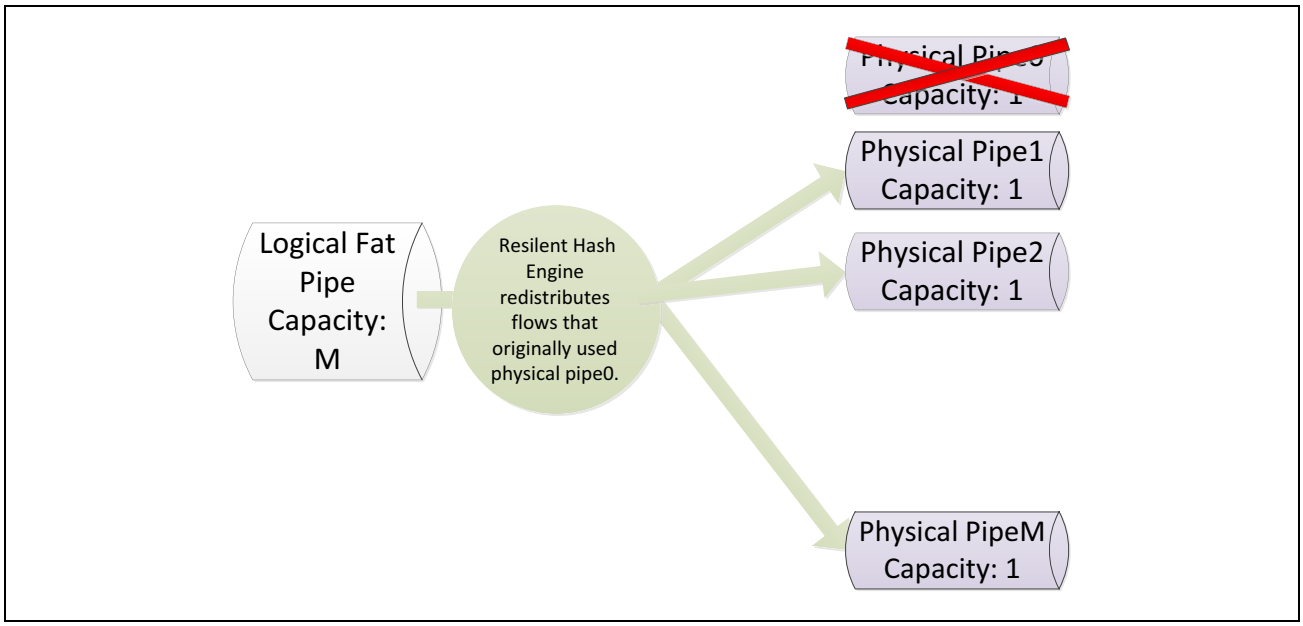

Figure 3: Resilient Hashing During Link Failure

Flexible Hashing

Data center features are evolving rapidly, as illustrated by the accelerated adoption of tunneling protocols such as VXLAN (Virtual Extended LAN) and NVGRE (Network Virtualization using Generic Routing Encapsulation). The common advantage of these tunneling schemes is that transit switches in the network primarily operate on the outer headers, so that only switches in the periphery need to treat these packets differently.

From the vantage point of the L2oL3 (layer 2 over layer 3) transit switch, the inner header of the packet changes more frequently than does its outer header. Traditional static hashing schemes are based on a standard set of fields in the outer packet header. With minimal variation in the outer header, it is difficult for network transit switches to distribute packets evenly across physical paths. Flexible Hashing enhances the hash to include more packet header fields at programmable offsets, and also supports L3, L3 Tunneling, and L4 packets (see Figure 4).

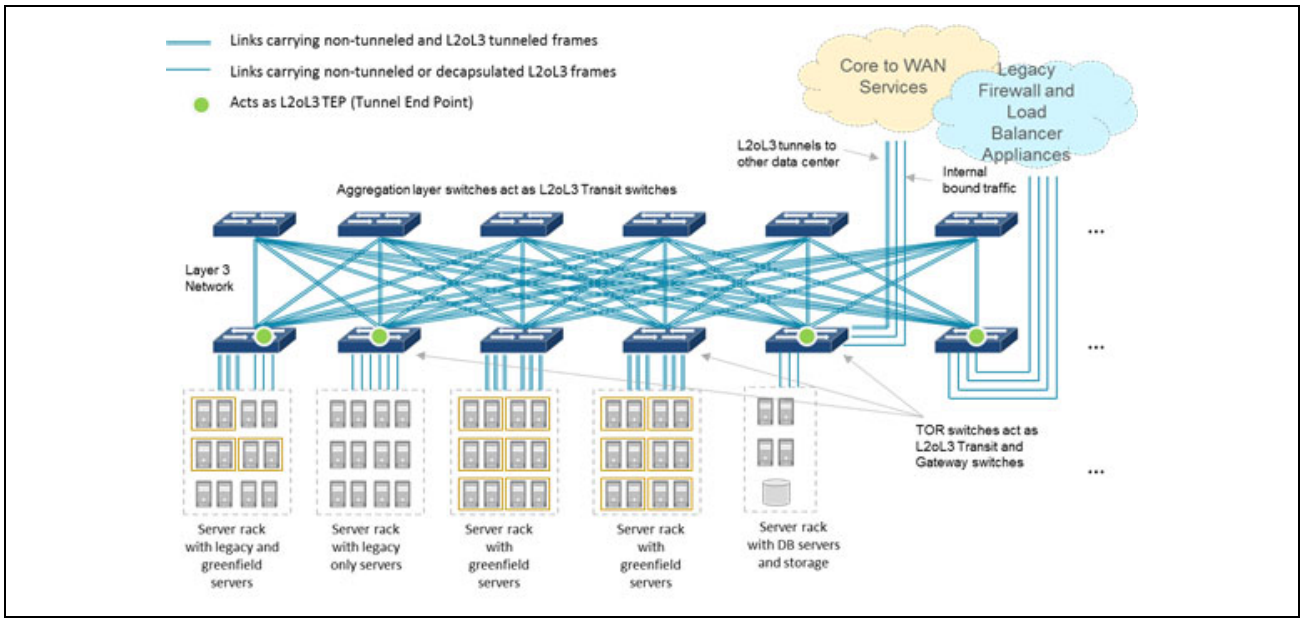

Figure 4: Typical Topology in a Data Center Using L2oL3 Tunneling

Symmetric Hashing

Symmetric Hashing ensures that packets belonging to the same bidirectional network communication travel the same physical paths in both directions. This requirement is necessary for intrusion detection systems that are placed in line on the network to analyze higher-level bidirectional transactions at the packet level. For network designers, this feature is also a useful and convenient debug feature. For example, protocol inspection devices can simply rely on the data captured on a single physical port. With this data, the analyzers will be able to reconstruct higher-level transactions.

Consider the example shown in Figure 5. Two service modules are connected to the ToR (top-of-rack) switch through an EtherChannel. Service module0 must see IP traffic flowing in both directions. Symmetric Hashing in the ToR normalizes the hash computation and yields the same hash value for packets flowing in both directions. As a result, packets flowing in both directions travel through port0 to the intrusion detection device.

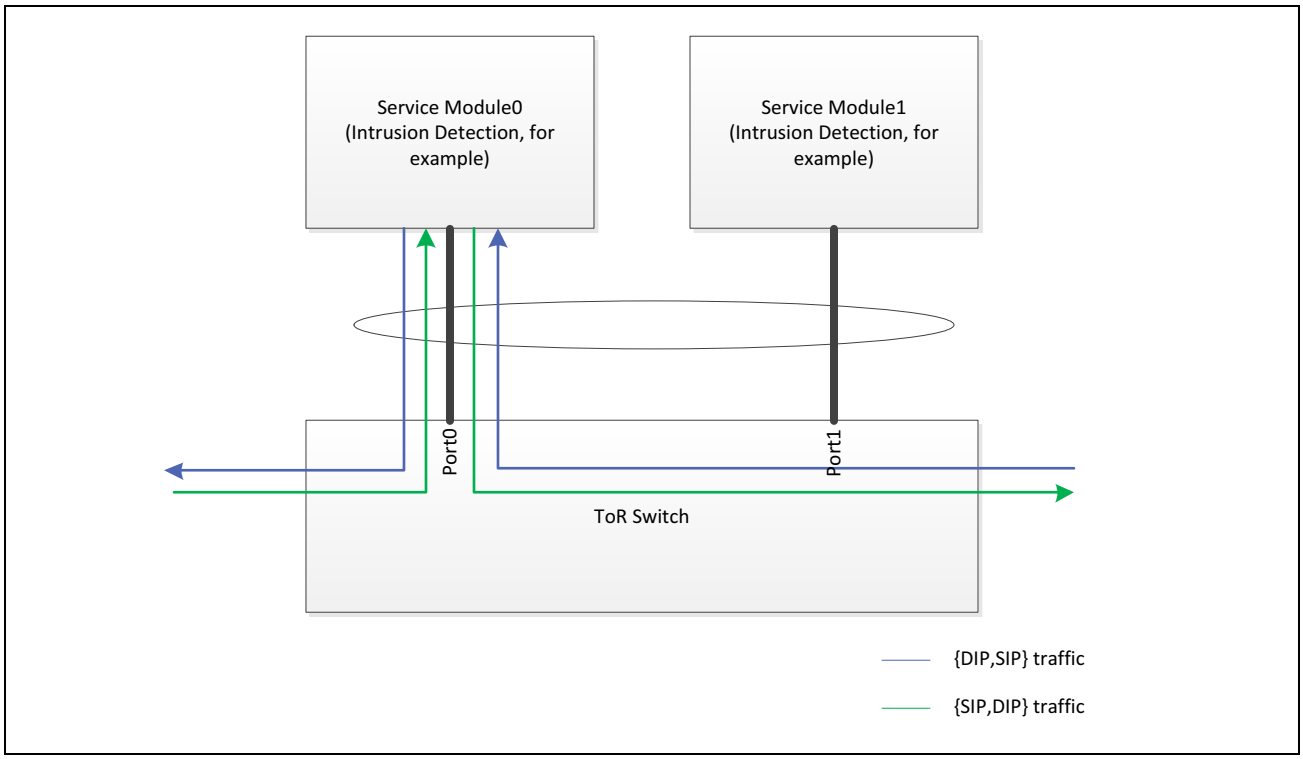

Figure 5: Example Use of Symmetric Hashing Technology

Summary

Traditional static hashing schemes work well for enterprise and carrier networks. The same approach has been adopted for data center networks, yet increased performance requirements prove it is no longer optimal for this type of network environment. Modern data centers are evolving very fast. Without smarter and more sophisticated hashing technologies that provide greater flexibility and network visibility, data centers of today's massive scale may suffer from inefficiencies and performance challenges when deploying new and rapidly evolving protocols. Broadcom's Smart-Hash technology introduces the Resilient Hashing, Flexible Hashing, and Symmetric Hashing enhancements necessary to effectively manage the requirements imposed by current trends in cloud and data center networking. This advanced technology offers an alternative to design approaches based on static hashing schemes that can lead to prohibitively poor application performance under typical data center traffic loads. Cloud network operators are already facing daunting challenges in scaling their network infrastructure to tomorrow's workloads — by understanding alternative hashing options, network operators can select designs that are future-proofed and ready to deliver on critical long-term performance requirements.

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-WP200-R August 2012