Broadcom Smart-NV Technology for Cloud-Scale Network Virtualization

Sujal Das
Product Marketing Director
Network Switching

April 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 SmartScale 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.

Network topology, performance metrics, and feature needs can differ dramatically in private and public cloud networks. Private cloud networks deployed in large enterprises are influenced by legacy use models, whereas public cloud networks are typically greenfield designs, built for cost-effective, equipment vendor-agnostic, commodity-like scaling. The intersection of the two, known as hybrid clouds, is creating new and interesting technologies and usage scenarios.

This white paper explores the network infrastructure virtualization requirements in private and public cloud networks, and how such requirements affect the design of data center network switches. It also describes features that are enabled by Broadcom's Smart-NV (Network Virtualization) technology, part of Broadcom’s SmartScale series of technologies, engineered specifically to meet current feature and scale requirements of private and public cloud networks. Smart-NV encompasses comprehensive best practices for today's high-performance data center switches, and addresses evolving needs of next generation cloud implementations.

Private Cloud Network Virtualization Needs

The proliferation of virtual machines (VM) used in the enterprise has traditionally been driven by the need for higher server utilization and the need to enable dynamic provisioning of IT infrastructure to meet changing business needs. As more and more applications are being virtualized, and the locality of server applications and their associated server hardware are becoming fluid, new network switch feature requirements are evolving. For example, the same virtual server and virtual network infrastructure must now support not only applications that are sensitive to network performance (such as database server and business logic server) but also those that are not (e.g., web server). These requirements affect two facets of network design: network performance, as experienced by applications running in VMs; and application mobility through VM migration and rack-to-rack performance (or performance for east-west network traffic).

Meeting Network Performance Needs of VM Applications

There is an increasing demand for native OS-based server-like performance in VM environments. Unlike native OS-based servers, virtualized servers featuring multiple VMs incur network performance penalties. This is due to the need for a virtual switch (or VEB, Virtual Ethernet Bridge) that runs in software and the need for two network stacks that must be traversed — one in the hypervisor kernel and one in the VM. Several standards-based switching technologies such as SR-IOV (Single Root I/O Virtualization by PCI Special Interest Group) and VEPA/EVB (Virtual Ethernet Port Aggregator/Edge Virtual Bridging by IEEE 802.1) have emerged, including some vendor-initiated proprietary versions. The goal of these network virtualization technologies is to improve the performance and scalability of applications that run in VMs. The implications of these technologies in network switches can be broadly classified as the need for “VM Switching.”
This white paper explores the network infrastructure virtualization requirements in private and public cloud networks, and how such requirements affect the design of data center network switches. It also describes features that are enabled by Broadcom's Smart-NV (Network Virtualization) technology, part of Broadcom's SmartScale series of technologies, engineered specifically to meet current feature and scale requirements of private and public cloud networks. Smart-NV encompasses comprehensive best practices for today's high-performance data center switches, and addresses evolving needs of next generation cloud implementations.

**Increasing Rack-to-Rack Performance**

Web, application, and database server applications running as VMs 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 the network. Such east-west traffic includes server-to-server, server-to-storage, and server rack-to-server rack. This trend is changing the inherent design of network topologies, from oversubscribed and tiered networks to fast, fat, and flat networks. These new topologies require new features in network switches.

**Addressing VM Migration Scale**

Live VM migration is important for increasing server utilization, improving disaster recovery, and meeting overall IT goals of implementing dynamic data centers. Live VM migration across servers, racks, or pods requires that they reside in the same layer 2 (L2) network segments (aka a flat network). To facilitate live VM migration across pods or sites within the data center or even across data centers, L2 networks must scale across racks, pods, and data centers.

In summary, private cloud networks require a network infrastructure design that addresses the VM application and network performance requirements and enables the deployment of fast, fat, and flat networks.
Addressing Private Cloud Networking Needs with Smart-NV

Smart-NV (Network Virtualization), part of Broadcom’s SmartScale series of technologies, effectively meets the feature and scale requirements of private cloud networks.

Addressing Networking Performance for VMs

Network switches must implement VM Switching to ensure adequate networking performance for VMs. Technologies such as SR-IOV and VEPA/EVB are relevant in this discussion (see Figure 1). VM switching consists of the following:

- Access layer switches in the data center network (e.g., top-of-rack switches, blade switches, or end-of-row switches) must support a large number of virtual switch ports (VSPs), directly servicing VMs running on the servers that are connected to the access layer switch. For example, assuming 20 VMs per server and 40 servers per rack, an access layer switch must support about 1K VSPs. An end-of-row switch servicing eight such server racks must support about 8K VSPs.

- Access layer switches in the data center network must also support an adequate number of queues that can be allocated on a per-VSP basis, delivering per-VM quality of service (QoS) and traffic shaping. An ideal implementation would be the availability of thousands of queues that can be dynamically allocated across the VSPs on an as-needed basis.

- VSPs in access layer switches in the data center network must support features such as link aggregation, load balancing, traffic mirroring, and statistics counters as available for physical switch ports. Such features enable VMs with the same level of reliability, performance, and monitoring as physical servers.

- Access layer switches in the data center network must further support an adequate number of ACL (access control list) rules so that they can be applied on a per-VM basis. For example, supporting 10 ACL rules per VM, assuming 20 VMs per server and 40 servers per rack as an illustration, means an access layer switch must support 10K ACL rules. An end-of-row switch servicing eight such server racks must support about 80K ACL rules.

Broadcom’s Smart-NV addresses these VM switching requirements as depicted in Figure 2. Virtual switch ports support link aggregation, queuing, ACL, statistics, and mirroring services similar to how those services are readily available for physical ports.
Implementing Fast, Fat, and Flat Networks

Smart-NV technology, coupled with Broadcom's industry-leading high density, low latency, and 10/40/100GbE line rate performance, serves as an excellent foundation for building fast, fat, and flat data center networks.

Fat implies full cross-sectional bandwidth across racks of servers. Through the very high bandwidth and high port density available in Smart-NV enabled Broadcom switch solutions, nonblocking or low oversubscription ratios (between downlink and uplink ports) can be supported. Nonblocking network switches have higher demands on cost and power per gigabit of bandwidth, a metric effectively met by Smart-NV. When complemented with multipathing of the links, full cross-sectional bandwidth or a fat network topology can be achieved. Multipathing can be achieved in different ways, depending on the design of the network. If the network is an L2 implementation, multipathing can be achieved using new technologies such as TRILL (Transparent Interconnection of Lots of Links) or SPB (Shortest Path Bridging). Smart-NV enables scaling of TRILL networks by supporting a large number of TRILL Routing Bridges. If the network is a layer 3 (L3) implementation, multipathing can be achieved using routing protocols such as OSPF (Open Shortest Path First) and ECMP (Equal Cost Multipathing). Further, to enable the L3-based scaling required by megascale networks, a very large number of L3 forwarding table entries and ECMP routes is supported.
Flat is used synonymously with two network design approaches. Flat is sometimes interpreted as an L2 network (either physical or virtual) that spans across multiple pods or sites within a data center or even across data centers. Smart-NV enabled Broadcom switch solutions can support industry-leading L2 address scaling. TRILL or SPB can be used to create scalable and large flat physical L2 networks without any constraints related to multipathing. When the underlying network is L3, a flat virtual L2 network can be achieved using L2oL3 (layer 2 over layer 3) network virtualization technologies, such as VXLAN (virtual extended LAN), NVGRE (Network Virtualization using GRE), or other similar overlay network technologies. These same high-performance switches can also support industry-leading L2oL3 network virtualization scaling.

Sometimes, flat is also interpreted as a single-tier or a two-tier network, rather than the traditional three-tier network, which consists of the access, aggregation, and core layers. Various “network flattening” technologies, both L2- or L3-based, as described above and supported by Smart-NV enabled Broadcom switch solutions, can be used to build single-tier or two-tier networks. Management of such flat networks is simplified when multiple access and aggregation switches in the network appear as a single switch. Technologies such as VN-tag or IEEE 802.1Qbr, as well as HiGig™, which is available in Broadcom data center switches, can be used to deliver a holistic data plane and control plane solution for flat networks.

Fast, fat, and flat networks can be implemented using high-bandwidth, high-density, fixed-configuration aggregation and access layer switches that are connected in a spine-leaf model. Figure 3 illustrates these highly scalable network designs as supported by Smart-NV technology.
Public Cloud Network Virtualization Needs

Public cloud networks are typically built from the ground up without any legacy equipment or usage model restrictions. As a result, the focus is on designing for scale, with off-the-shelf, easily replaceable, and cost-effective network equipment. The same fast, fat, and flat network topologies required for private cloud networks are equally important for public cloud networks. The same performance demands hold true, including the need for increased VM application performance, increased rack-to-rack performance, and higher VM migration scale. Yet the focus on massive scale and support for multitenancy typically leads to more vendor-agnostic equipment and standards-based design approaches. Some of the key differences include the following:

- **When the focus is on building megascale or Internet-scale data centers, implementations follow the success and scalability of the Internet, built on IP and a scalable hierarchical addressing scheme (rather than a flat L2 addressing scheme).** Such network design in typical public cloud data centers is L3-based. Advantages include address scalability, availability, and applicability of L3-based nonproprietary multipathing technologies such as OSPF and ECMP, as well as ready market availability of off-the-shelf, easily replaceable, cost-effective L3 switches. Use of L2-based technologies is rare, and is found only in small data centers. Use of technologies such as TRILL and SPB is nonexistent because such features are not available in off-the-shelf, easily replaceable, cost-effective network equipment.

- **To address the needs for VM migration at scale and multitenancy, L2oL3-based network virtualization technologies are used.** The physical network infrastructure is built on L3, as explained above; in turn, the natural way to create a flat L2 network for VM migration within and across data centers is through the use of L2oL3 technologies. Options include L2GRE (Layer 2 over Generic Routing Encapsulation), VXLAN (Virtual Extended LAN), or NVGRE (Network Virtualization using Generic Routing Encapsulation). L2oL3-based network virtualization technologies eliminate the VLAN-based scaling challenges (limited up to 4K VLAN IDs) that exacerbate scaling in multitenant networks. They also promise to detach network virtualization-related configuration from physical switches, enabling software-defined networks across multivendor equipment — an attractive benefit for public and hybrid cloud deployments.

- **Finally, in most use cases in the public cloud network, network acceleration technologies such as SR-IOV and VEPA/EVB are either used on a limited basis or are deemed unattractive because these features require specialized (and therefore expensive or vendor-specific) network equipment.** Technologies such as SR-IOV have found limited deployment because they hinder live VM migration technologies and are not fully supported by mainstream hypervisor vendors.

Addressing Public Cloud Networking Needs with Smart-NV

For a subset of the private cloud networking requirements — namely L3 networks — scale, ECMP scale, and L2oL3 network virtualization are key requirements for public cloud data centers. Because of the proven scalability of L3-based networking, large-scale data centers deploy L3 all the way down to the access layer switches. In such an environment, Smart-NV enabled switches support a very large L3 table scale and robust L3 features. Multipathing of the uplinks from these access switches to aggregation switches to create fat links with full cross-sectional bandwidth and load balancing is a critical requirement. Smart-NV technology in turn enables support of several thousands of ECMP routes to enable maximum scale of such fat L3-based networks.
Network Virtualization at Cloud-Scale Using L2oL3 Overlays

To enable network virtualization at cloud-scale, Smart-NV enabled Broadcom switch solutions support new and innovative L2oL3 overlay network technologies such as L2GRE, VXLAN, and NVGRE. Cloud-scale has several requirements, namely:

- It must extend the scale of virtual LANs.
- It must provide VM scale, network partitioning, and hybrid cloud enablement for multitenancy support.
- It must allow efficient VM-based workload placement through live VM migration across pods or sites in a single data center or across data centers.

L2GRE is implemented in the open source Open VSwitch (OVS) initiative. VXLAN and NVGRE are industry collaborations initiated by VMware and Microsoft, respectively, along with their partners. The VXLAN and NVGRE specifications have been submitted as RFC (Request for Comments) drafts to the IETF (Internet Engineering Task Force). Broadcom is a co-author of the VXLAN and NVGRE specifications and is working collaboratively with its partners, the IETF and OVS communities, in the development of its Smart-NV technology in data center switches. Although the packet format and control plane implementations for the three technologies may differ, their impact on Ethernet switches can be grouped together into a common set of requirements. Collectively in this article, the three technologies are referred to as L2oL3 overlay technologies. Figure 4 shows a generic frame format representing the three L2oL3 technologies. The L2oL3 component in the frame broadly represents the VXLAN, NVGRE, or L2GRE header and tunnel ID.

L2oL3 overlay technologies are a great match for large-scale data centers that prefer to use L3 because of the proven scalability of L3 addressing and multipathing technologies. They use the physical L3 network and multiple overlays or tunnels — each of them a virtual L2 network that can then be assigned to a tenant. As shown in Figure 5, such virtual L2 networks can also be used for VM-based applications that require L2-based adjacency, in a stretched cluster, for example. Virtual L2 networks can enable live VM migration across pods, sites, or data centers that are connected using L3 networks.
These benefits are achieved by placing the L2oL3 overlay endpoints, also referred to as tunnel endpoints (TEP), in the virtualized server, or more precisely in the hypervisors running in the virtualized servers. The virtual overlay networks then span across the physical networks, each such overlay network connecting a set of servers and VMs — for example, a tenant group, or an application group — together in the same L2 network segment. The physical network switches in the path must act as L2oL3 Transit Switches for the L2oL3 tunneled or encapsulated packets that traverse them.

The above scenario works well when all end nodes in the data center are capable of being TEPs, and all switches connecting them are partially or fully-featured L2oL3 Transit Switches. In practical deployments, however, there will be end nodes that are not capable of being a TEP — for example, a legacy hypervisor-based virtualized server, a native Linux or Microsoft Windows server, a firewall appliance, a load balancer appliance, and so on. In these instances, L2oL3 TEP Gateways are needed to enable connectivity with such nodes. Similarly, if there are islands of networks that are legacy and do not support L2oL3 overlays, then L2oL3 TEP Gateways can help bridge to such legacy networks.
In Figure 6, a data center network configuration is shown in which some servers are legacy servers (installed with hypervisors that do not support TEP capability), and others are greenfield servers (installed with newer hypervisors that support TEP capability). Also shown are database servers and storage nodes that are not virtualized and are not capable of being TEPs. A cloud of legacy firewall and load balancer appliances is shown connecting to an access layer switch acting as an L2oL3 TEP Gateway. In this example, TOR switches serve as L2oL3 Transit Switches or L2oL3 TEP Gateways. In this implementation, L2oL3 TEP Gateways are placed closer to legacy equipment, but they can also be placed elsewhere in the network.

Figure 7 illustrates a data center network from a public cloud service provider, supporting multitenancy and connecting to data centers belonging to multiple tenants. The use of L2oL3 Transit Switches and L2oL3 TEP Gateways is shown in this typical multitenant hybrid cloud use case. The shading represents a flat virtual L2 segment for Tenant A, enabling live VM migration and other L2-dependent applications to be deployed across the Tenant A and public cloud service provider's data centers. The shading for Tenant B implies the same. In the figure, for the sake of simplicity, a flat virtual L2 segment is shown to consist of a set of servers. Flat virtual L2 segments on a per-tenant basis, however, can be created at VM-level granularity as well.
Next, we discuss hardware level and data plane feature requirements for network switches to support the use cases depicted in Figures 6 and 7. Smart-NV includes support of such features in Broadcom data center switch solutions to enable these and other, more exhaustive use cases. Control plane features for various L2oL3 implementations require the hardware level support found in Smart-NV to enable scaling and performance.

**L2oL3 Transit Switch Feature Requirements**

Figure 4 shows a generic L2oL3 overlay network packet format. For the sake of simplicity, we use this generic packet format for the discussions in the rest of this article.

The L2oL3 packet consists of an outer header, an inner header, and payload. The format varies with respect to the specific contents of the L2oL3 header for each of the L2oL3 specifications, namely VXLAN, NVGRE, and L2GRE. Fueled by Smart-NV technology, the L2oL3 Transit Switch is able, at a minimum, to transport these three types of encapsulated packets, with support for all network switch services.
The inner header on the L2oL3 packet includes VM-specific network addressing and QoS requirements. Smart-NV enables VM-level granularity in the network services, such as class of service (CoS), security, load balancing, link aggregation, and so on, offered by the L2oL3 Transit Switch. The switch should be able to classify and hash, based on the contents of the inner header, gather and apply relevant entropy information to the outer header, and also apply queuing and shaping policies on a per-VM and per-tenant basis. It must accomplish these VM-level granular functions at line rate. For VMs belonging to multiple tenants, adequate levels of isolation and security must be maintained as network services are provisioned and deployed.

Smart-NV also enables network operators to monitor and analyze traffic on a per-tenant basis. The L2oL3 Transit Switch must be able to direct traffic on a per-tenant basis (traffic to and from a set of VMs belonging to a certain tunnel/tenant) to mirroring or analyzer ports.

Certain L2oL3 overlay network control plane implementations use an IP multicast-based address resolution and consolidation mechanism. This use implies that all L2oL3 Transit Switches in the path must support IP multicast features. The number of IP multicast entries supported by the L2oL3 Transit Switch determines the scale of the cloud network in terms of number of tunnels and tenants that can be supported in the network. For example, existing L2 mechanisms (flooding and dynamic MAC learning) can be used to discover remote MAC addresses and MAC-to-TEP mappings; IP multicast can be used to reduce the scope of the flooding to those hosts that expressed explicit interest in the tunneled frames. Every overlay segment may be mapped to a separate IP multicast address, limiting the L2 flooding to those hosts that have VMs participating in the same overlay segment. In a large-scale deployment, many such segments may be mapped or aggregated to a single IP multicast address if the number of IP multicast entries supported by the data center switches is limited. Broadcom switch solutions with Smart-NV technology support the maximum number of tenants per network, using the industry-leading IP multicast table size that is available in its data center switches.

Unlike traditional multicast usage in which information is distributed from a root (source) to a set of branches (receivers), in the case of L2oL3 overlay networks, a tunnel is a multicast tree with VMs as endpoints, and where every VM is both a traffic source and a receiver. As such, multicast trees serving as L2oL3 tunnels must support a scalable and performant PIM bidirectional multicast model. Smart-NV technology enables maximum scaling through hardware support of a highly scalable PIM bidirectional IP multicast mechanism.

### L2oL3 TEP Gateway Feature Requirements

The L2oL3 TEP Gateway must support all L2oL3 Transit Switch features. In addition, the function of the L2oL3 TEP Gateway is to enable connectivity between a new L2oL3 network overlay environment (with tunneled or encapsulated packets) and a legacy or non-network overlay environment. A network switch that sits at the edge of these two environments must support the L2oL3 TEP Gateway features. For packets going from a network overlay environment to a legacy environment, the L2oL3 TEP Gateway implemented in a switch must be able to decapsulate L2oL3 tunneled packets. For packets going from a legacy environment to a network overlay environment, the L2oL3 TEP Gateway implemented in a switch must be able to encapsulate L2oL3 tunneled packets. These packet traversal examples from one environment to another result in some L2 and L3 forwarding situations.
In L2-based forwarding scenarios between the L2oL3 overlay environments and legacy environments, the L2oL3 TEP Gateway implemented in the network switch must support a mapping of the L2oL3 tunnel ID (for example the VXLAN ID or the L2GRE ID) to a corresponding VLAN ID or a switch port and MAC address combination, and vice versa. Sometimes, the L2oL3 TEP Gateway may be needed to bridge between two network overlay segments. In this L2-based forwarding environment, the L2oL3 tunnel IDs used in the two segments are different. Such a situation may arise when the two segments are in two different data centers. In this case, the L2oL3 TEP Gateway, typically sitting at the edge of the data center, must be able to map one tunnel ID to another. At a minimum, a network switch implementing an L2oL3 TEP Gateway must support these L2 forwarding-related features. Maximum L2 forwarding-based connectivity is readily enabled by Smart-NV across such overlay and legacy network segments.

In more sophisticated L3-based forwarding scenarios addressed by Smart-NV technology, the L2oL3 TEP Gateway may terminate tunnel (encapsulated) frames and the L3 segment related to the tunnel. Forwarding outside the L2oL3 tunnel is based on the destination L3 information in the frame. In the reverse direction, the IP packet is encapsulated within the appropriate L2oL3 header. Finally, similar to inter-VLAN routing supported in today’s advanced data center switches, the L2oL3 TEP gateway may implement inter-L2oL3 segment routing.

An important network design consideration with deployment of L2oL3 TEP Gateway-capable switches is where such switches should be placed. For example, such switches may be deployed closer to where new hypervisors with L2oL3 overlay network TEP capabilities are deployed. In this case, connectivity of these new servers with all legacy end nodes and networks can be automatically ensured. In a second deployment example, L2oL3 TEP Gateway-capable switches can be deployed at the edges of legacy end nodes or networks. Figure 6 shows a typical network setting in which there is a mix of new and legacy hypervisor end nodes, and native OS-based end nodes in the server racks. This requires switches to support both personalities, Transit and TEP Gateway. A third deployment example is one in which data centers have only legacy end nodes and L2 networks. L2oL3 TEP Gateway-capable switches are deployed at the edge of these data center networks, where the L2-L3 boundary is at the core layer. This enables flat L2 segments across the data centers for cross data center live VM migration. Flexible deployments are essential; in turn, Broadcom switch solutions featuring Smart-NV technology offer multiple bandwidth and port density configurations for ideal placement in any location of the data center network.

Summary

Network virtualization complements server virtualization in private and public cloud data centers, helping higher ROI and business performance through dynamic resource allocation. Network switches designed for such data centers must support multiple virtualization technologies. Driven by both legacy and new use models, comprehensive network virtualization technologies are needed in enterprise private clouds. Public cloud data centers are designed and engineered for multitenancy, scale, and cost-effectiveness, requiring a subset of such virtualization technologies. Broadcom’s StrataXGS® architecture-based Ethernet switch processors with Smart-NV technology effectively meet all network infrastructure virtualization requirements — offering flexible comprehensive performance and scale for current and next-generation private, public, and hybrid cloud networks.