Enhancing Ethernet network controllers for server virtualization

By Dhiraj Sehgal, Abhijit Aswath, and Srinivas Thodati

Broadcom and Dell are working with a range of virtualization providers to enhance Broadcom® network controllers in Dell™ PowerEdge™ servers to help reduce the bottlenecks and overhead that can impair performance in virtualized environments.

In enterprise IT environments, virtualization offers numerous advantages—enabling organizations to consolidate and share computing resources, increase control and efficiency, and operate their data centers cost-effectively. Getting the most out of a virtualized environment, however, requires more than simply deploying virtual machines (VMs). An effective networking infrastructure is essential.

Like other hardware resources in Dell systems, Broadcom network adapters are virtualized to the VM. Virtualization platforms use a hypervisor-based architecture—also known as a VM monitor (VMM) architecture—that hides the physical characteristics of the computing platform and allows unmodified VMs to run concurrently on host platforms.

This design means that virtualization comes at the cost of reduced performance. Today’s virtualization architectures include VMs with device drivers, I/O stack, and applications running on top of a virtualization layer that includes device emulation, I/O stacks, and physical device drivers that manage the Ethernet network controllers. This virtualization layer adds overhead and degrades system performance, driving up processor utilization and reducing available bandwidth.

Broadcom, a networking leader in Gigabit Ethernet (GbE) and 10 Gigabit Ethernet (10GbE) network controllers, has been working closely with a variety of virtualization vendors—including VMware, Microsoft, Citrix, and Dell—to address these problems. Broadcom controllers are included across a range of Dell PowerEdge servers as LAN on Motherboards (LOMs), mezzanine cards, and standard network interface cards (NICs), and are used as part of Dell’s overall virtualization solutions. Broadcom is currently working in a two-phase effort to enhance its controllers for use in virtualized environments, helping organizations using Dell systems to maximize the benefits of their virtualization initiatives.

**Phase 1: Removing bottlenecks and increasing performance**

In the first phase of enhancements, Broadcom worked to remove virtualization bottlenecks and increase system performance by providing a number of additional features. For example, Broadcom Ethernet network controllers now support stateless offloads such as TCP checksum offload, which enables network adapters to compute TCP checksums on transmit and receive, and TCP large send offload, which allows the TCP layer to build a TCP message up to 64 KB long and send it in one call down the stack through IP and the Ethernet device driver, saving the host processor from having to compute the checksum in a virtualized environment.

In addition, Broadcom’s jumbo frame support in virtualized environments is designed to save processor utilization by reducing interruptions, and to increase throughput by allowing the system to concentrate on the data in the frames instead of the frames around the data. However, because of the...
The single-threaded nature of the hypervisor in processing I/O and the duplicate I/O copies in the virtualization layer, performance is still limited with this approach. Broadcom also supports the VMware® NetQueue and Microsoft® Hyper-V™ VM queue (VMQ) features, helping remove such single-queue bottlenecks and avoiding the need for stateful offloads such as TCP offload. And the Internet SCSI (iSCSI) host bus adapter (HBA) mode in Broadcom controllers, supported as part of Dell virtualization solutions, can provide excellent performance in virtualized environments.

Use of multiple queues

The trend toward increased processor core density is leading to the use of an increased number of VMs, which requires additional processor cycles to route packets to the VMs. Using hardware queues provided by the network controller, virtualization vendors have avoided the single-thread limitation of a traditional OS and optimized the hypervisor for multiple hardware threads.

In both the VMware and Microsoft Hyper-V platforms, packets must traverse the hypervisor or parent partition because a direct path is not available between the controller and the VMs. On egress, packets are first copied from the originating VM for processing in the virtual switch. The destination Media Access Control (MAC) address and virtual LAN (VLAN) ID are looked up to determine the route, and the packet is then copied to the receive queue of the other VMs and/or submitted to the network driver for transmission. On ingress, packets are indicated to the switch, which uses the destination MAC address and VLAN ID to determine which VM or group of VMs the packets can be copied to.

Route lookup, data copy, and filtering tasks represent additional processor load and latency absent in the non-virtualized environment. The associated overhead can significantly affect networking performance, especially at 10 Gbps; this problem is addressed by offloading these tasks into a network adapter, where the transport queue manager can transmit packets from multiple queues and steer the receive packets into multiple queues. The VMware NetQueue and Microsoft Hyper-V VMQ features enable Broadcom controllers to take on these tasks, thereby helping meet the demands of bandwidth-intensive applications that require high levels of performance and networking throughput in virtualized environments.

Storage offload

Networked storage is crucial in virtualized environments, enabling the smooth migration and failover of a VM from one physical server to another. iSCSI has emerged as a high-performance, accessible networked storage technology popular in many virtualization deployments. Broadcom NetXtreme II® iSCSI HBA functionality, with support for VMware, Microsoft Hyper-V, and Citrix® Xen® virtualization, provides the converged functionality needed in virtualized environments by offering on-chip processing to help free up processor resources and increase bandwidth and performance, as shown in Figure 1. Broadcom GbE and 10GbE iSCSI HBA functionality as shown in this figure is enabled by default in the VMware vSphere™ 4.1 platform on Dell PowerEdge servers.

Broadcom iSCSI HBA functionality enables on-chip processing of the iSCSI protocol (and of the TCP and IP protocols), helping free up host processor resources at 10 Gbps line rates over a single Ethernet port. This functionality provides extended performance benefits that help meet the demands of bandwidth-intensive applications requiring high-performance block storage I/O for the hypervisor, servicing all instances of the VM.

iSCSI boot

iSCSI boot allows a server to boot an OS over a storage area network (SAN), avoiding the need for local disk storage—a primary source of computer system failures. In addition to enhancing system reliability, the use of diskless servers helps simplify the IT administrator’s workload by centralizing the creation, distribution, and maintenance of server images; reducing the overall need for storage capacity through increased disk capacity utilization; and adding increased data redundancy through the use of data mirroring and replication.

As the use of SANs continues to grow in virtualized environments, and as administrators continue to recognize the advantages of moving local storage from individual servers to centrally managed storage arrays, network...
Networking

boot options such as iSCSI boot are poised to become an increasingly common feature within the data center and throughout the enterprise. Broadcom, VMware, Microsoft, and Citrix are working to create simple yet richly featured iSCSI boot options that use iSCSI to replace local storage in virtualized environments. Broadcom GbE and 10GbE iSCSI boot functionality is enabled by default in VMware vSphere 4.1 on Dell PowerEdge servers, and Broadcom iSCSI HBA mode and iSCSI boot for 10GbE controllers in hypervisors are both supported on Dell platforms.1

**Phase 2: Integrating SR-IOV enhancements**

The second phase of Broadcom’s enhanced virtualization effort, now under way, focuses on Single Root I/O Virtualization (SR-IOV); these enhancements are expected to be available in Dell systems using Broadcom NetXtreme II 57712 controllers. SR-IOV-capable Ethernet network controllers are designed to enhance I/O throughput and reduce processor utilization while increasing scalability and devices’ sharing capabilities. SR-IOV allows the direct I/O assignment of an Ethernet network controller to multiple VMs, helping maximize the network adapter’s full bandwidth potential.

The PCI Express (PCIe) SR-IOV specification for PCIe I/O virtualization and sharing technology is the basis for the SR-IOV implementation in SR-IOV-capable Broadcom controllers. The specification defines an extension to the PCIe specification that enables multiple system images or VMs to share PCIe hardware resources (see Figure 2). The Broadcom SR-IOV device presents a physical function (PF) having multiple virtual functions (VFs). A VF is a lightweight PCIe function, and resources associated with the main data movement of the function are available to the VM. The VF can be serially shared between different VMs—assigned to one VM, and then reset and assigned to another. In addition, a VF can be transformed from a VF into a PF.

Comprehensive support for PCIe SR-IOV involves both enhancing existing capabilities and adding new capabilities to the platform and OS. Network controller device drivers supporting SR-IOV also must be re-architected to support additional communication paths between PFs and VFs.

**VMware VMDirectPath**

High throughput and low latency are especially important in a distributed system, wherein the nodes’ I/O latencies affect the performance of the cluster and the overall application. Low latency is necessary to preserve data coherency in large database clusters implementing scalable SR-IOV network adapters.

Using the VMware VMDirectPath network plug-in architecture and an SR-IOV-capable Broadcom device, a VF can be directly assigned to a VM—yielding near-native performance, avoiding additional I/O copying in the hypervisor, and supporting a comprehensive array of virtualization features, including VMware vMotion™ technology and Microsoft Hyper-V live migration (see Figure 3). Direct assignment of PCI devices to VMs is necessary for I/O appliances and high-performance VMs.

With dynamic VMDirectPath or Uniform Pass-Through version 2 (UPTv2), the device interface is

---

1 For more information on iSCSI functionality in Broadcom adapters for Dell systems, see “A unified networking approach to iSCSI storage with Broadcom controllers,” by Dhiraj Sehgal, Abhijit Aswath, and Srinivas Thodati, in Dell Power Solutions, 2010 Issue 2. dell.com/downloads/global/power/ps2q10-20100411-broadcom.pdf.
split in two, enabling pass-through of performance-critical operations such as TX/RX producer index registers, interrupt mask registers, and emulated infrequent operations in the management driver running in VMware ESX. To implement live migration, the VF is acquiesced and switched to emulation mode from pass-through mode, allowing the minimal device state to be checkpointed or restored. Most of the state lives in the VM memory, and the guest operating systems are unaware of the migration.

Support for dynamic VMDirectPath requires re-architecting the OS platform and network device driver. The VMware platform implements a network plug-in architecture allowing pass-through of performance-critical parts by partitioning the vmxnet driver to include a VM-specific shell and hardware-specific module or network plug-in driver. The VM-specific shell implements the interface to the OS network stack and interacts with the hypervisor for configuration. The hardware-specific network plug-in driver interacts with hardware in the data path and uses the VM shell interface for OS-specific calls. VMware ESX controls the network plug-in used by the shell to load the plug-in into the VM based on the VF and to map the VF into VM address space.

**Effective and efficient virtualization**

The network is a key component of virtualized environments, and network controllers that are enhanced to meet the demands of virtualization can help maximize performance. Broadcom’s virtualization-focused efforts enable organizations using Dell systems to take advantage of offload technologies and flexible, real-time I/O—enhancements that facilitate effective and efficient virtualization. Broadcom capabilities in SR-IOV and I/O pass-through functionality for Ethernet network controllers with TCP and iSCSI offload can provide near-native performance and reduced latency. Looking ahead, Broadcom plans to continue to support and enable virtualization on Dell server platforms, helping IT organizations to accelerate their adoption of virtualized infrastructures while increasing performance and controlling costs.

**Dhiraj Sehgal** is a senior product line manager for Ethernet controllers at Broadcom.

**Abhijit Aswath** is a senior product line manager for Ethernet controller software at Broadcom.

**Srinivas Thodati** is a senior product marketing manager for PowerEdge M-Series servers at Dell.

**NIC embedded switch**

I/O virtualization and sharing are also required for point-to-point and switch-based configurations, enabling interoperability between VMs, VFs, chipsets, switches, endpoints, and bridges. A Broadcom NIC embedded switch enables Ethernet switching between VMs, from VF to VF, and to or from external ports.