

Virtual Switching in Solaris

Nicolas Droux
nicolas.droux@sun.com

Sun Microsystems, Inc.
Solaris Networking

April 2, 2007

Note: This document will be part of a larger VNIC design document, yet to be published.

1 Introduction

Virtual NICs, also known as VNICs, allow physical NICs to be shared by multiple Zones or virtual machines such as Xen domains. VNICs appear to the rest of the system as regular NICs. VNICs can be assigned a subset of the hardware resources (interrupts, rings, etc) made available by the underlying hardware.

In order to provide connectivity between the multiple Zones or virtual machines sharing a single physical NIC, the VNIC layer also provides a data-path between the VNICs defined on top of the same underlying NIC. This data-path is needed since switches will not loop a packet back to its originating port. The VNICs sharing the same underlying NIC appear to be part of the same segment, i.e. connected to a same *virtual switch*. The mapping between physical and virtual network components is illustrated by Figure 1.

The following sections describe the concept of virtual switches, the semantics they implement, and how they can be used in practice.

2 Virtual Switching with VNICs

Virtual switches are never directly accessed or visible by the user or system administrator. Instead, they are implicitly created when the first VNIC

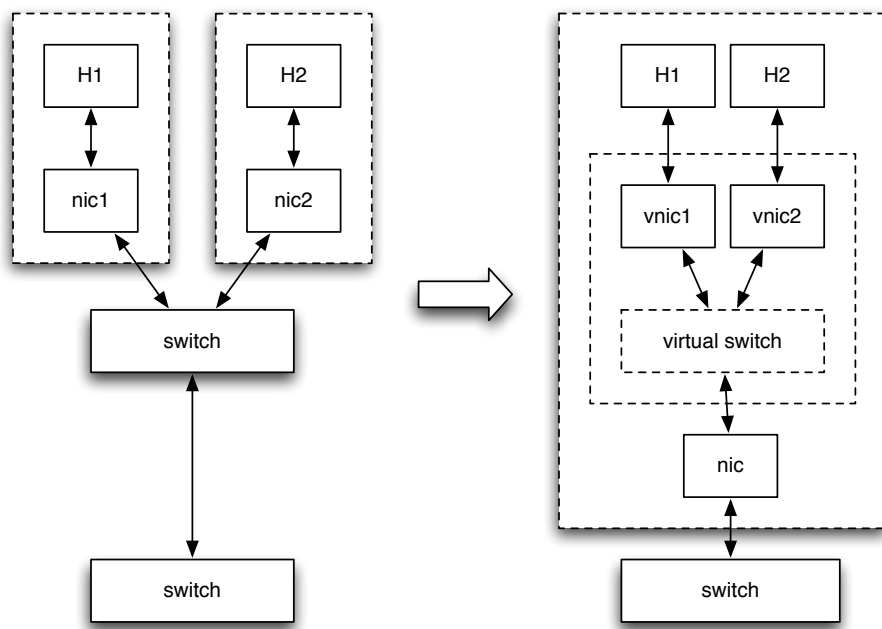


Figure 1: Mapping between switches and virtual switches.

is defined on top of a NIC. Virtual switches provide the same semantics expected by physical switches:

- VNICs defined on top of the same underlying NIC can send ethernet packets to each other using their MAC addresses
- Broadcast packets received by the underlying NIC are distributed to every VNIC defined on top of that NIC. Similarly, broadcast packets sent by one of the VNICs is sent to all VNICs defined on top of the same NIC (except of course the originating VNIC), and to the underlying NIC for transmission.
- Multicast group membership is tracked, and used to distribute multi-cast traffic to the appropriate VNICs.

Connectivity is only enabled between VNICs defined on top of the same underlying NIC. This is similar to having multiple hosts connected to multiple physical switches, as represented by Figure 2. Each VNIC (except for anchor VNICs described in Section 3) is part of a VNIC group. VNICs are part of the same group if they share the same underlying NIC, and the connectivity is allowed only between VNICs that are part of the same group.

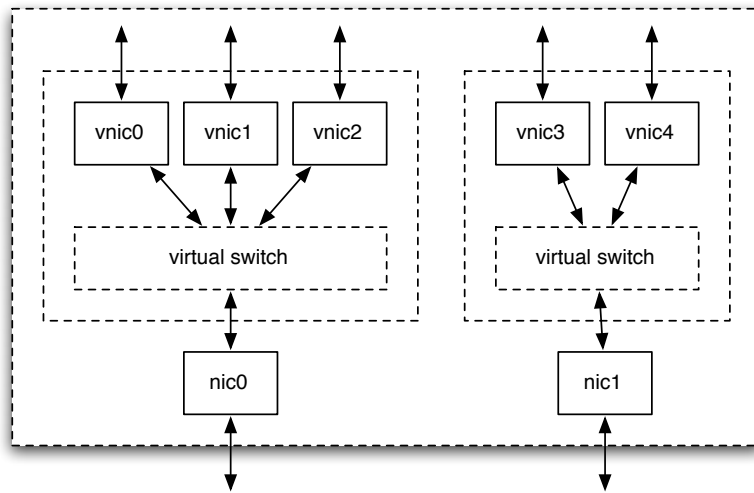


Figure 2: VNIC and virtual switches.

In practice, the virtual switch semantics are implemented by the transmit and receive entry points of the VNIC driver. On the transmit path, classification is done from the VNIC transmit entry point using the destination MAC address of the packet. The result of the classification is used to direct the traffic to the appropriate destination, which can be another VNIC defined on top of the same underlying NIC, or a group of VNICs in the case of a multicast or broadcast packet. If there is no classification match on the outbound packet, the packet is for an external host and it is passed to the underlying NIC for transmission.

On the receive path, packets are classified by the hardware classifier of the underlying NIC, or by the MAC layer software classifier. After classification, the callback function associated with the matching entry is invoked. The callback is responsible to pass the packet to a single VNIC in the case of a unicast MAC address, or to a group of VNICs in the case of a broadcast or multicast destination MAC address.

In order to facilitate the adoption of VNICs on existing configuration, the VNIC layer allows the underlying NIC to be plumbed while being used by VNICs. In order to maintain connectivity between the VNICs and the plumbed interface on top of the same underlying NIC, the plumbed interface appears to be connected to the same virtual switch used by VNICs. For example, a NIC can be used to create VNICs and assign them to non-global zones, while allowing the underlying NIC to be plumbed directly by the global zone, and connectivity is provided between the vNICs and the plumbed interface.

3 Anchor VNICs

As described in the last sections, the VNIC layer provides the virtual switching capabilities that allow the VNICs to communicate with each other. The VNICs are always created on top of an underlying NIC, and these associations define the allowed local data paths between the VNICs. These local data paths provide the same semantics as a layer 2 switch to which the VNICs are virtually connected.

In some cases, it is desired to be able to create virtual networks on the same machine without the use of a hardware NIC. Some examples are described in section 4. Since a virtual switch is implicitly created when a VNIC is created on top of a NIC, one solution would be to have the ability to create pseudo NICs, and define VNICs on top of these pseudo NICs.

Since VNICs implement the behaviors of regular NICs, it is possible

to implement such pseudo NICs as a special type of VNIC. These special VNICs are called *anchor VNICs*. Anchor VNICs are created and deleted using the same interfaces as those used to manage regular VNICs.

VNICs can be created on top of either physical NICs, or anchor VNICs. When multiple VNICs are implemented on top of the same anchor VNIC, they can send traffic to each other through the virtual switch corresponding to the anchor VNIC. Anchor VNICs are different than regular VNICs in a number of ways:

- Anchor VNICs are not associated with an underlying NIC.
- Anchor VNICs do not send or receive data.
- Anchor VNICs do not have a MAC address.

Since anchor VNICs provide a subset of the VNIC functionality, they can be easily supported by the common VNIC layer. Figure 3 shows how VNICs (vnic0, vnic1, vnic2) can be connected to a virtual switch on top of an anchor VNIC (vnic5) without the need of dedicated hardware.

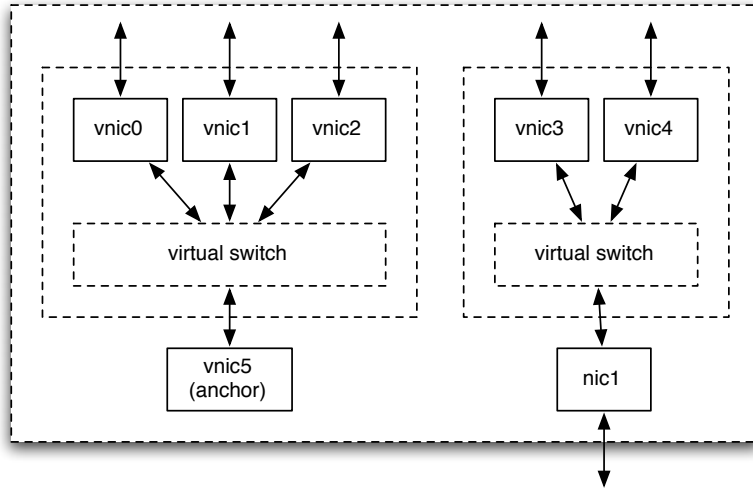


Figure 3: VNIC and virtual switches with an anchor VNIC.

4 Example

Figure 4 represents a zones environment where a virtual network is created between zone1-zone4. That virtual network is defined by creating VNICs vnic1-vnic4 on top of the same anchor VNIC vnic5. In addition, the physical NIC bge0 is assigned to zone1. With this configuration, the virtual network anchored by vnic5 is kept separate from the physical network. zone1 can use IPfilter, NAT, and DHCP to assign the addresses to vnic2-vnic4, and allow the permitted traffic to be exchanged between zone2-zone4 and the physical network attached to bge0.

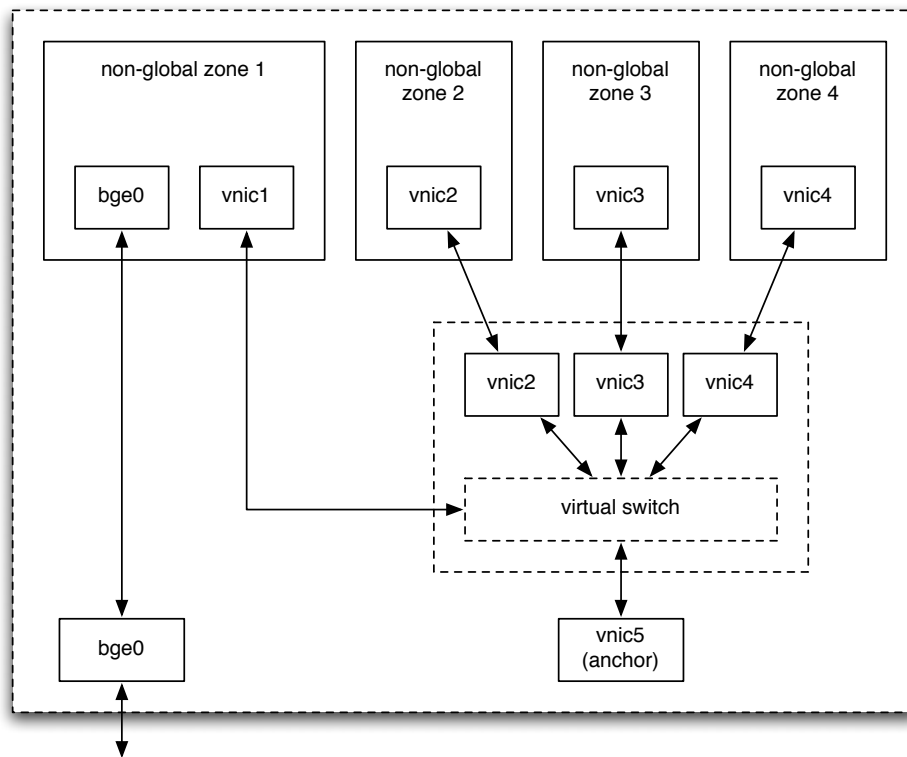


Figure 4: VNIC and virtual switches with an anchor VNIC.