ORACLE®
SOLARIS

An Oracle Technical White Paper
November 2011

Oracle Solaris 11 Network Virtualization and Network Resource Management

ORACLE®

Executive Overview

This paper describes the wide reaching re-architecture of Oracle Solaris 11 network stack to bring three key advantages:

- Network virtualization—Provide a fully virtualizable network environment for more effective sharing of networking resources and increasing the scope for server consolidation projects.

- Network resource management—Provide networking resource management capabilities to allow organizations to meet quality of service goals for networking.

- Increased Performance—Provide a parallelized data path enabling increased scalability on a large number of processors, cores, and hardware threads.

Introduction

The network stack for Oracle Solaris 11 has been substantially re-architected from that of Oracle Solaris 10 in an ambitious effort, called the Crossbow project, to accomplish the three goals listed above. This section provides more detail about those goals and how Oracle Solaris 11 enables them.

Network Virtualization

The addition of facilities for network virtualization enhances the ability to consolidate server workloads. In the networking equipment market the term 'network virtualization' typically has a limited reference and focuses on specific issues like Virtual LANs and aggregation techniques. In this paper, network virtualization is a more general abstraction in which all aspects of a network topology are virtualized within a server virtualization framework.

There are many virtualization aspects of the Crossbow project:

- Virtualizing the hardware Network Interface Controller (NIC) into Virtual NICs (VNICs) provides the direct benefit of more effective sharing of networking resources. The VNIC construct allows dividing a physical NIC port into multiple virtual interfaces to create kernel-enforced isolated and dedicated network stacks from physical interface to application.

- The network virtualization enhancements for Oracle Solaris 11 also allow virtual switching between VNICs without depending on any networking hardware, enabling the construction of virtual network topologies within a single Oracle Solaris instance.

- Virtualization is also applicable for link aggregation. By aggregating two physical NICs and configuring multiple VNICs over the aggregation, network resources are shared more efficiently. In addition, redundancy assures continued network availability even if one of the physical links fails.

- The industry standard Virtual LAN (VLAN) construct is supported, allowing NICs and/or VNICs to be assigned to a VLAN. Thus, in an environment with switches and routers that support VLANs, end–to-end traffic can be isolated even though the traffic may be running on a shared physical link.

Additionally, other networking elements can be brought into play, particularly the router, firewall, and the latest addition, a load balancer, all of which are included with Oracle Solaris 11. Taken together, these elements enable the configuration of an entire network topology within one physical server that can be used for architecting/prototyping, testing and even deployments. Figures of sample configurations are provided later in this paper.

Network Resource Management

Network resource management allows organizations to meet quality of service goals for networking. In Oracle Solaris 10, resource management is implemented by specifying guaranteed resource levels and resource maximums to system processes. Administrators control usage of CPU and memory resources to ensure that designated applications and/or zones do not exceed usage limits that have been defined. At the same time, these applications and zones can get minimum resource levels regardless of the demand for those resources. The new Oracle Solaris 11 network stack architecture extends resource management to networking in two ways:

- Specific CPU resources can be assigned to a NIC port or Virtual NIC such that greater CPU resources are allocated to high priority and high bandwidth traffic while more limited resources are assigned to low priority traffic. This capability functions in conformity with Oracle Solaris Zones and CPU pools. If a CPU pool is assigned to a zone, then VNICs defined for that zone will inherit the same pool, and CPU resource limits placed on the VNIC will be from that CPU pool.

- Bandwidth limits can be set for a NIC port or Virtual NIC. This capability is most useful in ensuring that one interface does not exceed its expected use of the network, and negatively impact other traffic. Bandwidth limits should be assigned with some care to ensure that the sum of the VNIC bandwidths reasonably matches the physical bandwidth of the underlying port.

These networking resource management capabilities enable creation of enforceable organizational network sharing policies. These policies can be enforced by data center operations' staff. There is no requirement for Oracle Solaris 11 network resource management awareness to be built into the applications.

Increased Performance

The changes made to the networking stack through project Crossbow can increase network throughput by parallelizing and better controlling the scheduling and handling of packets.. The best performance gains typically come with the latest generation intelligent NICs that enable packet classification and have multiple receive and transmit ring buffers that the new networking stack code can manage. There are many aspects of the design that facilitate increased efficiency, but one of the major ones is how it deals with inbound packets. See the Architecture section below for details.

Virtualization

Oracle Solaris 11 network virtualization and network resource management also enhances the value of Oracle Solaris virtualization technologies for server consolidation.

Operating System Virtualization

Oracle Solaris Zones[1] are a virtualization technology that allows one operating system instance to offer multiple virtual isolated OS environments. The key advantage of this approach is that while applications see an environment that looks like a dedicated OS, in reality multiple Oracle Solaris Zones can all run on a single instance of Oracle Solaris 10 (or Oracle Solaris 11). As a result, zones, when compared to hypervisor virtualization technologies, can make much more efficient use of system resources because one OS oversees the CPU, memory, and network resource allocation. Zones also have excellent scaling properties because of the extremely small system overhead they place on the OS. And finally, creation and destruction of zones is a lightweight task that facilitates their use in dynamic environments. For example, for highly secure environments using Oracle Solaris Trusted Extensions (built into Oracle Solaris 10 and 11), the separation of security levels is accomplished through zones.

In Oracle Solaris 10, applications running in zones could access network interfaces. However, if a dedicated network stack for the zone's network interface were desired in Oracle Solaris 10, a dedicated physical NIC or dedicated VLAN would then be required. With the networking changes for Oracle Solaris 11, a zone can be assigned as many Virtual NICs as needed and each will have it's own dedicated stack whose bandwidth and CPU allocation can be managed.

With zones and network virtualization, one could consolidate multiple servers (and services) on to one instance of Oracle Solaris. See the section below for scenarios that illustrate examples involving zones.

Hypervisor Virtualization

Oracle VM Server for SPARC is a virtualization technology for T-Series CMT SPARC systems based on a hypervisor architecture. Each guest domain is assigned its networking interfaces (vnets) through the management interface. The Oracle VM Server for SPARC network virtualization architecture is tightly integrated with Oracle Solaris 11 virtualized data link layer. Zones as well as Oracle Solaris 11 VNICs and flows can also be used from within guest domains.

Architecture

From the preceding it should be clear the impact the Oracle Solaris 11 networking architecture has on network virtualization and network resource management. This section now turns to the details about the architecture to enable readers to understand how key features are delivered and why the new

---

[1] Known as Oracle Solaris Containers in Oracle Solaris 10.

Oracle Solaris 11 network stack is about more than just virtualization. The new architecture impacts the entire network stack and particularly the data link layer to provide the foundation for the next generation network stack.

The fundamental building blocks of this new architecture are Virtual NICs or VNICs—a construct for dividing a physical NIC into multiple virtual ones. A VNIC device is accessed, from the applications viewpoint, exactly like a physical NIC. The characteristics of a VNIC—the bandwidth and what CPU resources are assigned to handle it—can be dynamically controlled.

The new networking stack is designed as a fully parallelized network stack structure. Think of a physical network link as a road. The new stack design allows dividing that road into multiple lanes. Each lane represents a hardware classification of packets, and is architected to be independent of other lanes. There are no common queues, threads, locks, or counters. Tying this architecture to a modern NIC is illustrated in Figure 1.
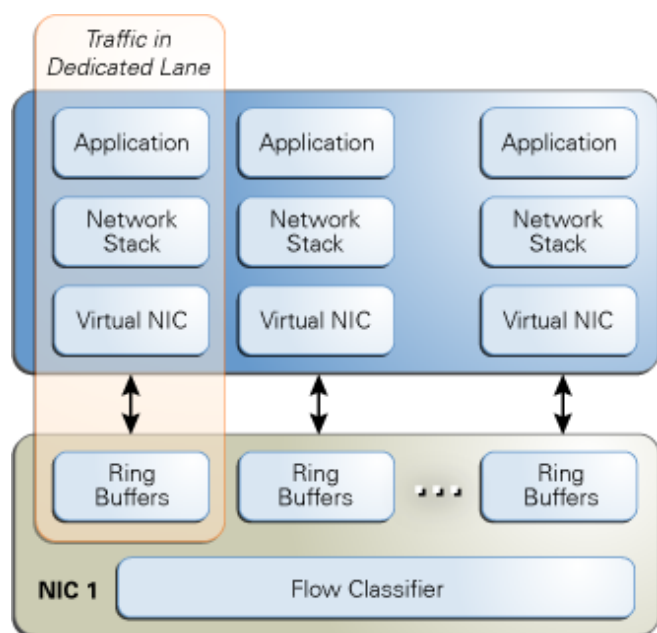

Figure 1. VNIC traffic lanes can map to their own specific ring buffers in a modern NIC.

A key element of the design is hardware classification. In the case of Figure 1, the classification is done in the NIC. Once the incoming packet has been classified, it enters it's own private lane indicated by the shaded area. The NIC has a Transmit (Tx) and Receive (Rx) ring buffer dedicated to each lane, or more specifically to the MAC address of the VNIC to which it belongs. When an Rx ring is full, packets are dropped. Of course, packets are generally not dropped and it will be shown below that if the load is so great that packets must be dropped, it's much better to drop them in the NIC rather than have the OS expend CPU resources to decide to drop them.

On the transmit side, the buffers facilitate the ability to support parallel DMA transfers, allowing multiple CPU threads to queue packets for transmission by the hardware. This gives an example of how Tx scaling for SPARC systems is achieved.

Not every NIC has a hardware classifier, and even if it did, the number of lanes required by the OS may be more than can be supplied by the NIC. For that reason, the new network stack design also includes a software layer for dealing with NICs that do not have multiple on-board Transmit and Receive buffers and filtering capabilities. See Figure 2 below. For the purposes of this paper, NICs without those capabilities will be referred to as "dumb" NICs to differentiate from the latest generation "smart" NICs.

Figure 2. Traffic lanes with classification by the OS.

In the case shown in Figure 2, all the architecture for supporting VNICs is part of the operating system kernel, which includes a software classifier with dedicated Transmit and Receive rings to create the lanes entirely in software. This approach allows supporting applications with requirements for 20 VNICs, for example, on a system with a classifier that only feeds 16 hardware rings. The software enables creating real lanes from the hardware as well as virtual ones that share a real lane.

Virtual Network Resource Management Capabilities

Network resource management is another important aspect of the new network stack architecture and the ability to specify the characteristics of these lanes is an important element of the design. As mentioned above, Oracle Solaris 11 facilities allow administrators to set bandwidth limits and assign the appropriate number of CPUs to handle the traffic.

These resource management capabilities tie cleanly to the architecture. Here are two examples:

• Pushing the control of the flow of packets close to the NIC is important. In traditional flow control implementations, just bringing a packet into an operating system queue from which it will potentially be dropped expends a great deal of the total processing cost of packet handling. If the flow is

metered at the NIC, there is less impact on the system from a dropped packet. For NICs that manage their own Tx/Rx rings, dropped packets incur no CPU overhead.

• Other Quality of Service (QoS) implementations are typically a layer inserted into the network stack—a choke point that doesn't take into account contention for resources elsewhere. With Oracle Solaris 11 resource management, the resource management capabilities are designed into the stack architecture. It is possible to specify resource management controls for an application based on CPU, memory, and networking resources, enabling a holistic enforcement of QoS requirements. This is particularly useful in an Oracle Solaris Zones environment where it makes good policy sense to set caps on resource levels to minimize the chance that applications in one zone will negatively impact other zones by using too much of the network resources.

Network resource controls are not limited to VNICs. They can also be applied to other data links, such as physical NICs, link aggregations, IP over IB (InfiniBand), and so on. Network resource controls can also be applied on a more fine grain basis than "all the traffic through a VNIC or data link." To identify a subset of traffic, Oracle Solaris 11 introduces the concept of a "flow." A flow allows identifying a subset of traffic based on a range of Layer 3 or Layer 4 attributes, including:

• Source or destination IP address

• Protocol

• Port

Operationally, this means that all HTTPS traffic could be assigned a higher bandwidth than that of default traffic, FTP traffic could be assigned a lower bandwidth, or communication to a specific IP address could be given a different bandwidth than the default. One can also use flows for accounting purposes, such as to track the history of certain types of traffic.

The architecture also supports monitoring capabilities. By enabling history logging, the amount of traffic on a link can be monitored for capacity and other planning purposes. If a flow has been defined, further drill down is possible.

One other element of the architecture is worth highlighting—the way the new stack architecture manages interrupt handling. In low utilization mode, packets are handled in the traditional interrupt manner. This is an adequate approach for lower speed traffic and provides the best latency. In high speed (10GbE), high load networks, interrupts can have a negative effect on overall system throughput. There are various techniques to mitigate this, but fundamentally an interrupt per packet is simply not efficient on a busy network. The re-architected network stack automatically switches from interrupt mode to polling mode when the packet arrival rate exceeds a threshold. Polling has a key advantage for busy networks. One poll by the driver can potentially return a chain of many packets in one operation, which is far more efficient than one packet per interrupt.

Virtual Networking Scenarios

The value of virtual networking as described above, should be apparent. The combination of virtual NICs and the ability to manage those resources makes an excellent match to virtual server

technologies. For example, the ability to carve up a 10Gb or 1Gb physical interface into smaller 'lanes' and assign those to an Oracle Solaris Zone is compelling because the operating system enforces the bandwidth and/or network CPU resources assigned, preventing one zone from using more network resources than expected.
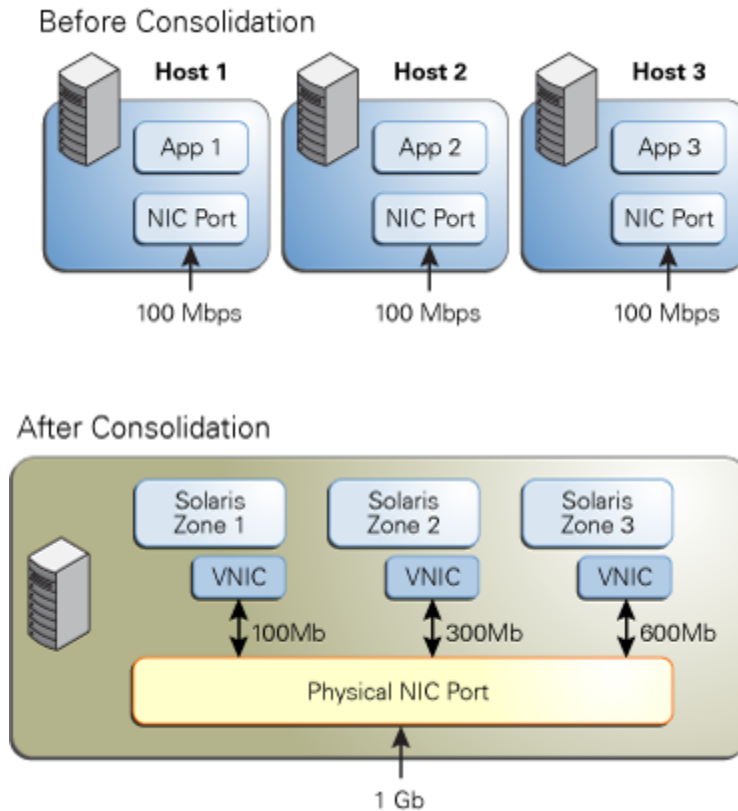

Figure 3 - Simple Virtualization Scenario

In the case of Figure 3, three systems and their applications have been consolidated onto one server running Oracle Solaris 11 and using three Oracle Solaris Zones. The consolidated server was moved to a 1 Gigabit per second (Gbps) network. Note that for two of the Zones, the same bandwidth limits prior to consolidation are enforced after consolidation. For the third zone, a new higher speed VNIC is provided. Many different bandwidth scenarios are possible. The new network stack architecture in Oracle Solaris 11 gives an administrator much more control of the network resources while enjoying the advantages of server consolidation.

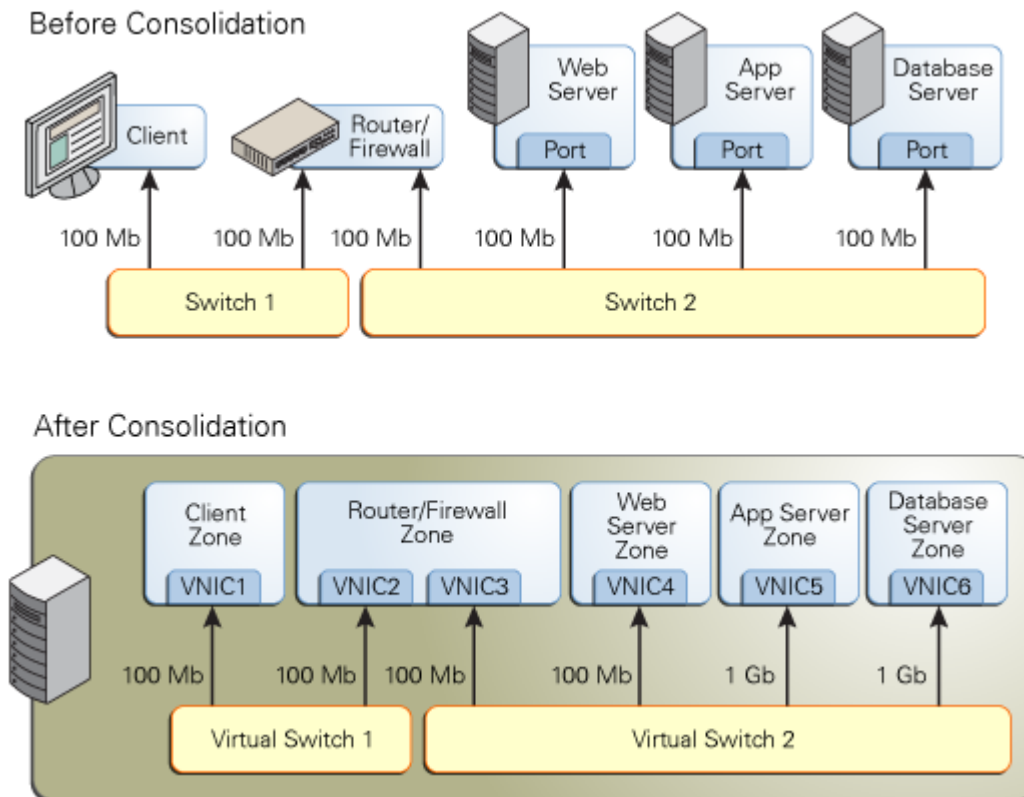A more sophisticated virtualization project is pictured in Figure 4.

Figure 4 – Sophisticated Virtualization Scenario

In this classic three tier architecture case, one host runs a Web server, another the application server, and the third runs the database. One client is shown. A developer might consolidate this environment onto one system to work on some element of the interaction between these three systems. The testing organization might also use this consolidation to simplify the hardware requirements for testing the entire application environment. Note this scenario adds two new virtualization elements, a switch and a router. The virtual switch makes it possible for the zones to communicate directly between each other. The virtual router is the open source Quagga project, included with Oracle Solaris 11. Oracle Solaris 11 also includes the IP Filter firewall, new load balancer, plus bridging support and VRRP functionality. These features can be combined with virtual networking features to build virtual networking topologies.

As in the first scenario, resource management makes many different solutions possible. Some examples include:

• A new architecture where the App Server and the Database Server would communicate via a 1Gb link

• Matching the bandwidths of the pre-consolidation environment

• A production environment that includes some additional external network connections and thus has no reason to put any bandwidth limitations on the inter-server communication

Simulating topologies such as a three-tier architecture on one system with hypervisor-based virtualization would not be as efficient due to the "per virtual machine" resources required. However, thanks to the low overhead of zones, as well as the resource management and virtualization features of Crossbow, such generic virtual topologies can be easily deployed on a single host with limited physical resources.

A final example (illustrated in Figures 5 and 6) focuses on network resource management controls. As mentioned, these controls do not need to be built into the application. For example, suppose that a data center is using network tape backups. It is not uncommon for those services to soak up an inordinate amount of network bandwidth.

Figure 5 illustrates a network before a tape backup commences, with traffic flowing between Server 2 and the other servers. The size of the colored arrows showing the flows between servers represents the relative percentage of network bandwidth consumed. In the case of Figure 5, the consistent size of the colored arrows represents that bandwidth is evenly distributed for traffic between the different servers.
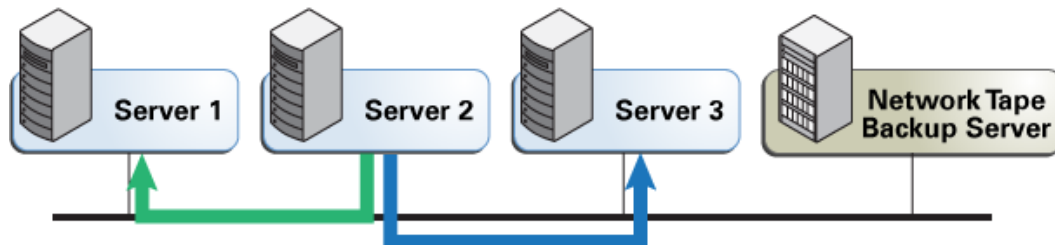

Figure 5 – Before backup starts

After starting the backup of Server 2, the situation may be as pictured in Figure 6 where much of the bandwidth is being soaked up by the backup (the thick arrow) and the other traffic is negatively impacted (much thinner lines).
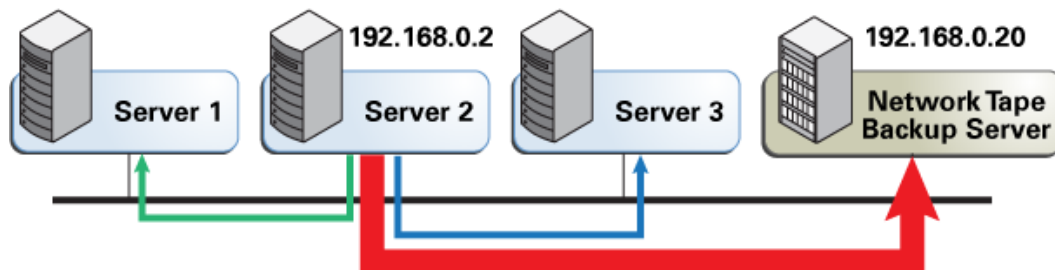

Figure 6 – After Backup starts

With network resource management, the operations personnel could address this problem by defining what Oracle Solaris calls a "flow" and then categorizing the traffic by a filter. In this case traffic between 192.168.0.2 and 192.168.0.20 would be restricted so as not to adversely affect other server traffic. Once the flow has been defined, the next step is to set the bandwidth for the flow. In this case, the operations personnel would dial down the bandwidth until they observe normal traffic levels as pictured below in Figure 7.
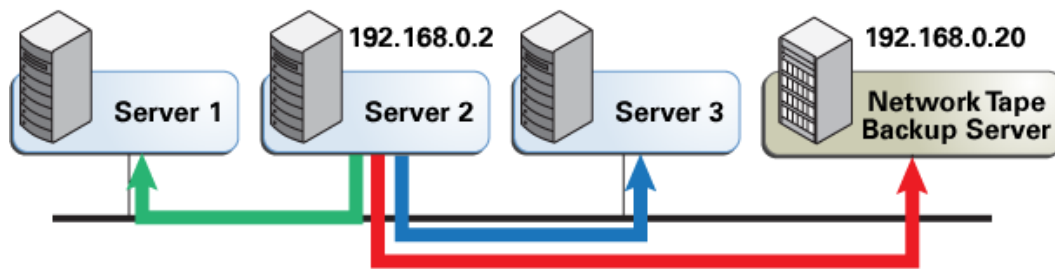
Figure 7 – After applying network resource management controls

To accomplish this resource management task, the application did not have to be modified in any way. Flow controls can be applied to the NIC port (or VNIC) by the operations staff through a simple command line operation.

Network Resource Management as a Cloud Enabler

One of the key themes for Oracle Solaris 11 is "Built for clouds." Zones and resource management clearly play a large role in creating an environment conducive to building public or private clouds. Zones offer lightweight, no overhead virtualization, application isolation, and resource management. Resource management provides the control points to both set resource limits as well as enable creating tiered resource offerings for CPU, memory and network bandwidth and network I/O processing. Other less obvious cloud advantages from the networking perspective include

- Automatic binding of network I/O resources (interrupts, kernel threads) of a VNIC or data link to the CPU pool assigned to a zone allows for improved separation between zones. The network processing for a zone is performed by the CPUs of the zone, and also allows for improved performance, by co-localizing the network processing of a zone with the application that is sending or receiving that traffic.

- Fine grain real-time and historical observability allows the metering of each zone's actual network usage for billing purposes.

- Networking for zones can be configured so that all the information about the network interfaces is contained within the zones, facilitating the ability to move them without undue restrictions on the configuration details of the receiving system.

Summary

Adding the Oracle Solaris 11 capabilities for network virtualization and network resource management represents the next step in the evolution of the Oracle Solaris networking stack. Network resource management gives data center operations the ability to control CPU, memory, and networking resources to meet quality of service goals. Network virtualization adds a critical piece of the system virtualization story by providing the ability to virtualize not only servers but their networking topology, including firewalls, routers, load balancers, and switches. And finally the new Oracle Solaris 11 network

stack offers an architecture that dovetails with current trends in NIC design to extract the best performance from those devices.

Other Resources

- How to get started creating Oracle Solaris Zones in Oracle Solaris 11
http://www.oracle.com/technetwork/articles/servers-storage-admin/o11-092-s11-zones-intro-524494.html

-How to Script Oracle Solaris 11 Zones Creation for a Network-In-a-Box Configuration
http://www.oracle.com/technetwork/articles/servers-storage-admin/o11-118-s11-script-zones-524499.html

- How to restrict your application Traffic Using Oracle Solaris 11 Network Virtualization and Resource Management
http://www.oracle.com/technetwork/articles/servers-storage-admin/o11-095-s11-app-traffic-525038.html

-Oracle Solaris 11 Networking Virtualization Technology – All Resources
 http://www.oracle.com/technetwork/server-storage/solaris11/technologies/networkvirtualization-312278.html

-

ORACLE®

Oracle Solaris 11 Network Virtualization and
Network Resource Management
November 2011, Version 1.0
Author: Jeff McMeekin

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200

oracle.com

Oracle is committed to developing practices and products that help protect the environment

Hardware and Software, Engineered to Work Together