

Oracle Solaris Networking

Oracle Solaris 11 Express

Introduction

In today's information driven times the network is the lifeblood of the system, providing it with the connections and the data to allow the applications to do their work. With Oracle Solaris 11 Express, Oracle takes another big leap forward in networking technologies providing a reliable, secure and scalable infrastructure to meet the growing needs of today's datacenter implementations. In one step Oracle Solaris 11 Express can become the foundation for an open networking platform or allow the building of "datacenters in a box" providing all the savings that consolidation, flexibility and TCO savings can bring.

Oracle continued innovations across the networking stack, including the introduction of an outstanding new network stack architecture, fit into the following five areas:

- Virtualization
- Performance
- Observability
- Security
- Configurability

Virtualization

Networking has always had a part to play in terms of virtualization, the adoption of VLANs into almost every network infrastructure is self evident. However with the advent and adoption of virtualization technologies, for example Oracle Solaris Zones, there is the opportunity for network virtualization to move further within the servers and operating systems themselves.

Oracle Solaris 11 Express, introduces a new network stack architecture also known as the Crossbow project. This architecture allows, amongst other features, the creation and use of Virtual NICs (VNICs). A VNIC allows the division of a physical NIC (or for that matter an aggregation of physical NICs), enabling the sharing of network resources. In addition to allow the virtualization of a physical NIC through VNICs, the new network stack also provides virtual switching between VNICs, and allows the creation of virtual switches (known as 'etherstubs') that are completely independent from the underlying hardware.

The power of VNICs has been tightly integrated with Oracle Solaris Zones. An Exclusive-IP Zone provides a complete, tunable, manageable and independent networking engine to each zone; such a zone can configure DHCP, IP Routing, IP multipathing, IPsec, and more. However, in Oracle Solaris 10, exclusive-IP zones have a fundamental limitation: you need to dedicate a physical network interface (NIC) to the zone. Oracle Solaris 11 Express resolves this limitation by introducing the ability to create high performance VNICs used by zones. The result is that you can create an unlimited number of exclusive-IP zones, even on a system with a single physical NIC. In addition this capability is also extended to Oracle Solaris 10 Containers (the ability to run Solaris 10 instances inside a branded zone on top of Oracle Solaris 11 Express).

The integration into virtualization continues with the introduction of Infiniband IPoIB support for zones. The integrations allows multiple IPoIB instances to be created on the same IB port/P_Key. Now each zone could talk to each other by utilising their own IPoIB instances.

These features combine to become extremely useful in the Virtualization space. For example in an environment where there are different end users or customers on the

Oracle Solaris Networking

Oracle Solaris 11 Express

same system, Oracle Solaris 11 Express allows the building of independent network infrastructures, allowing you control over the environment and your customers the flexibility they need to be successful.

Performance

The Oracle Solaris 11 Express network architecture also delivers performance. This nicely complements the low overhead, high performance, high scalability capabilities already present in Oracle Solaris Zones. Just like virtualization, these new networking features are built-in. In addition, these features are designed to work together simplifying administration with fewer, integrated commands and working together as efficiently as possible.

The new network stack architecture introduces network resource management. This allows setting bandwidth limits and assigning the number of CPUs to handle the traffic. For example, a 10Gb or 1Gb physical interface could be bandwidth limited into smaller 'lanes', these constrained lanes could then be assigned to different zones. The operating system will enforce the bandwidth and/or network CPU resources assigned, preventing one zone from using more network resources than expected. For instance you may want to make sure that a specific zone gets a higher proportion of network traffic, maybe it is servicing your OLTP traffic.

Complementing resource management, Oracle Solaris 11 Express also introduces the idea of network flows. A flow can be defined for a subset of the network traffic, for example traffic for a specific service or IP destination(s). This becomes particularly powerful when used in combination with resource controls. A good example here is periodic backup operations. Before the advent of bandwidth limitations and flow control it would have been possible for the monthly backup operation to 'swamp' the

external network to the detriment of other, more important, applications. By defining a flow from say, the source server IP address to the backup server IP address, and applying resource management to that flow the undesired behavior is easily prevented.

In addition, the sockets implementation has been improved and no longer uses STREAMS. This not only means performance improvements but also a new, simplified developer interface for adding new socket types. The architecture also keeps an eye on network traffic volume allowing it to shift from interrupt driven to polling mode which is much more efficient when dealing with high network traffic volumes.

The Infiniband stack in Oracle Solaris 11 Express has also undergone some significant improvements. Lending itself to the success of the Oracle RAC and Oracle Exadata products the Infiniband stack has improved Sockets Direct Protocol (SDP) allowing the transparent redirection of TCP/IP usage to SDP and the efficiencies that brings. It also now implements the RDSv3 protocol providing better performance and observability for Oracle RAC databases.

Observability

It's very good to have a flexible and highly performing network infrastructure but also critical is the ability to observe that environment. Specifically with Oracle Solaris 11 Express users are now able to use or develop common packet sniffing tools such as etherreal, snoop and wireshark to view all IP traffic sent on real and virtual paths. All network traffic can be observed, including IP packets that are looped back in the IP stack and traffic flowing through Oracle Solaris Zones. Administrators of Zones can also observe packets from within Zones.

In addition in Oracle Solaris 11 Express the **d1stat** tool provides statistics information for data links, for example

Oracle Solaris Networking

Oracle Solaris 11 Express

detailed per-VNIC usage of a physical NIC, all the way down to per-hardware rings statistics. Also newly introduced the **flowstat** tool provides detailed, real-time statistics for flows. The **dladm** and **flowadm** commands are used to provide configuration information for datalinks and flows respectively. Now you can clearly see not only how your network is performing but how it is configured.

Security

As already described Oracle Solaris 11 Express adds the capability to create one or more fully virtualized and dedicated network stacks per Oracle Solaris Zone. This provides flexibility to the end user but does not prevent the guest itself from inadvertently misbehaving, for example sending harmful packets onto the network. Link protection provides additional security by denying some basic outbound threats such as IP, DHCP, MAC and L2 frame spoofing.

The dedicated network stacks also allow administrators to create fine grained network security policies on a per zone basis. Examples are the configuration of zone specific IP routing, DHCPv4 and IPv6 stateless address configuration, IP filter and NAT configurations and IP security (IPsec) and Internet Key Exchange (IKE) automating the provision of authenticated keying material for IPsec security associations.

Finally with Automatic Secure by Default, network services are disabled by default or set to listen for local system communications only.

Configurability

Oracle Solaris 11 Express doesn't stop there, the networking stack has also been improved to allow Oracle Solaris 11 Express to be highly versatile in it's abilities.

We have introduced support for the Virtual Router Redundancy Protocol (VRRP) which provides high availability to our integrated L4/L3 load balancing and also our integrated router technologies.

Improved datalink management now allows vanity naming, **dladm** support for legacy DLPI drivers, unified driver property configuration and simply makes the life of administrators that much easier.

IP Multipathing (IPMP) has been re-architected, providing improved administration and observability. Now it can work transparently with all IP based applications including core technologies such as DHCP.

The network driver framework now has a kernel plugin mechanism that provides an architecture for implementing distinct MAC layers such as Ethernet, WiFi, Infiniband, and IP tunnels. In addition GLDv3 core driver APIs are now available for use by 3rd Party driver writers.

Network Auto-Magic (NWAM) simplifies and automates network configuration by allowing users to automatically discover and connect to networks based on their network conditions and profiles.

And finally in the Infiniband space, new storage protocols including SRP target (SCSI RDMA protocol) and iSER (iSCSI enhanced by RDMA) have been introduced. We have implemented Ethernet on InfiniBand (EoIB) a new protocol to support using the our new gateway products. With this protocol, a host can use the remote 10Gb Ethernet NICs in the Nano-Magnum 2 Gateway. Open Fabric User Verbs, a port of the OFED 1.3 Linux user space libraries, commands and utilities has been completed allowing support of new OS-bypass IB applications. Also the port provides many new commands and utilities that are familiar to Linux InfiniBand users.