ORACLE

SOLARIS

An Oracle White Paper
September 2010

Highly Available and Scalable Oracle RAC Networking with Oracle Solaris 10 IPMP

ORACLE

Overview

Businesses world wide rely on network connectivity in order to service clients, disruption in network services can lead to loss of revenue, and loss of productivity. Business continuity and minimizing planned downtime involves ensuring that the application and systems are designed for resilient to failures. Highly available networking infrastructure could be leveraged by configuring multiple physical paths to the network to eliminate single points of failure. Oracle Solaris supports high availability at the network layer out of the box. Oracle Solaris network high availability options include IP multipathing (IPMP) and Link level aggregation.

Oracle RAC configuration has specific networking requirement, choices for IPMP configuration will be discussed for addressing those requirement, and where Oracle RAC with Clusteware can take benefits of them in seamless fashion. This paper contributes to the existing Oracle RAC and IPMP knowledge base by bringing out the best practices for configuring a fault resilient IPMP network for Oracle RAC with Oracle Clusterware.

Introduction

Oracle Solaris IPMP ships with robust failure detection and fail-over mechanism for network interfaces on a system that are attached to the same link. Oracle Solaris IPMP supports both IPv4 and IPv6 IP addresses.

Failure Detection Mechanism

Link-based

In the Link-based failure detection mechanism network failure detection is always enabled. Physical NICs with network failure detection support are used in link-based configuration. When the link fails or the state of link changes, NIC driver informs changed state to above networking subsystem. Its configuration is a very simple, and would have no additional IP addresses to manage.

A list of supported NIC drivers for the current release of Oracle Solaris 10 can be found at http://docs.sun.com/app/docs/doc/816-4554/emqqq?l=en&a=view .

Probe-based

Probe-based mechanism inherits link-based failure mechanism. The 'in.mpathd' daemon performs probe-based failure detection on each interface in the IPMP group that has a test address. Probe-based failure detection involves the sending and receiving of ICMP probe messages that use test addresses. These messages go out over the interface to one or more target systems on the same IP link.

The 'in.mpathd' daemon determines which target systems to probe dynamically. Routers that are connected to the IP link are automatically selected as targets for probing. If no routers exist on the link, 'in.mpathd' sends probes to neighbor hosts on the link. A multicast packet is sent to the all hosts multicast address, '224.0.0.1' in IPv4 and 'ff02::1' in IPv6, determines which hosts to use as target systems. The first few hosts that respond to the echo packets are chosen as targets for probing. If 'in.mpathd' cannot find routers or hosts that responded to the ICMP echo packets, 'in.mpathd' cannot detect probe-based failures.

This is useful even in scenarios where the NICs are not supported with link failure detection. It attempts to detect failures of any link in between until a probe reaches the target IP address. Configuration of probe-based IPMP group involves planning a set of test IP addresses, configuring and managing them.

IPMP Groups

Both of the failure detection mechanisms allow configuring active-standby and active-active IPMP groups. Only one failure detection mechanism can be configured at a given time.

IPMP daemon in.mpathd monitors the configured NICs to take the appropriate action. On platforms that support Dynamic Reconfiguration (DR) of NIC's, IPMP can be used to transparently fail over network access, providing uninterrupted network access to the system.

Active-Standby IPMP group

In the event of a NIC failure, the network driver recognizes the changed link state and propogates the information to sub-system. IPMP responds to the change by triggering fail-over process of IP address from the failed NIC to a healthy NIC in an IPMP group.

Once the healthy NIC is back online, and the 'FAILBACK' property is set to 'yes' in the /etc/default/mpathd configuration file then the fail-back process is triggered. If the standby NIC fails, IPMP does not take any action on the active link.

Active-Active IPMP group

Active-Active IPMP group is scalable and load spreading configuration since both the NIC IPs can be leveraged to communicate with different clients. This IPMP group achieves load spreading by having two links enabled. If one of the NIC fails, the IP fails-over to the surviving NIC, and fails-back in the event of the recovery of the failed NIC.

If the 'FAILBACK' property is set to 'no' then there is no fail-back. On the recovery of failed NIC it will be part of the active-active IPMP group without any IP assigned to it.

For Oracle RAC public networking, link-based active-active IPMP group is configured with 'FAILBACK' option set to 'no' in /etc/default/mpathd.

Oracle Real Application Cluster (RAC)

Oracle RAC is a highly available database where data is accessed by many Oracle instances. In an Oracle RAC environment, two or more systems concurrently access a single database. This allows an application or user to connect to either system and gain access to a single coordinated set of data. The cache fusion mechanism of Oracle RAC ensures a consistent view for all data consumers.

VLAN tagged interfaces

VLAN (Virtual Local Area Network) tagged NICs are used to over-come the problem of limited available NICs on a system. VLAN tagged interfaces are created using VLAN tags, there by a given NIC will be part of different VLANs on network. These VLANs need to be configured on a given port of a switch for the traffic to pass through. Most of the layer 2 supported switches would offer VLANs.

For example, if the physical NIC is nxge0, VLAN tagged interface name would be nxge131000. Here 'nxge' is the NIC driver, '131' is the VLAN tag and '000' is the interface. This shows that on a given system there can be one virtual vlan tagged NIC for a given network. However there could be many virtual IPs hosted on the same VLAN tagged interface and those VIPs would be hosted as

nxge131000:1 (where :1 is the first interface which hosts a VIP). These VLAN tag NICs are configured with exclusive-IP type of Oracle Solaris Containers. They offer an TCP/IP stack, there by Oracle CRS/clusterware can plumb/unplumb VIP. These VLAN tag NICs are as independent of physical NICs in its configuration. Plumb/Unplumb of these NICs doesn't have any impact on the actual physical NIC.

Oracle RAC Networking with Oracle Solaris IPMP

This section details the public and private networking of Oracle RAC and how it leverages IPMP groups using VLAN-tagged interfaces. There are no configuration and operational differences in IPMP for physical NICs and VLAN-tagged interface usage.

Public networking

Oracle RAC needs one IP address per node is called Virtual IP (VIP). VIP is typically part of the same TCP/IP network as that of the hosts IP. Oracle Clusterware plumbs VIP on each node before it brings up other resources such as Oracle listeners.

For VIP high availability, Oracle Clusterware depends on underlying operating environment to provide high availability of network. In the event of complete VIP failure on a system, Oracle Clusterware brings down the system and the VIP and fails over to other healthy system.

A single NIC failure is handled by IPMP group within a system, entire IPMP group failure is handled by VIP service of Oracle Clusterware.

Private networking

Oracle RAC uses one private IP address per node for both cluster heartbeats and Oracle RAC database instance communication. IPMP group provides highly available NIC for the cluster private interconnect traffic. IPMP offers availability without any performance overhead.

Oracle RAC can be configured to leverage one or more than one NIC for its private network to scale the private network traffic by using active-active IPMP group configuration.

Make sure that the interface(s) at the RDBMS/RAC layer are as defined in the installation process, are the interfaces monitored by the CRS HA framework. If they are redundant interfaces, as with IPMP, make sure that the IPMP interface is used by the RDBMS and is monitored by CRS. Identify this in a number of places, RDBMS from v$cluster_interconnects, from the alert.log, and from AWRs. From the CRS perspective in the ocrdump, the css log file and oifcfg output has all the private interconnect configuration details.

In the event of a NIC failure, IPMP manages the availability of the private IP by failing over to surviving NIC. Entire IPMP group failure is handled by Oracle Clusterware, as Oracle Clusterware stops sending and receiving heartbeat over the private interconnect, it triggers an action to manage the state.

Use cases of IPMP with Oracle RAC

This section covers the different IPMP failure scenarios, and highlights the link-based active-active IPMP group configuration as a best available option.

Best performing and simple IPMP configuration

There are two different IPMP groups possible with link-based failure detection mechanism, namely link-based Active-Standby IPMP group and link-based Active-Active IPMP group which is simple and easy to configure.

Active-Standby configuration and Active-Active configuration can be configured with one IP. For quick fail-over process, configure active-active IPMP group with one IP., It operates like Active-Standby configuration. As other link is active, in the event of a failure of a healthy NIC, fail-over takes place quickly without changing the NIC state.

Disabling the 'FAILBACK' option turns down the fail-back operation in the event of recovery of failed NIC. Link-based Active-Active IPMP group with disabled 'FAILBACK' option reduces the operating overhead of IPMP for Oracle Clusterware environment.

Hence link-based Active-Active IPMP group configuration is best suitable for Oracle RAC private and public networking.

Advanced IPMP configuration

There are two different IPMP groups possible with probe-based failure detection mechanism, namely probe-based Active-Standby IPMP group and probe-based Active-Active IPMP group which is an advanced configuration, as it involves configuring multiple test IPs for each active NIC. As probe-based failure detection mechanism sends ICMP packets over the network to detect the failure, adds additional overhead as compare to link-based mechanism. At the same time offers failure detection of remote network devices.

Consider using the probe-based mechanism for Oracle RAC's public network when high availability of remote network equipment matters. On the other hand consider probe-based mechanism for Oracle RAC public and private networking when the underlying physical NIC does not support link-based failure.

Example configuration

The following example configuration demonstrates the link-based active-active IPMP configuration with Oracle RAC. Setup and configuration details are given below.

Setup

This study was carried out on Sun SPARC T5220 servers running Oracle Solaris 10 (Release 10/09) and Oracle DB 10gR2 (10.2.0.4). The configuration described in this study is applicable to T-series, M-series, and x86 based systems running Solaris 10. The configuration can be used without any changes in both Container and non-virtualized environment called global zone of Oracle Solaris operating system. Furthermore, this configuration can be used on both physical and VLAN-tagged interfaces.

Figure 1 illustrates an example setup used in the configuration of link-based active-active IPMP group for Oracle RAC public and private network.

As shown in Figure 1, a link-based active-active IPMP group 'pub_ipmp0' is configured to manage the high availability of public network. It has got e1000g131000 and e1000g131001 VLAN-tagged interfaces assigned. The Container's IP is brought up by its network services, and IPMP manages it for high availability. VIP is brought up by Oracle Clusterware before bringing up its dependent services and IPMP manages high availability of VIP.
Another link-based active-active IPMP group 'priv_ipmp0' is configured to manage the high availability of private network. It has got e1000g111002 and e1000g111003 VLAN-tagged interfaces assigned. The Container's network services brings up the private IP address as the container is brought up, IPMP offers high availability. Oracle RAC leverages this hosted IP address on 'priv_ipmp0' IPMP group. Fail-over process of this IP address within the IPMP group is transparent to the Oracle RAC private network.

The fail-back option is disabled in the '/etc/default/mpathd' to reduce the down times to the fail-back operation.

Example configuration uses VLAN-tagged interfaces, replacing VLAN-tagged interfaces with physical NICs does not change any IPMP group configuration, except VLAN-tagged interface names has to be replaced with physical NIC.

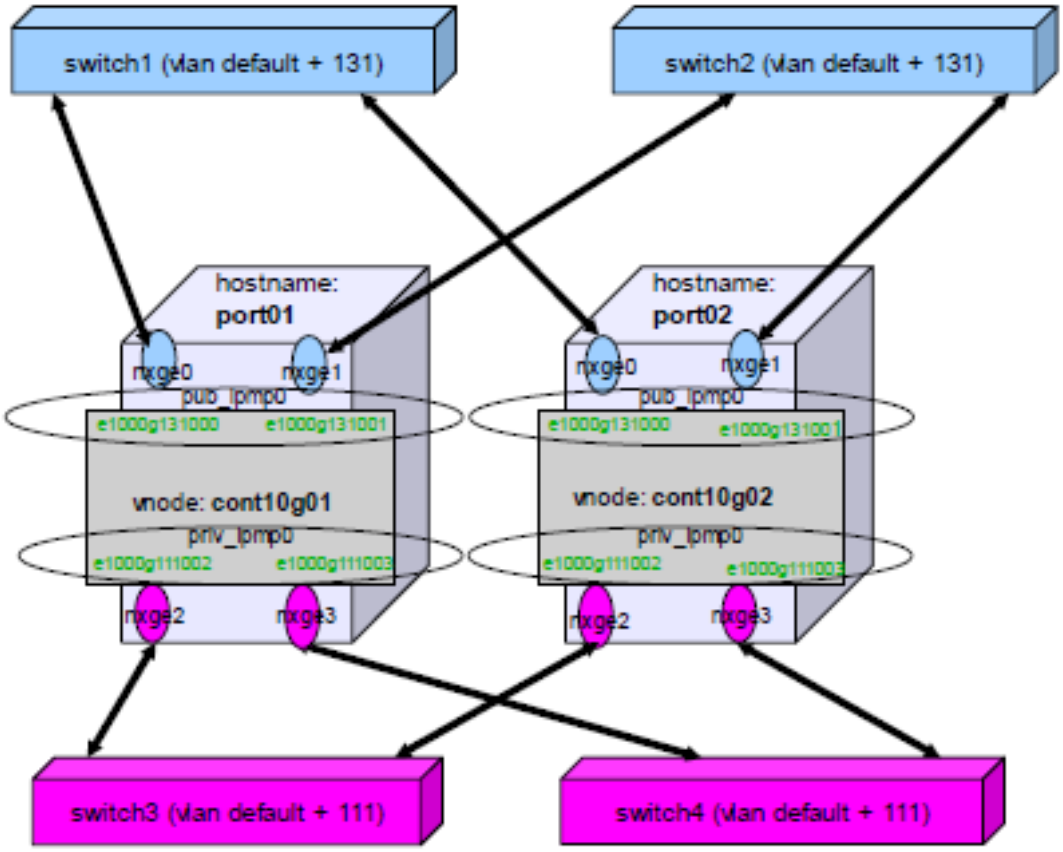

Figure 1: Oracle RAC Networking setup with IPMP

Configuration

Configure node1 hostname "cont10g01"

Public IPMP group: "pub_ipmp0":

Physical NIC: e1000g0 has vlan 131 tagged NIC on each container as e1000g131000

Physical NIC: e1000g1 has vlan 131 tagged NIC on each container as e1000g131001

1. 1 IP as physical node's IP hostname: cont10g01 IP: 199.199.131.101 NIC: e1000g131000
2. 1 IP for the Oracle vip hostname: cont10g01-vip IP: 199.199.131.111 NIC: e1000g131000:2
3. No IP for the other active NIC hostname: N/A - IP: N/A NIC: e1000g131001

Private IPMP group: "priv_ipmp0":

1. Physical NIC: e1000g2 has vlan 111 tagged NIC on each container as e1000g111002
2. Physical NIC: e1000g3 has vlan 111 tagged NIC on each container as e1000g111003
3. 1 IP as private NIC hostname: cont10g01-priv - IP: 199.199.111.101 - NIC: e1000g111002
4. No IP for the standby NIC hostname: N/A - IP: N/A - NIC: e1000g111003

Configure "cont10g01"

create/edit the following files:

```
::::::::::::::
/etc/hostname.e1000g111002
::::::::::::::
cont10g01-priv group priv_ipmp0

::::::::::::::
/etc/hostname.e1000g111003
::::::::::::::
group priv_ipmp0 up

::::::::::::::
/etc/hostname.e1000g131000
::::::::::::::
cont10g01 group pub_ipmp0
::::::::::::::
/etc/hostname.e1000g131001
::::::::::::::
group pub_ipmp0 up


::::::::::::::
/etc/hosts
::::::::::::::
#
# Internet host table
#
::1     localhost
127.0.0.1       localhost
199.199.131.101 cont10g01       loghost
199.199.131.102 cont10g02
199.199.131.111 cont10g01-vip
199.199.131.112 cont10g02-vip
199.199.111.101 cont10g01-priv
199.199.111.102 cont10g02-priv

root@cont10g01:/# svcadm restart network/physical

This is how it looks post configuration on "cont10g01":

root@cont10g01:/# ifconfig -a
root@cont10g01:/# ifconfig -a
lo0: flags=2001000849<UP,LOOPBACK,RUNNING,MULTICAST,IPv4,VIRTUAL> mtu 8232 index 1
        inet 127.0.0.1 netmask ff000000
e1000g111002: flags=201000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4,CoS> mtu 1500 index 2
        inet 199.199.111.101 netmask ffffff00 broadcast 199.199.111.255
        groupname priv_ipmp0
        ether 0:21:28:4f:5d:a2
e1000g111003: flags=201000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4,CoS> mtu 1500 index 3
        inet 0.0.0.0 netmask ff000000 broadcast 0.255.255.255
        groupname priv_ipmp0
        ether 0:21:28:4f:5d:a3
```

```
e1000g131000: flags=201000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4,CoS> mtu 1500 index 4
        inet 199.199.131.101 netmask ffffff00 broadcast 199.199.131.255
        groupname pub_ipmp0
        ether 0:21:28:4f:5d:a0
e1000g131000:1: flags=201040843<UP,BROADCAST,RUNNING,MULTICAST,DEPRECATED,IPv4,CoS> mtu 1500 index
4
        inet 199.199.131.111 netmask ffffff00 broadcast 199.199.131.255
e1000g131001: flags=201000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4,CoS> mtu 1500 index 5
        inet 0.0.0.0 netmask ff000000 broadcast 0.255.255.255
        groupname pub_ipmp0
        ether 0:21:28:4f:5d:a1
root@cont10g01:/#
```

Configure Node2 hostname "cont10g02"

Public IPMP group: "pub_ipmp0":

Physical NIC: e1000g0 has vlan 131 tagged NIC on each container as e1000g131000

Physical NIC: e1000g1 has vlan 131 tagged NIC on each container as e1000g131001

1. 1 IP as physical node's IP hostname: cont10g02 IP: 199.199.131.102 NIC: e1000g131000
2. 1 IP for the Oracle vip hostname: cont10g02-vip IP: 199.199.131.112 NIC: e1000g131000:2
3. No IP for the other active NIC hostname: N/A - IP: N/A NIC: e1000g131001

Private IPMP group: "priv_ipmp0":

1. Physical NIC: e1000g2 has vlan 111 tagged NIC on each container as e1000g111002
2. Physical NIC: e1000g3 has vlan 111 tagged NIC on each container as e1000g111003
3. 1 IP as private NIC hostname: cont10g02-priv - IP: 199.199.111.102 - NIC: e1000g111002
4. No IP for the standby NIC hostname: N/A - IP: N/A - NIC: e1000g111003

Configure "cont10g02":

Create/edit the following files on

```
::::::::::::::
/etc/hostname.e1000g111002
::::::::::::::
cont10g02-priv group priv_ipmp0

::::::::::::::
/etc/hostname.e1000g111003
::::::::::::::
group priv_ipmp0 up

::::::::::::::
/etc/hostname.e1000g131000
::::::::::::::
```

```
cont10g02 group pub_ipmp0

::::::::::::::
/etc/hostname.e1000g131001
::::::::::::::
group pub_ipmp0 up

::::::::::::::
/etc/hosts
::::::::::::::
#
# Internet host table
#
::1     localhost
127.0.0.1       localhost
199.199.131.102 cont10g02       loghost
199.199.131.101 cont10g01
199.199.131.111 cont10g01-vip
199.199.131.112 cont10g02-vip
199.199.111.101 cont10g01-priv
199.199.111.102 cont10g02-priv

root@cont10g02:/# svcadm restart network/physical

This is how it looks post configuration on "cont10g02":

root@cont10g02:/# ifconfig -a
lo0: flags=2001000849<UP,LOOPBACK,RUNNING,MULTICAST,IPv4,VIRTUAL> mtu 8232 index 1
        inet 127.0.0.1 netmask ff000000
e1000g111002: flags=201000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4,CoS> mtu 1500 index 2
        inet 199.199.111.102 netmask ffffff00 broadcast 199.199.111.255
        groupname priv_ipmp0
        ether 0:21:28:5e:1b:b4
e1000g111003: flags=201000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4,CoS> mtu 1500 index 3
        inet 0.0.0.0 netmask ff000000 broadcast 0.255.255.255
        groupname priv_ipmp0
        ether 0:21:28:5e:1b:b5
e1000g131000: flags=201000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4,CoS> mtu 1500 index 4
        inet 199.199.131.102 netmask ffffff00 broadcast 199.199.131.255
        groupname pub_ipmp0
        ether 0:21:28:5e:1b:b2
e1000g131000:1: flags=201040843<UP,BROADCAST,RUNNING,MULTICAST,DEPRECATED,IPv4,CoS> mtu 1500 index
           4
        inet 199.199.131.112 netmask ffffff00 broadcast 199.199.131.255
e1000g131001: flags=201000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4,CoS> mtu 1500 index 5
        inet 0.0.0.0 netmask ff000000 broadcast 0.255.255.255
        groupname pub_ipmp0
        ether 0:21:28:5e:1b:b3
root@cont10g02:/#
```

Conclusion

The combination of Oracle Solaris IPMP and Oracle RAC provides high availability and scalability for private and public networking. IPMP handles fault detection and recovery of NICs failures in a manner transparent to Oracle RAC. Configuration of link-based active-active IPMP group is best for Oracle RAC private network as it supports configuring multiple IPs. IPMP dynamically manages high availability of VIPs hosted on Oracle RAC on public network interface.

This paper has demonstrated link-based IPMP group configuration for public and private networking of Oracle RAC as the best suitable configuration. Link-based IPMP has many advantages over probe-based IPMP group as it simplifies the configuration and management of IP addresses. It does not need additional IP addresses as the case of probe-based IPMP groups. Disabling of 'FAILBACK' option under /etc/default/mpathd prevents fail-back by which higher up time of hosted service or application is achieved.

References

- IPMP Overview http://docs.sun.com/app/docs/doc/816-4554/mpoverview?l=en&a=view

- Configuring IPMP groups http://docs.sun.com/app/docs/doc/816-4554/emybr?l=en&a=view

- Oracle 10gR2 High Availability Overview http://www.oracle.com/pls/db102/to_toc?pathname=server.102%2Fb14210%2Ftoc.htm&remark=portal+%28Administration%29

- Oracle 10gR2 Clusterware and Oracle Real Application Clusters Administration and Deployment Guide http://www.oracle.com/pls/db102/to_toc?pathname=rac.102%2Fb14197%2Ftoc.htm&remark=portal+%28Administration%29

- "Virtualization Options for Oracle Database deployments on Sun SPARC Enterprise T-series systems" http://www.oracle.com/technetwork/articles/systems-hardware-architecture/virtualization-options-t-series-168434.pdf
  Roman Ivanov, Mohammed Yousuf [2010]

- "Best practices for deploying Oracle RAC in Oracle Solaris Containers" http://www.oracle.com/technetwork/articles/systems-hardware-architecture/deploying-rac-in-containers-168438.pdf
  Mohammed Yousuf [2010]

Highly Available and Scalable Oracle RAC Networking with Oracle Solaris 10 IPMP
Septem 2010
Author: John Mchugh, Mohammed Yousuf

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

SOFTWARE. HARDWARE. COMPLETE.