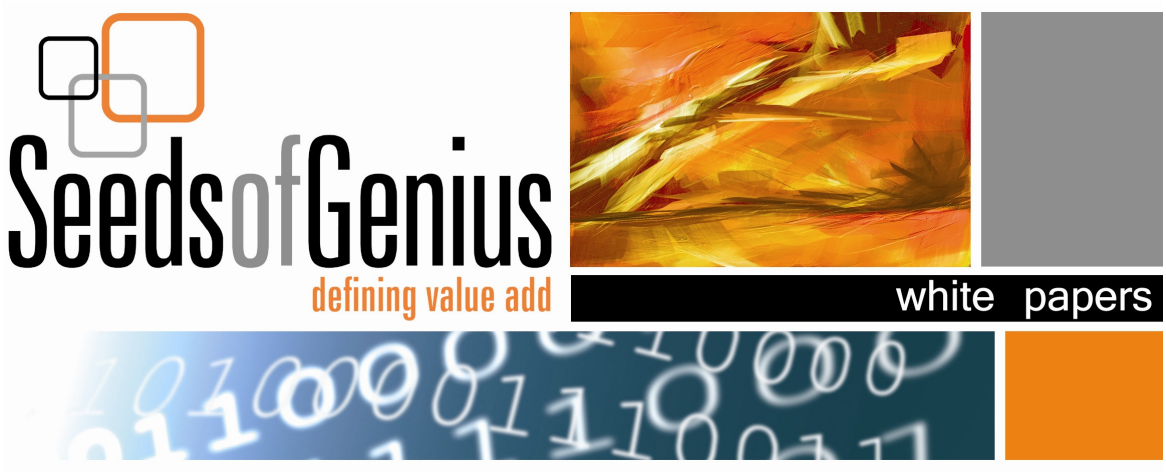


Solaris IPMP
(IP Multipathing)

Brad Isbell, Seeds of Genius
January 2008

Edition: 1.0 January 2008

Purpose:

In a highly available server configuration it is important to eliminate any single point of failure. IPMP, or IP Multipathing, provides a mechanism for building redundant network interfaces to guard against failures with NIC's, cables, switches or other networking hardware. It also provides a method to do maintenance on system hardware without losing network connectivity in environments that make use of the DR features in high end Sun server. This document explains the concepts of IPMP and provides a commonly used configuration. For more information on IPMP please visit the following URL. http://docs.sun.com/app/docs/doc/816-4554/6maoq027r?a=view

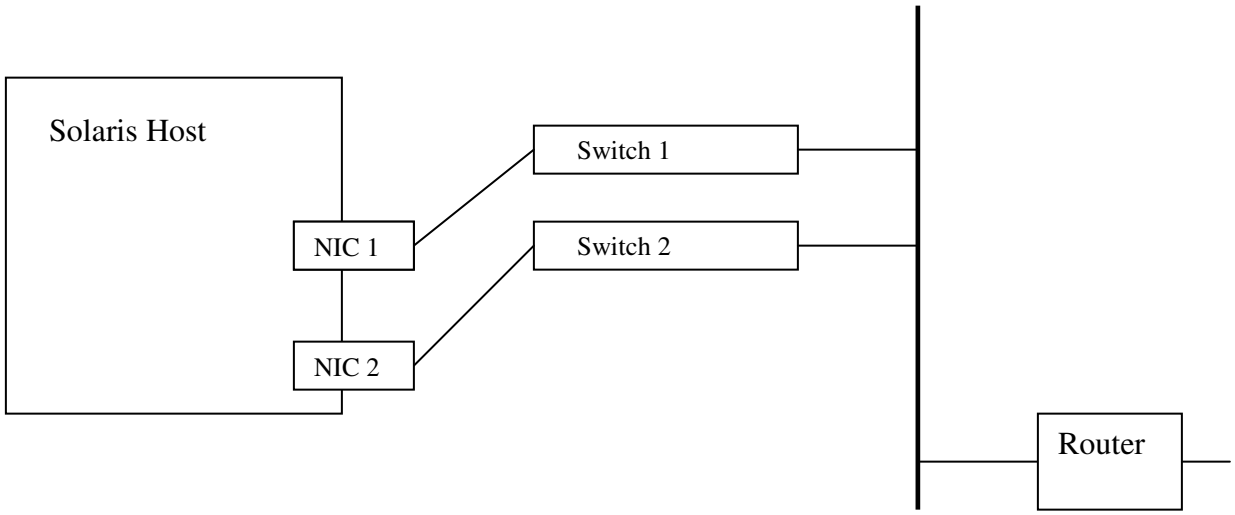
Requirements:

- 2 network interface cards
- 3 IP addresses

Technical Overview:

When configuring IP Multipathing for your Solaris host you will combine two or more physical network interfaces into an IPMP group. For each physical network interface that is in an IPMP group one IP address needs to be allocated for failure testing. These IP links will be used to periodically send an ICMP echo request to a target system and listen for the response. If no response occurs within a given number of tries the link is considered dead and will cause a failover of all application IP addresses currently configured on that physical interface to another physical interface within the IPMP group.

In the diagram below NIC 1 and NIC 2 will be configured into an IPMP group. NIC 1 will be configured with a test IP address of 192.168.1.105 and NIC 2 will be configured with a test IP address of 192.168.1.205. A third IP address, 192.168.1.5, will be assigned to the IPMP group for application traffic. This IP address will be initially bound to NIC 1 as a virtual interface. If NIC 1 fails, then the application IP will transfer to NIC 2 since they are both in the same IPMP group.



Configuration:
The configuration of IPMP depends primarily on the setup of the /etc/hostname.xxxN files that are used to configure network interfaces during the system boot. There are three options to the *ifconfig* command that need to be explained.

deprecated: This option specifies that an IP address should not be used for the transfer of application data. This means that this IP address will only be used as a test interface to determine if the network link is active and alive.

-failover: No failover. The minus sign followed by the word 'failover' indicates that this IP address will not failover to another NIC in the IPMP group when a failure is detected. This is to keep an IP address assigned to a failed NIC so there will be a method for detecting when the failure has been fixed.

group: The IPMP group that this interface belongs to. This IPMP group does not need to already exist, the first member of the group will create the group. The IPMP group name should not contain spaces.

If the two NIC's on a Solaris host are *ce0* and *ce1* the following files would configure IPMP across these two NIC's.

/etc/hostname.ce0

```
host1-ce0 netmask + broadcast + deprecated -failover group IPMP-1
addif  host1  netmask + broadcast +
```

/etc/hostname.ce1

```
host1-ce1 netmask + broadcast + deprecated -failover group IPMP-1
```

/etc/hosts

```
192.168.1.5       host1
192.168.1.105     host1-ce0
192.168.1.205     host1-ce1
```

In this configuration the two interfaces are configured as an IPMP group each with a test IP address assigned to it. The test IP addresses will not carry application data (deprecated) and will not failover to the other NIC in the event of a failure (-failover). The application IP address will be assigned to *ce0* as a virtual interface, *ce0:1*.
If the interface *ce0* fails then the *host1* address will be migrated to *ce1* as virtual interface *ce1:1*.

In the following example, the command *if_mpadm* is used to cause a failover of ce0 to ce1 for demonstration purposes. The following options are used with *if_mpadm*.

if_mpadm:
-d: detach or offline an interface
-r: reattach or online an interface that has been previously offlined with the -d option.

```
# ifconfig -a
lo0: flags=2001000849<UP,LOOPBACK,RUNNING,MULTICAST,IPv4,VIRTUAL> mtu 8232 index 1
        inet 127.0.0.1 netmask ff000000
ce0: flags=9040843<UP,BROADCAST,RUNNING,MULTICAST,DEPRECATED,IPv4,NOFAILOVER> mtu 1500 index 2
        inet 192.168.35.93 netmask ffffffc0 broadcast 192.168.35.127
        groupname ipmp1
        ether 0:14:4f:5a:32:5e
ce0:1: flags=1000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4> mtu 1500 index 2
        inet 192.168.35.113 netmask ffffffc0 broadcast 192.168.35.127
ce1: flags=9040843<UP,BROADCAST,RUNNING,MULTICAST,DEPRECATED,IPv4,NOFAILOVER> mtu 1500 index 3
        inet 192.168.35.103 netmask ffffffc0 broadcast 192.168.35.127
        groupname ipmp1
        ether 0:14:4f:5a:32:5e

# if_mpadm -d ce0

# ifconfig -a
lo0: flags=2001000849<UP,LOOPBACK,RUNNING,MULTICAST,IPv4,VIRTUAL> mtu 8232 index 1
        inet 127.0.0.1 netmask ff000000
ce0: flags=89040842<BROADCAST,RUNNING,MULTICAST,DEPRECATED,IPv4,NOFAILOVER,OFFLINE> mtu 1500
index 2
        inet 192.168.35.93 netmask ffffffc0 broadcast 192.168.35.127
        groupname ipmp1
        ether 0:14:4f:5a:32:5e
ce1: flags=9040843<UP,BROADCAST,RUNNING,MULTICAST,DEPRECATED,IPv4,NOFAILOVER> mtu 1500 index 3
        inet 192.168.35.103 netmask ffffffc0 broadcast 192.168.35.127
        groupname ipmp1
        ether 0:14:4f:5a:32:5e
ce1:1: flags=1000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4> mtu 1500 index 3
        inet 192.168.35.113 netmask ffffffc0 broadcast 192.168.35.127

# if_mpadm -r ce0

# ifconfig -a
lo0: flags=2001000849<UP,LOOPBACK,RUNNING,MULTICAST,IPv4,VIRTUAL> mtu 8232 index 1
        inet 127.0.0.1 netmask ff000000
ce0: flags=9040843<UP,BROADCAST,RUNNING,MULTICAST,DEPRECATED,IPv4,NOFAILOVER> mtu 1500 index 2
        inet 192.168.35.93 netmask ffffffc0 broadcast 192.168.35.127
        groupname ipmp1
        ether 0:14:4f:5a:32:5e
ce0:1: flags=1000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4> mtu 1500 index 2
        inet 192.168.35.113 netmask ffffffc0 broadcast 192.168.35.127
ce1: flags=9040843<UP,BROADCAST,RUNNING,MULTICAST,DEPRECATED,IPv4,NOFAILOVER> mtu 1500 index 3
        inet 192.168.35.103 netmask ffffffc0 broadcast 192.168.35.127
        groupname ipmp1
        ether 0:14:4f:5a:32:5e
```

Optional Configuration:

Test System:

The *in.mpathd* daemon, responsible for handling IPMP, will select a system on the network to send IPMP echo requests for all test interfaces in the IPMP group. The *in.mpathd* daemon will try pinging the host's routers by first checking all host routes, and then checking the default router. If a default router is configured then no further configuration needs to be made as long as frequent and constant ICMP echos to your default router are tolerable. If you have disabled ICMP responses on your router, or want to keep this traffic off this device you can specify an alternate, dedicated IP address to use as a test system by creating a host route.

 # route add -host xxx.xxx.xxx.xxx xxx.xxx.xxx.xxx

If no routes are configured then the host will determine its test system by sending a packet to the multicast address of *224.0.0.1* and waiting for a response.

Failover Detection Time:

The time it takes for an interface to determine if it is failed is configured in the file /etc/default/mpathd. By default the *FAILURE_DETECTION_TIME* setting is set to 10000ms, or 10 seconds. The *in.mpathd* daemon is considered failed if 5 consecutive ICMP packets fail during this time period. So, *in.mpathd* will send 5 ICMP echo requests during this time period, or 1 packet every 2 seconds.
If the *FAILURE_DETECTION_TIME* is set to 20000ms, or 20 seconds, then an ICMP packet is sent every 4 seconds.

Failback:

In order to failback to a previously failed interface the FAILBACK setting in /etc/default/mpathd must be set to *yes*. This is the default setting. After an interface has been failed, *in.mpathd* needs to receive 10 consecutive ICMP echo responses. The default setting of 10000ms will result in a 20 second failback to an interface that has been brought back online.

If you need any help please contact the Seeds of Genius support center at support@seedsofgenius.com.