fineDoc Number:	EDCS-<XXXXX>
Last Revision Date:	December 15, 2010
Created by:	Louis Watta
Template Ver. Number:	EDCS-XXXX Rev X

CAE

Nexus 1000V 1.4 Network State Tracking (NST) White Paper

TABLE OF CONTENTS

1 Introduction

The Cisco Nexus 1000V, is a Cisco developed server virtualization switching architecture for VMware ESX environments. The Nexus 1000V enables policy based virtual machine (VM) connectivity, mobility of security and network properties, and a non-disruptive operational model for both Server and Network administrators.

Offering a set of network features, management tools and diagnostic capabilities consistent with the customer's existing physical Cisco network infrastructure and enhanced for the virtual world, the Nexus 1000V allows customers to accelerate their adoption of VMs through the unification & simplification of the physical and virtual networks. The n1000v also secures & simplifies the deployment & movement of VM's to increase service velocity while maintaining and enforcing security policy.

1.1 White Paper

The purpose of this white paper is to walk the user through the Network State Tracking feature.

1.2 Assumptions

The assumptions of this white paper are that the reader has

- Installed VMware VC 4.0U1/U2 or VC 4.1
- Installed Cisco Nexus 1000V 1.3/1.3a/1.3b on an ESX VM in HA mode
- At least 2 ESX 4.0U2/U1 or 4.1 boxes with VEM module already loaded
- Created a Nexus 1000V Distributed Virtual Switch (DVS) under vCenter
- Added the ESX boxes to the Nexus 1000V DVS

2 Network State Tracking (NST or Beaconing)

2.1 Introduction

Network State Tracking (formally Beacon Probing) is new in this release of the Cisco Nexus 1000V. Network State Tracking is a network failover detection mechanism that sends out and listens for beacon probes on all NICs in the team and uses this information, in addition to link status, to determine link failure. Network State Tracking detects failures, such as cable pulls and physical switch power failures, but not configuration errors.

Similar to beacon probing feature provided by VMware, however we do not require 3 NICs to do Network State Tracking. NST can function with only two NICs. However, when there are only two NICs in the configuration it is difficult to determine which NIC is really having the problem. Thus NST will only take a NIC out of service if it sees no ingress traffic on the NIC. We will highlight this in the configuration example below

Network State Tracking only works with VPC-HM and VPC-MAC Pinning port-channels. It will not work with LACP port-channel.

This release of the Nexus 1000V Network State Tracking also addresses a specific issue when Nexus 1000V is run on HP Virtual Connect. The issue is as follows

> "The problem is that one physical link from the Flex-10 fabric appears as four Flex-10 NICs (physical NICs) to the VMkernel. Link states of all the Flex-10 NICs have to change together so that link-state of individual Flex-10 NIC can be reported correctly. Due to this issue, ESX may not be able to detect failures when a corresponding external uplink is down, even if SmartLink is turned on and the Ethernet Connection Mode is connected to disable failover.
>
> There are two known work-arounds at this point:
>
> 1. The problem is mainly observed with Flex-10 firmware 2.30 and above. Downgrade to Flex-10 firmware 2.10 (you need to check with HP on the feasibility of this option)
>
> 2. It is also not observed if you use "Beacon Probing" for Network Failure Detection instead of "Link Status only" in the Nic Teaming tab of the Port Group properties."

2.2 Requirements

You need to be running the Nexus 1000V 1.4 code for Network State Tracking to be enabled.

You will need at least one ESX host with network connections to two different switches. The ESX host needs to be added to the running Nexus 1000V and the uplink port-profile needs to be configured with either vPC-HM or vPC MAC-Pinning.

2.3 Configure Network State Tracking

By default Network State Tracking is turned off.

The default timer for probing when enabled is 5 seconds (configurable 1-10 seconds)
The default threshold miss count is 5 seconds (configurable 3-7 seconds)
The default split mode action is no repin (configurable repin/no repin)

Network State Tracking on the Cisco Nexus 1000V works on each sub-group by selecting one 1 uplink and 1 vlan from the sub-group. It is the first uplink and the first forwarding VLAN added.

How it works.
Tracking broadcasts are sent out on all sub-groups on one interface on one VLAN. We expect the broadcast packets to be received on other sub-groups. We track consecutive miss count. If the miss count >= the threshold we declare the port-channel to be in split mode. Once a beacon is received we declare a recovery from split-mode for that port-channel. All of this is tracked via syslog. A NIC is only failed when no ingress traffic is received.

If repin is turned on then when the system recovers from split-mode it will pin all the veths back to the original sub-group that was failed.

Verify Network State Tracking is off

```
n1000v-MV(config)# show network-state tracking configuration
Tracking mode        : disabled
```

```
Tracking Interval    : 5 sec
Miss count threshold : 5 pkts
Split-network action : re-pin
```

Network State Tracking is turned on as follows

```
n1000v-MV(config)# track network-state ?
  <CR>
  interval   Set timer interval at which tracking packets need to be sent
  split      Configure the settings for split-network mode
  threshold  Set max threshold value

n1000v-MV(config)# track network-state
n1000v-MV(config)#
```

Verify that it is turned on

```
n1000v-MV(config)# show network-state tracking configuration
Tracking mode        : enabled
Tracking Interval    : 5 sec
Miss count threshold : 5 pkts
Split-network action : re-pin
```

Note that Network State Tracking only works with vPC-HM and vPC-Mac Pinning port-channels

Once it is enabled Network State Tracking will be used on every VEM that is configured with a vPC-HM or vPC-Mac Pinning uplink port-profile

Defaults are as follows

```
n1000v-MV(config)# show network-state tracking configuration
Tracking mode        : enabled
Tracking Interval    : 5 sec
Miss count threshold : 5 pkts
Split-network action : re-pin
```

Beacon Interval – time between each beacon being sent. Default is 5 seconds
Miss count threshold – the number of beacons that can be missed in succession before a split network is detected
Split-network action – what the Nexus 1000V will do if a split network is detected. No re-pin means veths stay where they are. Re-pin means it will try to pin traffic to another uplink.

2.4 Verify Network State Tracking on the ESX host (VEM)

You need to be running the Nexus 1000V 1.4 code for Network State Tracking to be enabled.
Requirements:

- You have turned on NST on the VSM.
- Created an uplink port-profile with vPC-HM (either CDP or Mac Pinning)
- Added 2 nics from ESX host into the uplink port-profile

To verify Beacon Probing on the ESX host itself

```
[[root@cae-esx-196 ~]# vemcmd show tracking
Network-state tracking information:

Port-channel index 0 (port ltl 305):
Split-network mode: 0
```

```
Tracking vlan: 1

From  using  To  Threshold   Present
SG    ltl    SG  miss count  state
----  -----  --  ----------  -------
0     17     1   0           Normal
1     18     0   0           Normal

       Time      Total ingress
SG_ID  Interval  Packets
-----  --------  --------
0      0         31480
0      1         33266
0      2         35076
0      3         27891
0      4         29686
1      0         145808
1      1         147605
1      2         149420
1      3         142151
1      4         143973
```

Note that the tracking VLAN is VLAN 1
This can be changed manually if desired with the below commands.

First run “vemcmd show port” to get the bndl id of your port-channel

```
[[root@cae-esx-196 ~]# vemcmd show port
  LTL   VSM Port  Admin Link  State  PC-LTL  SGID  Vem Port
   17    Eth12/1     UP   UP    FWD     305     0  vmnic0
   18    Eth12/2     UP   UP    FWD     305     1  vmnic1
   49      Veth4     UP   UP    FWD       0     1  vmk0
   50      Veth5     UP   UP    FWD       0     1  vswif0
  305        Po4     UP   UP    FWD       0
```

Next run the vemset tracking command with the bndl id from above

```
[root@cae-esx-196 ~]# vemset tracking vlan 2 ltl 305

[root@cae-esx-196 ~]# vemcmd show tracking config
Network-state tracking configuration
  Enabled                       : Yes
  Tracking interval             : 5 sec
  Threshold miss count          : 5 pkts
  Split network action mode     : repin

Port-channel ltl   Tracking vlan
----------------   -------------
       305                 2
```

Now vlan 2 will be used for tracking probes on the this ESX host only.

For troubleshooting purposes tracking can also be turned off/on the ESX host as well with the following command.

```
[root@cae-esx-196 ~]# vemset set tracking enable
[root@cae-esx-196 ~]# vemset set tracking disable
```

You can also clear the statistics with the below command

```
[root@cae-esx-196 ~]# vemset clear tracking stats
```

Lastly you can also get statistics from the VSM with the following commands

```
n1000v-MV# show network-state tracking module 12
Port-    Network  Tracking  SG  SG        Tracking   SG
Channel  Mode     Vlan      ID  State     Interface  Members
-------  -------  --------  --  --------  ---------  -----------------
Po4      ok       2         0   Active    Eth12/1    Eth12/1
Po4      ok       2         1   Active    Eth12/2    Eth12/2

n1000v-MV# show network-state tracking interface port-channel 4
Port-    Network  Tracking  SG  SG        Tracking   SG
Channel  Mode     Vlan      ID  State     Interface  Members
-------  -------  --------  --  --------  ---------  -----------------
Po4      ok       2         0   Active    Eth12/1    Eth12/1
Po4      ok       2         1   Active    Eth12/2    Eth12/2
```

2.5 Fail upstream link and monitor

Fail a link on the upstream switch that will cause the NST to see a failed link. Remember not to fail the link from the ESX host to the upstream switch as that will immediately get recognized as a failure by the port-channel. You want to fail a link upstream not directly connected to the ESX host.

An easy way to simulate a failure is to remove all the VLANs from one of the upstream links on the ESX host. This will prevent traffic from flowing to the port but not hard fail the port so that the port-channel fails.

Once the upstream link is failed you will see the following.

```
[[root@cae-esx-196 ~]# vemcmd show tracking
Network-state tracking information:

Port-channel index 0 (port ltl 305):
Split-network mode: 1
Tracking vlan: 2

From  using  To  Threshold   Present
SG    ltl    SG  miss count  state
----  -----  --  ----------  -------
0     17     1   5           Normal
1     18     0   5           Normal
```

When the threshold count hits 5 packets the link will go into split-network mode. If no packets are received on the failed link then the VEM module will fail that link and force all packets to go up the active link until broadcasts are received again.