CHAPTER 7

Resiliency and High Availability

The Virtualized Multiservice Data Center (VMDC) 2.3 design provides a number of High Availability (HA) features and is a highly resilient network. The following sections provide an overview of network resiliency and also summarize the validation results of convergence around service impacting failures as tested in the lab configuration.

This section presents the following topics:

- Resiliency Against Link and Node Failure, page 7-1
- Convergence Test Results, page 7-2

Resiliency Against Link and Node Failure

HA has different aspects that are implemented at different layers in the network. The VMDC2.3 design does not have any single point of failure, and the service impacting failures are minimized by ensuring quick convergence around the failing link or node. In terms of converging around the failing link or node, this may be required as part of planned maintenance or an unplanned failure event. Planned events are most commonly done to upgrading software on various nodes in the Data Center (DC) and other maintenance reasons on power plants, and to address facilities issues.

In VMDC 2.3, the network portion has dual paths, with two nodes supporting each path, in an active/active configuration with load balancing of traffic achieved by using Border Gateway Protocol (BGP). During maintenance events on one node, wherein the node is taken down, the traffic and services can continue to be provided using the other path, however, there could be local congestion during such events, as one node going down would cause all traffic to use the other path. For example, when the Provider Edger (PE) node is down, all traffic uses the surviving PE and WAN link, which causes the bandwidth available for the entire DC to be reduced to half. This can be avoided by using dual-redundant route processors and the Encapsulating Standard Protocol (ESP) on the ASR 1006 DC-PE, and by using dual supervisors on the Nexus 7004 DC-Agg routers, which is our recommendation. In addition to the benefit of being able to perform In Service Software Upgrade (ISSU), any unexpected failure of the supervisors when configured with redundancy will cause automatic switchover to the the redundant RP/supervisor, and forwarding is minimally impacted. Similarly, it is highly recommended to deploy other services appliances and compute infrastructure, as well as the Nexus 1000V Virtual Supervisor Module (VSM) and Virtual Security Gateway (VSG) in a HA configuration with a pair of devices to support failover. Additionally, for redundancy on the link level on the Nexus 7004, two modules are used, and port-channels with members from both modules are used to provide service continuously for planned or unplanned events on each module.

Table 7-1 lists the redundancy model for the ASR 1006 and Nexus 7004.

Table 7-1 Redundancy Model for ASR 1006 and Nexus 7004

Event Type	Planned/Unplanned	Redundancy	Mechanism	Service Impact
Software upgrades	Planned	Not redundant	Routing convergence	Yes, convergence event during link/ node shut
Software upgrades	Planned	Redundant	HA/SSO	Minimum, zero packet loss in most conditions *
Software or hardware error	Unplanned	Not redundant	Routing convergence	Yes, convergence event
Software or hardware error	Unplanned	Redundant	HA/SSO	Minimum, zero packet loss in most conditions *

Note The ASR 1000 is impacted by CSCuc51879. This issue causes packet drops during RPSO or during ISSU on an ASR 1000 PE with a highly scaled up configuration and is still under investigation as of this publication.

For other nodes used in the VMDC 2.3-based DC, Table 7-2 lists the redundancy model to support not having a single point of failure.

Table 7-2 Redundancy Model for Services and Compute

Node Type	Role/Services	HA
ASA 5585	Firewall	FT using active/standby pair
ASA 5555-X	VPN	FT using active/standby pair
ACE 4710	SLB	FT using active/standby pair
Nexus 1010	Virtual Service Blades	Paired Nexus 1010 in active/ standby
Nexus 1000V VSM	Virtual Supervisor Module	Paired VSM in active/standby
UCS Fabric Interconnect 6248	Fabric Interconnect	Pair of FI/6248 devices in active/standby for management, active/active for data
ICS Switch Nexus 5000	ICS Access switch	Pair in virtual port-channel, no stateful sync
Compute Cluster	Compute	VMware HA/DRS cluster
FC Storage/NAS Storage	Storage	NetApp dual controllers
VSG	Compute Firewall	Pair of active/standby, statefully synced
VNMC	Compute Firewall Management	Use VMware HA

Convergence Test Results

Table 7-3 and Table 7-4 detail convergence results for ASR 1006 DC-PE, and Nexus 7004 aggregation switch convergence events.

1. For the network test scenarios, traffic was sent using traffic tools (IXIA, Spirent TestCenter) for all tenants north to south.

2. A convergence event was triggered for all tenants north to south.
3. MAC scale was set to 13,000 - 14,000 MAC addresses on the Nexus 7004 devices.
4. Additional traffic was sent between east/west tenants for the first 50 tenants.
5. The worst case impacted flow amongst all the flows is reported. It should be noted that all flows are not impacted due to alternate paths, which are not impacted during tests.

Table 7-3 ASR 1006 DC-PE Convergence Events

	Event	N-S	S-N	Comments	Issues
1	Node fail	2.7-5.3 sec	0.3 sec		
2	Node recovery	zpl	zpl		
3	Link fail ASR 1000 to Nexus 7004 Agg	2.502 sec	4.628 sec		
4	Link restore ASR 1000 to Nexus 7004 Agg	zpl	zpl		
5	ASR 1000 RP switchover CLI	0 sec	0.5 sec	With scaled up configuration on the ASR 1000 PE, some losses were seen instead of zpl	CSCuc51879[1]
6	ASR 1000 RP switchover - RP module pull	0	2.6-5 sec	With scaled up configuration on the ASR 1000 PE, some losses were seen instead of zpl	CSCuc51879
7	ASR 1000 ISSU	15 sec	23 sec	With scaled up configuration on the ASR 1000 PE, some losses were seen instead of zpl	CSCuc51879
8	ASR 1000 ESP module pull	zpl	zpl		
9	ASR 1000 ESP switchover via CLI	zpl	zpl		

1.The fix for this issue is still under investigation at the time of this publication.

Table 7-4 Nexus 7004 Aggregation Switch Convergence Events

	Event	N-S	S-N	Comments	Issues
1	Nexus 7004 AGG module fail	3.6 sec	3.12 sec		2 3
2	Nexus 7004 AGG module restore	9.9 sec	10.9 sec		2 6
3	Nexus 7004 AGG node fail	1-2 sec	1-2 sec		
4	Nexus 7004 AGG node recovery	5 sec	8 sec	See Layer 3 Best Practices and Caveatsfor more information. Additional steps are needed to workaround the issue. 4 5	1 2 4 5

Table 7-4 Nexus 7004 Aggregation Switch Convergence Events (continued)

5	Nexus 7004 AGG vPC peer link fail	13 sec	2 sec	See Layer 3 Best Practices and Caveats for more information. BGP convergence will move traffic off the Nexus 7004 path. Future fixes to help with convergence.	1 2 5
6	Nexus 7004 AGG vPC peer link restore	3.5 sec	6.3 sec		2
7	Nexus 7004 AGG link to ICS Nexus 5548 SW fail	0.2 sec	0.2 sec	Fiber pull	
8	Nexus 7004 AGG link to ICS Nexus 5548 SW restore	0.2 sec	0.2 sec	Fiber restore	
9	Nexus 7004 AGG supervisor fail - module pull	zpl	zpl		
10	Nexus 7004 AGG supervisor switchover - CLI	zpl	zpl		
11	Nexus 7004 ISSU	zpl	zpl		

Note The following issues are being fixed in the Nexus 7000, but are not available in the release tested. These fixes are currently planned for release in the 6.2-based release of NX-OS for the Nexus 7000.

- [1]CSCtn37522: Delay in L2 port-channels going down
- [2]CSCud82316: vPC Convergence optimization
- [3]CSCuc50888: High convergence after F2 module OIR The following issue is under investigation by the engineering team:
- [4]CSCue59878: Layer 3 convergence delays with F2 module - this is under investigation. With scale tested, the additional delay is between 10-17s in the vPC shut case. The workaround used is to divert the traffic away from N7K Agg as control plane (BGP) does converge quicker and traffic bypasses the N7K agg. Also use the least amount of port groups possible to reduce the number of programming events. Alternatively, consider using M1/M2 modules for higher scale of prefixes and better convergence.

The following issues are closed without any fix, as this is the best convergence time with the F2 module after the workarounds are applied:

- [5]CSCue67104: Layer 3 convergence delays during node recovery (reload) of N7k Agg. The workaround is to use L3 ports in every port group to download FIB to each port-group.
- [6]CSCue82194:High Unicast convergence seen with F2E Module Restore

Table 7-5, Table 7-6, Table 7-7, and Table 7-8 detail convergence results for Nexus 5500 Series ICS switch, ACE 4710 and Nexus 7004, ASA 5585, and other convergence events.

Table 7-5 Nexus 5500 Series ICS Switch Convergence Events

	Event	N-S	S-N	Comments	Issues
1	Nexus 5548 ICS SW vPC peer link fail	1.7 sec	2.5 sec		
2	Nexus 5548 ICS SW vPC peer link restore	7.14 sec	9.35 sec		
3	Nexus 5548 ICS SW Node fail	1.7 sec	0.36 sec		
4	Nexus 5548 ICS SW Node Recovery	9.8 sec	8.5 sec		

Table 7-6 ACE 4710 and Nexus 7004 Convergence Events

	Event	N-S	S-N	Comment	Issues
1	ACE failover with CLI	0.072 sec	0.072 sec		
2	ACE node fail	9.3 sec	9.3 sec	ACE FT configured 10 sec	
3	ACE node recovery	5.7 sec	3.1 sec		
4	ACE Single link fail	0.5 sec	0.5 sec		
5	ACE Single link restore	0.05 sec	0.05 sec		
6	ACE Dual links to same Nexus 7000 fail	2.5 sec	2.5 sec		
7	ACE Dual link to same Nexus 7000 restore	2.5 sec	1.2 sec		
8	ACE Port-ch fail	13 sec	13 sec	ACE FT configured 10 sec	
9	ACE port-ch restore	10 sec	10 sec	ACE FT configured 10 sec	

Table 7-7 ASA 5585 Convergence Events

	Event	N-S	S-N	Comment	Issues
1	ASA FT link fail	zpl	zpl	Failover is disabled	
2	ASA FT link restore	zpl	zpl	Failover is disabled	
3	ASA reload	4-6 sec	4-6 sec	Failover pull time/hold time 1/3 sec	
4	ASA recovery	0.2-3 sec	0.2-3 sec	Default preemption	

Table 7-8 Other Convergence Events

	Events	Traffic Impact	Comments	Issues
1	UCS FI 6248 fail	2.20648 sec		
2	UCS FI 6248 restore	0.6909 sec		
3	UCS FT switchover	2.2954 sec when FT failed, 0.3893 sec when FT restored.	No packet drop if only the control/ management plane of the FT switchover	

Table 7-8 Other Convergence Events (continued)

4	Ethernet link fail between 6248 and 5500	Minimal impact seen.	VM running script continuously accessing a file (read/write) on NFS data store continues after a momentary stop	5
5	Ethernet link Restore between 6248 and 5500	No impact seen.	VM running script continuously accessing a file (read/write) on NFS data store sees no impact	
6	FC link fail between 6248 and 5500	Minimal impact seen.	VM running script continuously accessing a file (read/write) on FC data store continues after a momentary stop	
7	FC link restore between 6248 and 5500	No impact seen.	VM running script continuously accessing a file (read/write) on FC data store sees no impact	
8	FC link fail between 5500 and NetApp	No impact seen.	VM is up and running and able to browse data store	
9	FC link restore between 5500 and NetApp	No impact seen.	VM is up and running and able to browse data store	
10	Link fail between FT and TOM	3.9674 sec		
11	Link restore between FT and TOM	0.2098 sec		
12	Fail IOM	1.12765 sec		
13	Restore IOM	0.25845 sec		
14	Fail Nexus 1010	0 sec	Nexus 1010 not in the forwarding path of data traffic.	
15	Restore Nexus 1010	0 sec	Nexus 1010 not in the forwarding path of data traffic.	
16	CLI switchover Nexus 1010	0 sec	Nexus 1010 not in the forwarding path of data traffic.	
17	Nexus 1000V VSM switchover	0 sec	VSM not in the forwarding path of data traffic.	
18	Nexus 1000V VSM failure	0 sec	VSM not in the forwarding path of data traffic.	
19	Nexus 1000V VSM restore	0 sec	VSM not in the forwarding path of data traffic. The failed node booted up the the standby node.	
20	Nexus 1000V VSG failure	0 sec for established flow. 7.98 sec for new connection setup.		
21	Nexus 1000V VSG restore	0 sec	The failed node booted up the the standby node.	

Table 7-8 Other Convergence Events (continued)

22	UCS blade fail	A few minutes.	All VMs on the failed blade will fail. vSphere HA will restart the VM on another blade, guest OS boot up takes a few minutes.	
23	UCS blade restore	0 sec	No impact, unless vSphere DRS vMotions the VMs to the restored blade because of the load on other blades. vMotion would cause a packet drop of 1-2 sec. This depends on which VMs are moved and the resource usage.	

Authors

- Sunil Cherukuri
- Krishnan Thirukonda
- Chigozie Asiabaka
- Qingyan Cui
- Boo Kheng Khoo
- Padmanaba Kesav Babu Rajendran