APPENDIX A

Best Practices and Caveats

This appendix provides the best practices and caveats for implementing the VMDC 2.3 solution.

Compute and Storage Best Practices and Caveats

UCS Best Practices

- When using UCSM configuration templates, be aware that some configuration changes will either cause server reboot or service disruption. Multiple templates of the same type should be used to prevent any single change to cause service disruption to all blade servers.
- When configuring server pools, select servers from multiple chassis to avoid single chassis failure bringing down all servers in the pool.
- Disable fabric failover for all vNICs configured for the blade servers, and let the Nexus 1000V manage the vNIC failure.
- UCSM does not support overlapping VLANs in disjoint L2 networks. Ensure that each VLAN only connects to one upstream disjoint L2 network.
- UCS FI uses LACP as the port-channel aggregation protocol. The opposing upstream switches must be configured with LACP active mode.
- A vNIC (VMNIC in the vSphere ESXi hypervisor or physical NIC in the bare metal server) can only communicate with one disjoint L2 network. If a server needs to communicate with multiple disjoint L2 networks, configure a vNIC for each of those networks.
- UCSM implicitly assigns default VLAN 1 to all uplink ports and port-channels. Do not configure any vNICs with default VLAN 1. It is advisable not to use VLAN 1 for carrying any user data traffic.

Storage Best Practices

- If using NetApp OnCommand System Manager 2.0 to configure storage filers, it is recommended to configure the following using the command line:
  - Configuring VIF and VLAN interfaces for NFS port-channel.
  - Configure security style (Unix or Windows) permissions when a volume is exported as NFS.
- To take advantage of Thin Provisioning, it is recommended to configure Thin Provisioning on both volumes/LUNs in storage and in VMFS.
- Configure Asymmetric Logical Unit Access (ALUA) on the filers for asymmetric logical unit access of LUNs.
- Enable storage deduplication on volumes to improve storage efficiency.

- Nexus 5000 is the storage switch in this design. It is mandatory to enable NPIV mode on the Nexus 5000, and also configure soft zoning (enables server mobility) that uses WWPNs.

vSphere ESXi Best Practices

- vSphere Auto Deploy makes use of PXE and gPXE. The PXE/gPXE bootloader does not support 802.1Q tagging of DHCP frames. Configure the VLAN where the ESXi management vmk interface resides as the native VLAN.
- vSphere Auto Deploy makes use of DNS. Configure both forward and reverse DNS resolution for the ESXi hostname on the DNS server.
- When using vSphere Auto Deploy, make sure that the vCenter server, Auto Deploy server, DHCP server, and TFTP server are made highly available.

vSphere ESXi Caveats

- For the UCS blade server with the Cisco VIC adapter (Cisco UCS VIC 1280, Cisco UCS VIC 1240, Cisco UCS M81KR VIC, etc.), the ESXi host boot time will be much longer than those with other adapters. See CSCtu17983 for more details.
- In ESXi version 5.0, the ESXi Network Dump Collector feature is supported only with Standard vSwitches and cannot be used on a VMkernel network interface connected to a vSphere Distributed Switch or Nexus 1000V Switch. See VMware Knowlesge Base for more details.

Nexus 1000V Series Switches Best Practices

- Make sure that the SVS domain IDs for the Nexus 1010 VSA and the Nexus 1000V VSM are unique.
- Configure port profiles for management and vMotion vmknic as **system vlan**.
- Make use of port-profile inheritance to enforce consistent configuration and ease of management.

Layer 2 Best Practices and Caveats

vPC Best Practices

- A vPC peer link is recommended to use ports from different modules to provide bandwidth and redundancy.
- "ip arp synchronize," "peer-gateway," and "auto-recovery" should be configured in the vPC configuration.
- LACP should be used if possible
- It is recommended to disable LACP graceful convergence when the other end of port-channel neighbors are non NX-OS devices.
- Pre-provision all VLANs on MST and then create them as needed.
- On the Aggregation layer, create a root or a secondary root device as usual. Design the network to match the primary and secondary roles with the spanning-tree primary and secondary switches.
- If making changes to the VLAN-to-instance mapping when the vPC is already configured, remember to make changes on both the primary and secondary vPC peers to avoid a Type-1 global inconsistency.

Layer 3 Best Practices and Caveats

Best Practices

1. To accelerate L3 convergence, spread the L3 ports on different SoCs on the F2 module. This is due to the fact that on the F2 module, each port is mapped to a VRF instance and then the FIB for that VRF is downloaded. If an SoC has all ports as L2, then during reload and possibly other conditions, when the ports come up, FIB download is delayed until the SVI to VRF mapping is done, and hence FIB download happens after the port comes up and L2 convergence and mapping of VLANs to that port is complete. In VMDC 2.3 implementation, the L3 ports to the DC PEs and the VPC peer links were spread across five SoCs per module to get the benefit of FIB download immediately on reload. Refer to Cisco Nexus 7000 F2-Series 48-Port 1 and 10 Gigabit Ethernet Module Data Sheet for more information about F2 card and SoCs. Also, see CSCue67104 below.
2. To reduce traffic loss after system reload, delay the time that it takes for VLAN interface and vPCs to come online. By default, VLAN interfaces are brought online 10 seconds after the peer link is up, and vPCs are brought online 30 seconds after the VLAN interfaces are brought up. Based on scale characteristics of this validation, we delay VLAN interfaces and vPCs from coming online by 90 seconds each.
3. The ACE 4710 appliances do not support LACP, and hence their port-channels to the Nexus 7000 switches are static with mode on. We expect to see some traffic loss when the system comes online after a reload. To protect against this loss, carrier delays can be configured on the ACE GigabitEthernet interfaces to prevent this interface from coming online. Using this scheme will introduce a carrier-delay time during a vPC shut/no shut test or similar negative event.
4. Carrier delay can be configured on the ASR 1000 interfaces to the Nexus 7000 aggregation routers to delay the L3 interface from coming up. This ensures that these L3 interfaces are brought up at a time when the Nexus 7000 routers are ready to successfully set up and establish BGP sessions. In this validation, the carrier delay on the ASR 1000 PE was set to the maximum of 60 seconds.
5. By default, the ACE 4710 appliance will renew ARP entries for a configured host every 300 seconds. We increase the ARP rates to 1440 seconds to reduce the possibility of the ACE ARP request being lost as the system comes online after a reload.
6. To get better convergence performance, use BGP policy to divert traffic away from the Nexus 7004 aggregation switch under certain conditions such as VPC peer link fail or secondary shutdown. This is because the FIB programming on the F2 card is slower, leading to additional packet losses of up to 10 seconds in the scale validated, and this can be higher with a high-programmed prefix count. BGP configuration on the ASR 1000 and Nexus 7000 aggregation routers is set up so that the ASR 1000 reroutes traffic to an alternate path if the vPC peer link fails and shuts down the VPC secondary. This eliminates up to 10 seconds of traffic loss that occurs due to the F2 FIB programming delay. If the peer link fails, expect up to 13 seconds of traffic convergence, which is due to up to 8 seconds being required for the VLAN interface to go down, and due to up to 5 seconds being required for the BGP and RIB update on the Nexus 7000 aggregation routers. The causes of this convergence delay in FIB programming is under investigation. See CSCue59878 below. For overall vPC convergence, there are a few enhancements targeted for the next NX-OS software release 6.2.
7. BGP PIC, BGP graceful restart, and other routing optimization should be enabled on the ASR 1000 PE devices for faster convergence. BGP PIC and graceful restart are enabled by default on the Nexus 7000 aggregation routers.

Caveats

1. CSCud23607 was an HSRP programming issue seen if the MAC address table size limits are reached. This is fixed in NX-OS 6.1.3. Prior to NX-OS 6.1.3 , the workaround was to manually flap the affected HSRP interfaces.
2. CSCue59878 was filed to investigate the FIB programming delay after routing convergence during a vPC shut test or similar scenarios. This issue is under investigation. The reason for delay is due to the FIB programming mechanism used for the F2 module. The module has to program TCAM for all 12 SoCs, and as the number of prefixes gets higher, it takes additional time to calculate and program each of the SoCs. The workarounds are to reduce the number of SoCs used, i.e., less number of ports and to reduce the number of prefixes per SoC (by mapping specific VRF instances (ports) to SoCs so that the total prefix is less per SoC). If convergence times need to be quicker, and with a larger number of prefixes, consider using M2 or M1 series modules.
3. CSCue67104 was filled to investigate convergence delays due to packet losses after system reload of the nexus 7000 aggregation router. These losses are seen as FIB losses when the vPC port-channels are brought up and can last 10 or more seconds. This issue was closed as this is expected. On F2 modules, which have an SoC design, each SoC needs to map all of its ports into VRF instances, and then download the FIB. When all of the ports on an SoC are L2 only, the L2 ports need to come up and the SVIs need to be mapped to VRF instances before downloading the FIB for those VRF instances. This takes additional time after the port comes up (see CSCue59878 above, F2 FIB convergence is slow). To work around this issue, have a mix of both L2 and L3 ports on the same SoC. The L3 ports being on the SoC will cause all FIBs for the VRF instances on the L3 port to be downloaded as soon as the module comes up. In VMDC 2.3, all VRF instances used are allowed on the L3 port, so all FIBs will be downloaded to any SoC that has L3 ports. Since there are two L3 uplinks and four L3 peer links for iBGP per box, this provides one L3 port for uplink and two iBGP ports for peer per module. These ports should be spread on three different SoCs. Additionally, we can also spread the vPC peer link ports in different SoCs. Since there are four ports in the vPC peer link, two ports from each module, this covers two more SoCs. This helps with the reload case, as the vPC peer link will come online first and have SVIs mapped to it followed by FIB download, before the actual vPC port-channels come up, however, this will not help in the module restore case, as the vPC peer link port SoCs and FIB download will still be delayed. Additional L3 ports can help, if they are configured on any additional SoCs used. The goal with this workaround is to have all SoC FIBs programmed by the time the vPC port-channels come online.
4. CSCuc51879 is an issue seen during RP failover either due to RPSO or In-Service System Upgrade (ISSU). This is an issue related to traffic loss seen during RPSO or during ISSU on an ASR 1000 PE with a highly scaled up configuration.
5. The following performance fixes are expected in the 6.2 release of NX-OS. These fixes are expected to help with convergence.
   a. CSCtn37522: Delay in L2 port-channels going down
   b. CSCud82316: VPC Convergence optimization
   c. CSCuc50888: High convergence after F2 module OIR

Services Best Practices and Caveats

ASA Firewall Appliance Best Practices

- The ASA FT and stateful links should be dedicated interfaces between the primary and secondary ASA.

- Failover interface policies should be configured to ensure that the security context fails over to the standby ASA if monitored interfaces are down.
- Configure an appropriate port-channel load-balancing scheme on the ASA to ensure that all port-channel interfaces are used to forward traffic out of the ASA.

Copper Implementation Best Practices

- Configure all Copper tenants' servers with either public or private IP addresses, not a mix of both types. If both types are needed, use seperate ASA context for all public addressed tenants and a separate context for all private addressed tenants.
- Private IP addresses for servers can be overlapped for different tenants, and requires the use of NAT with separate public IP addresses per tenant for outside.

Compute Firewall Best Practices

- The VSG does not support vSphere HA and DRS. On clusters dedicated for hosting VSG virtual appliances, disable vSphere HA and DRS. On clusters hosting both VSG virtual appliances and other VMs, disable HA and DRS for the VSG virtual appliances.
- For a VSG HA-pair, the primary and secondary nodes should be hosted on the same ESXi host. Use vSphere anti-affinity rules or DRS groups to ensure this.
- Each VSG virtual appliance (be it active or standby node) reserves CPU and memory resources from the ESXi host. Make sure the ESXi host has enough unreserved CPU and memory resources, otherwise, the VSG virtual appliance will not power on.
- Make sure that the clocks on the VNMC, VSGs, and Nexus 1000V are synchronized. The VSGs and Nexus 1000V will not be able to register to the VNMC if the clocks are too out of sync.
- Enable IP proxy ARP on the router interface(s) on the subnet/VLAN facing the VSG data interfaces.
- On the VNMC, compute firewalls should be added at the tenant level or below, and not at the Root org level.
- For the tenant, the DMZ should have its own VSG compute firewall, separate from the firewall used on the PVT zone.
- When configuring security policies/rules on the VNMC, the attributes used for filtering conditions should be preferred in the following order:
  - Network attributes, most prefer, providing highest performance for VSG
  - VM Attributes
  - vZone, lest prefer, lowest VSG performance

Compute Firewall Caveats

- The VSG FTP/TFTP protocol inspect on the VSG fails to open the pinhole required for data connection when the source and destination vNICs are under the same VSG protection, but on different VLANs. See CSCud39323 for more details.

ACE 4710 Appliance Best Practices

- The FT VLAN should be configured using the **ft-port vlan <vlan-id>** command to ensure that FT packets have the right QoS labels. This ensures that proper treatment is given to ACE FT packets in the network.
- Configure an appropriate port-channel load-balancing scheme to ensure that all port-channel interfaces are used to forward traffic out of the ACE appliance.

- To avoid MAC collision among operational ACE appliances on the same VLAN, use an appropriate shared-vlan host-id <1-16> to ensure that each ACE appliance has a unique MAC address on a shared VLAN.

QoS Best Practices and Caveats

Nexus 1000V QoS Best Practices

- Configure QoS policies to classify, mark, police, and prioritize traffic flows. Different traffic types should have different network treatment.

UCS QoS Best Practices

- Reserve bandwidth for each traffic type using QoS system class. Each type of traffic should have a guaranteed minimum bandwidth.
- For UCS servers deployed with the Nexus 1000V, it is highly recommended to do the CoS marking at the Nexus 1000V level. Configure UCS QoS policy with **Host Control Full** and attach the policy to all vNICs of UCS servers.

ACE QoS Best Practices and Caveats Caveats

- **CSCtt19577:** need ACE to preserve L7 traffic dot1p CoS
  - QoS transparency requires that DSCP not be touched, and that only CoS be used to support DC QoS in the VMDC system. The tenant uses DSCP for their markings, and the DC operator can use independent QoS markings by using dot1P CoS bits. To support this, both DSCP and dot1p CoS need to be preserved as packets transit the ACE, however, the ACE does not currently support CoS preservation for L7 traffic. This enhancement requests support for CoS preservation and DSCP preservation for all scenarios including L7 traffic.

Nexus 7000 QoS Best Practices and Caveats Best Practices

- The Nexus 7000 series uses four fabric queues across modules, and CoS values are mapped to these four queues statically, i.e., they cannot be changed. The priority queue for CoS5,6, and 7 is switched with strict priority, and the other three queues are switched with equal weights. The F2 cards used in VMDC 2.3 use the 8e-4q4q model, which class-maps that map to the CoS values in the same way as the fabric queues. This is particularly important as the F2 card uses buffers in the ingress card, and back pressure from the egress interface congestion is mapped to ingress queues. Packets are dropped at ingress when such congestion happens. It is important to use the 8e-4q4q model to track each class separately. This model is supported from NX-OS release 6.1.3 onwards.

Caveats

- **CSCue55938:** duplicating policy-maps for egress queuing.
  - Attaching two queuing policies for the same direction under a port is allowed under some conditions.
- **CSCud46159:** all interfaces in the module are gone after reboot
  - When a scaled up configuration with many interfaces is configured with the same policy-map that includes egress policing, upon reload, the Nexus 7004 aggregation switch loses its configuration of all interfaces. This workaround is to configure multiple policy-maps with the same policy and divide the total number of subinterfaces into three or four groups and attaching a different policy-map to each group.
- **CSCud26031:** F2: aclqos crash on configuring QoS policy on subinterfaces

  - ACLQOS crash is observed when attaching a service policy that includes egress policing on a large number of subinterfaces. The workaround is to use different policy-maps (with the same underlying policy) so that the number of subinterfaces using the same policy-map is reduced.
- **CSCud26041:** F2: scale QoS configs by not allocating policer stats when no policing
  - Qos per class stats use hardware resources that are shared with policers. On the F-series card, this is restricted to a small amount, i.e., currently 1024, which is the total of all classes in policies multiplied by attachments. For example, with an eight-class policy, only 128 attachments can be done on 128 subinterfaces on the same SoC. This bug requests disabling default per-class statistics collection and providing proper error messaging to indicate the actual issue. Statistics are enabled by default, and hence the workaround is to add **no-stats** to the service policy attachments.

Nexus 5000 QoS Best Practices and Caveats Best Practices

- Use all six classes of traffic for the Ethernet class if no FCoE traffic is expected.
- Account for NFS traffic at this layer of the DC, and provide a separate class and queuing to provide a BW guarantee.

Caveats

- **CSCue88052:** Consistency between Nexus 5000 and Nexus 7000 QoS config
  - Nexus 5500 Series switches currently have different semantics of similar sounding QoS configuration items, and this bug tracks specifically the fact that the Nexus 5500 allows the configuration of bandwidth percent for a class in a policy-map where priority is configured. Also, the bandwidth percent semantics in a policy-map that has priority class is actually called "bandwidth remaining." This is confusing and not consistent with the Nexus 7000 semantics, which have checks in place to prevent priority and bandwidth percent configuration for the same class in a policy-map.

ASR 1000 QoS Best Practices and Caveats Best Practices

- QoS on port-channel interfaces is not supported. For the MPLS-Core facing interfaces, port-channels are not recommended, as the VMDC 2.3 QoS policies cannot be implemented.
- QoS on port-channel subinterfaces have restrictions. For example, ingress QoS cannot be done in flow-based mode, and egress QoS requires a QoS configuration on the member links. The recommendation for VMDC 2.3 is to use multiple links between the DC-PE and DC-AGG if more than 10GE is required.
- NetFlow on the ASR 1000 series with custom NetFlow records can impact the switching performance. The recommendation is to use default NetFlow record formats. While this is not exactly a QoS best practice, this can impact QoS due to dropping of packets earlier than expected due to switching performance rather than actual link congestion.
- Mapping of priority traffic based on CoS and MPLS-TC to the high-priority queue between SIP and the ESP is required to provide priority traffic low latency treatment.
- Calculation of bandwidth requirement for both normal and failure cases should be accounted for, as the ASR 1000 is a centralized switching platform and all traffic is funneled and switched at the ESP. In this design, a SIP-40 is used with 4x10GE shared port adapters, and with ESP-40, which can handle 40 Gbps of switching. This provides 10 Gbps of traffic from north-south, and 10 Gbps of traffic from south-north, for a total of 20 Gbps for normal conditions. Different failure scenarios will not cause any oversubscription at the ESP-40.

Caveats

- **CSCud51708:** wrong calc for bytes w ms based queue-limit config after random-detect
  - If the queue-limit is configured in milliseconds after configuring random-detect, the bytes calculation is wrong for the specified number of milliseconds in the queue-limit. The workaround is to first configure the queue-limit in milliseconds and then configure random-detect.