HP System Dimensioning Guide, CEE R6

Cloud Execution Environment

CONFIGURATION MODEL

2/192 02-AZE 102 01/5-V1 Uen A

Contents

1 Introduction

This document describes the characteristics of Cloud Execution Environment (CEE) to enable dimensioning and understanding the limitations of CEE. It also describes requirements on HW combinations required for running CEE. The application can have additional requirements.

Storage is measured in gibibyte (GiB), tebibyte (TiB), and mebibyte (MiB) in this document.
1 GiB is equivalent to 1.074 GB.

The following words are used in this document with the meaning specified below:

vNIC	A virtual network interface card (vNIC) provides connectivity between the Cloud SDN Switch (CSS) and a VM. A configuration can provide several vNICs to a VM.
Interface	A network interface. Can be either a physical NIC (PHY) providing CSS with board external connectivity, or a virtual NIC connecting CSS to a VM.
PMD thread	CSS uses a Poll Mode Driver (PMD) technique that continuously polls incoming packets from the NICs, that is, interrupts are not used. To be able to reliably handle all incoming packets, a software continuously polls the NIC queues for packets to be handled. This software is executing in one or more threads that are called PMD threads. The execution environment for the PMD threads is isolated from the Linux scheduler to be able to reliably handle a high sustained packet flow without interrupts or delays caused by being scheduled out.
vCIC host	A host running Compute and vCIC. There are three vCIC hosts in Multi-Server deployments of CEE.
Compute host	A host running Compute without vCIC

1.1 Target Group

Cloud Infrastructure providers and application designers.

1.2 System Characteristics

The characteristic features of CEE are the following:

- High Availability cloud system
- Support for up to 80 physical blades
- OpenStack® SW deployment (for deviations, see Section 5.1 on page 32)

For information on system limitations, see Section 5 on page 32.

2 CEE System

This section describes CEE system configurations and capabilities.

2.1 System Configurations

CEE is a scalable system used with HW products from different vendors. CEE offers a set of configurations.

2.1.1 Reference Configurations

Reference configurations, as shown in Table 1, offer the possibility to build and scale a cloud system on specified HW, such as servers, switches, and storage.

In CEE R6, HP c7000 HW can scale depending on the switch configuration used:

- A small reference configuration consists of shared LAN/SAN Extreme X670V switches supporting up to 24 server blades (2 enclosures).
- A medium reference configuration consists of separate LAN/SAN Extreme X670V switches supporting up to 47 server blades (3 enclosures).
- A large reference configuration consists of separate LAN/SAN Extreme X770 switches supporting up to 80 server blades (5 enclosures).

Table 1 HP Reference Configurations

Scaling	Number of Blades	Number of Enclosures
Small Configuration	24	2
Medium Configuration	47	3
Large Configuration	80	5

2.1.2 Certified Configuration

Certified configurations are a subset of reference configurations. Certified CEE configurations have been documented and tested by the Ericsson Cloud organization.

The certified configuration for CEE R6 consists of the following HW components:

- HP c7000 with 48 server blades (3 enclosures)
- Separate LAN/SAN Extreme X770 switches
- Extreme X460 control switches
- EMC VNX5400 centralized storage (with a specific disk drive configuration and a specific iSCSI I/O module configuration)

3 HW Requirements

This section describes generic hardware and firmware requirements for CEE, based on the certified Ericsson CEE R6 HW.

Complying with the outlined HW requirements does not in any way imply Ericsson CEE certification. HW deployment outside the already certified ones requires system integration work.

The server, switching, and optional storage components of the HW environment are shown in the figures below. Two storage options are available:

- The hardware with optional VNX storage is shown in Figure 1.
- The hardware with optional ScaleIO storage is shown in Figure 2.

The two different storage solutions cannot be combined. Only one of them can be used in the same CEE system.

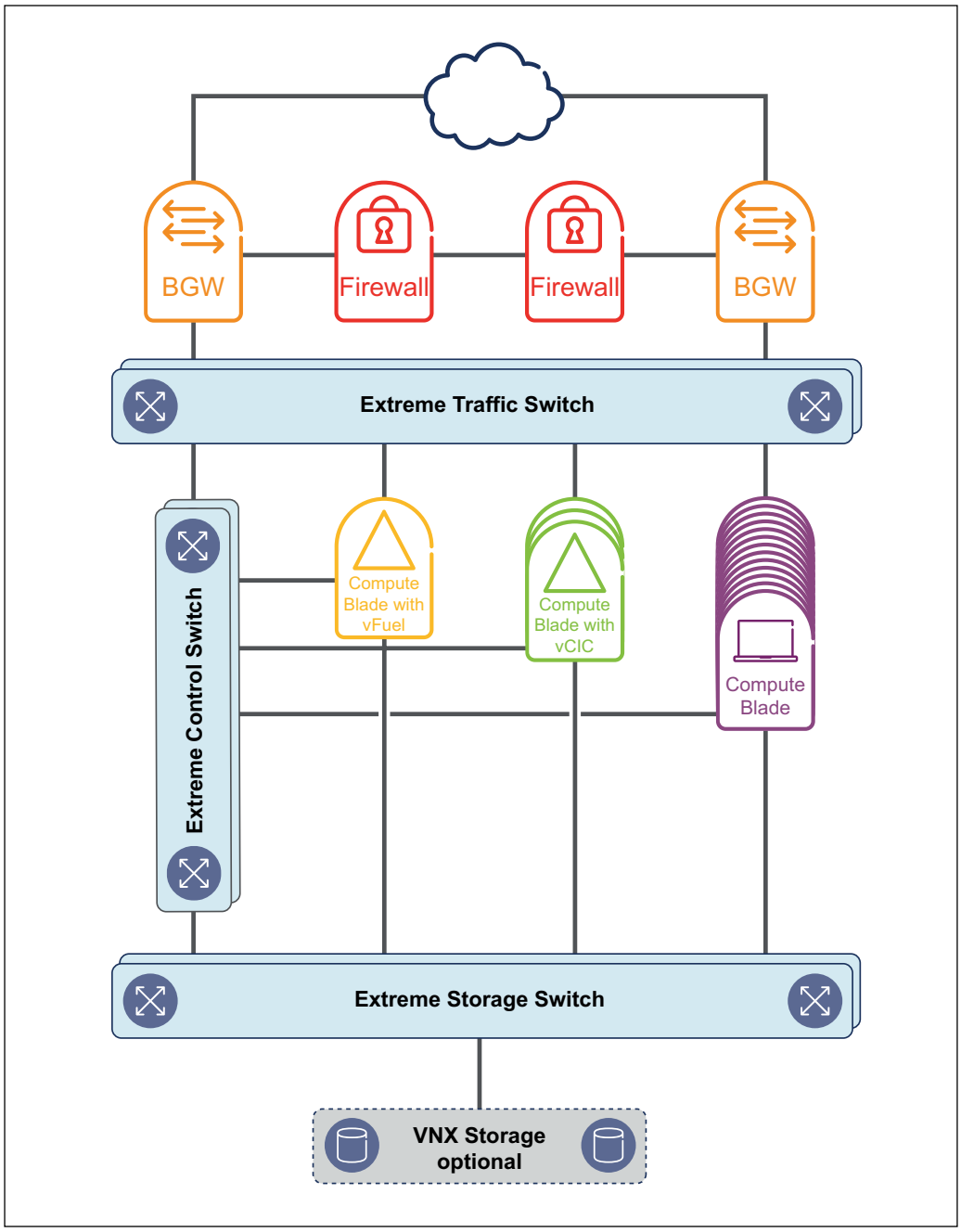

Figure 1 CEE Hardware Environment with Optional VNX Storage

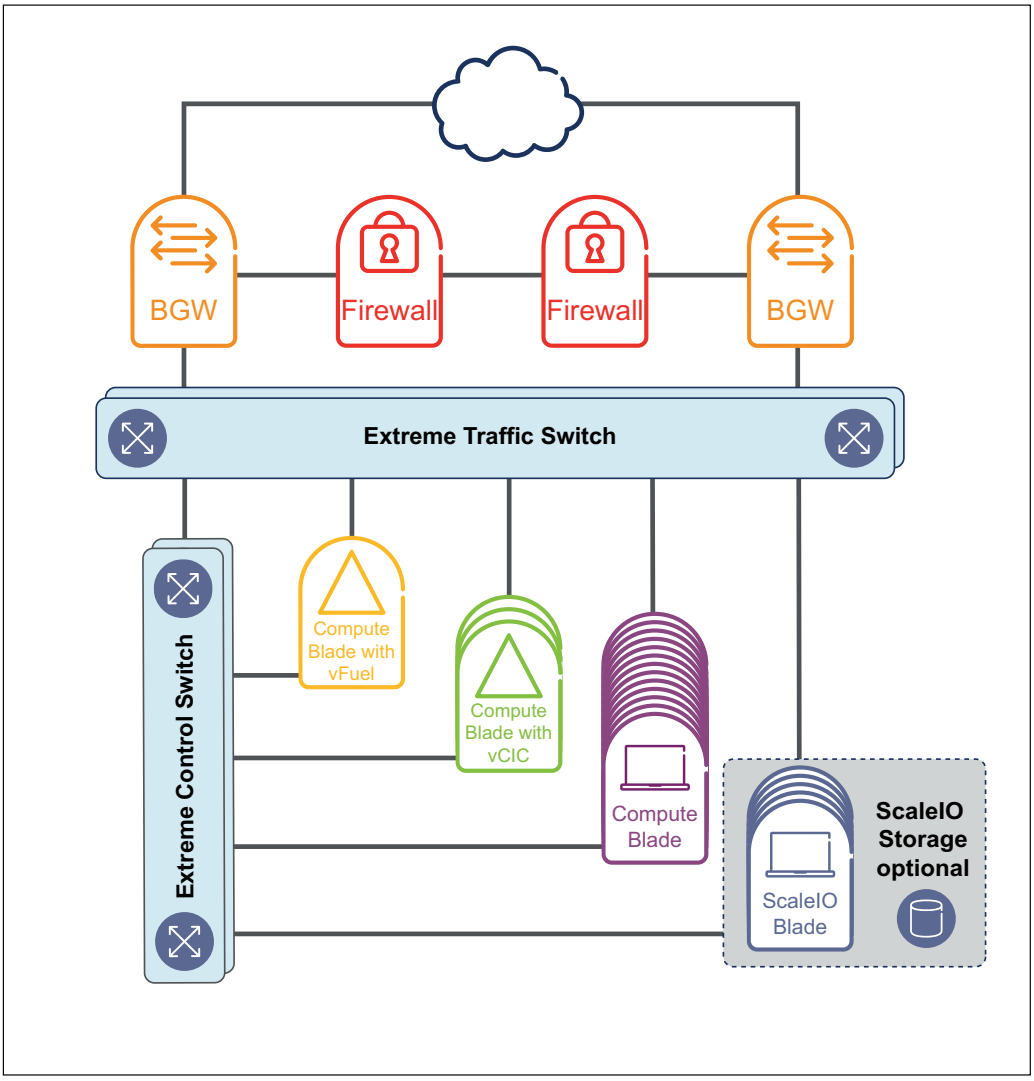

Figure 2 CEE Hardware Environment with Optional ScaleIO Storage

The supported HW is listed below:

Table 2 Supported HP HW

Item	Status
Blades	
HP C7000 BL460c Gen8	Reference
HP C7000 BL460c Gen9	Certified
Networking	
Extreme x670V	Reference
Extreme x770	Certified
Extreme x460	Certified

Item	Status
Storage	
EMC VNX 5400	Certified
Other EMC VNX® series 2 models	Reference
RAID-0 for local disk(1)	Certified
RAID-1 for local disk	Certified

(1) The use of RAID-1 is recommended.

3.1 Summary of Blade Configuration

This section describes HW requirements for blades.

Table 3 HW Configuration Example, Blade

Aspect	Requirement	Limitation
CPU(1)	2x Intel Xeon Processor E5-2680 v3 (12 cores per processor, 2 Hyperthreads per core), 48 Hyperthreads available	Max 21 VMs/2HTs non-oversubscribed NFVs
RAM	At least 64 GiB. Recommended 128 GiB or more, up to maximum capacity.	
NIC	4*10GE Intel Niantic and 2*1GE unspecified (Management)(2)	
Onboard Disk	At least 600 GB (558 GiB) SSD or HDD	
Management Interace	HP iLO/OA	

(1) 3 cores, 6 HTs used for infrastructure, additional cores used by VMs. Requirement depends on application.
(2) Additional NICs are needed if SR-IOV is used.

Note: SR-IOV is not supported for HP HW.

3.2 Network Configuration

This section describes HW configuration for networking.

Table 4 HW Configuration, Network

Aspect	Requirement	Limitation
LAN/SAN Switch(1)	4 × X770, use for up to 80 blades	HW allows up to 80 blades to be connected.
	4 × X670V, use for up to 47 blades	
	2 × X670V where LAN and SAN are mixed in each switch, use for up to 24 blades	
DC Firewall	See Section 3.3 on page 8 for information about selecting the DC Firewall	
Border Gateway	BGW is essentially a router with the following HW and SW requirements: • Two BGW units for redundancy • At least four 10 GE or one 40 GE interface for the configuration • Support of Virtual Router Redundancy Protocol (VRRP) v3	

(1) Bandwidth requirements towards BGw and storage influences switching configuration and number of blades that can be connected. Required and foreseen expansion scenarios must be considered because reconfiguration means a rebuild of the switch solution.

3.2.1 Network Requirements for Distributed Storage, ScaleIO

The following requirements must be fulfilled:

Network:

- Bandwidth:
  - Minimal: 2 x 10 Gbit
  - Optimal: 4 x 10 Gbit, to separate front-end and back-end traffic
  - Additional 2 x 1 Gbit network for management traffic
- All the ScaleIO components are accessible via the 10 Gbit network.
- Network bandwidth and latency between all nodes is acceptable, according to application demands.
- The Ethernet switch supports the required bandwidth between network nodes.
- MTU settings are consistent across all servers and switches.

- The following TCP ports are not used by any other application, and are open in the local firewall of the server:
  - Meta Data Manager (MDM): 6611 and 9011
  - ScaleIO Data Server (SDS): 7072, for multiple SDS: 7073-7076
  - ScaleIO Gateway (including REST Gateway, Installation Manager, and SNMP trap sender): 80 and 443
  - Light Installation Agent (LIA): 9099

Parameters and configuration in ScaleIO that will affect dimensioning:

- SDS network limits can be set to avoid overloading the storage network with huge amount of rebuild or rebalance traffic. Refer to the **SDS network limits** section in the *EMC ScaleIO Version 2.0.x User Guide*.

3.3 Firewall Requirement

During test of the CEE Certified Configuration, a Juniper EX 4550 has been used as BGW and a Juniper SRX 3400 as Firewall.

All Juniper SRX High-End series Firewalls provide a similar set of features. SRX can be equipped with a flexible number of Services Processing Cards (SPC), I/O Cards (IOC), and Network Processing Cards (NPC), allowing the system configuration to support a balance of performance and port density. Suitable firewall HW setup depends on expected traffic volumes and sessions. More information can be found in supplier datasheets, on the Juniper Networks homepage, Reference [1].

3.4 CPU Configuration

Refer to the *Configuration File Guide* for more information about the configuration procedure.

The number of available CPU IDs depends on the CPU model.

3.4.1 Compute Host

The allocation examples below are valid for HP BL460c G9 with Intel Xeon E5-2680 v3 processor.

- Table 5 and Figure 3 give an example for automatic CPU allocation for VMs, OVS, vCIC, and vFuel on a Compute node. CPUs are automatically allocated for OVS on both Numa nodes, although both NICS for tenant traffic are connected to Numa 0. The CPU allocation must be adjusted manually.

- Figure 4 shows the manually adjusted CPU allocation on a Compute node where CPU 23 and CPU 47 changed from OVS to Tenant VM.

If the number of cores in the used processor type differs from the number of cores in this example, the number of cores allocated to the VMs must be adjusted.

The examples describe the CPU allocation at a server that runs vCIC and vFuel. If vCIC or vFuel are not used on the server, their CPUs are allocated for the tenant VMs.

Table 5 Automatic CPU Allocation on HP BL460c G9 with Intel Xeon E5-2680 v3 Processor for VMs, OVS, vCIC, and vFuel in a Compute Host

CPU Owner	**Allocated CPU ID**
Tenant VM	8,32, 9,33, 10,34, 11,35, 12,36, 13,37, 14,38, 15,39, 16,40, 17,41, 18,42, 19,43, 20,44, 21,45, 22,46
OVS(1)(2)	1,25, **23,47**(3)
OVS control process(4)	0(4)
vCIC(5)	4,28, 5,29, 6,30, 7,31
vFuel	3,27
Host OS(6)	0,24, 2,26
If ScaleIO is installed, then 1–5 CPU threads are used for SDC.	

(1) OVS configuration requires at least one PMD thread on each NUMA node with a physical interface, and at least one PMD thread on the NUMA node where the control threads are located.
(2) To achieve a more predictable performance, allocate only one CPU per core for OVS. See Section 3.4.1.1 on page 11 for more information.
(3) The allocation of CPU 23 and 47 must be manually changed from OVS to Tenant VM since Numa 1 has no NICs for tenant traffic.
(4) The process does not get a CPU for its exclusive use. A configuration parameter specifies one of the host OS CPUs to be used by the OVS control process.
(5) The vCIC can be allocated on cores that are over allocated and shared with the application. In such cases, it must be ensured that the application sharing resources with the vCIC dont exhaust the vCIC resources.
(6) In some use-cases only one core is allocated to the host OS as explained in Section 3.4.1.2 on page 11.

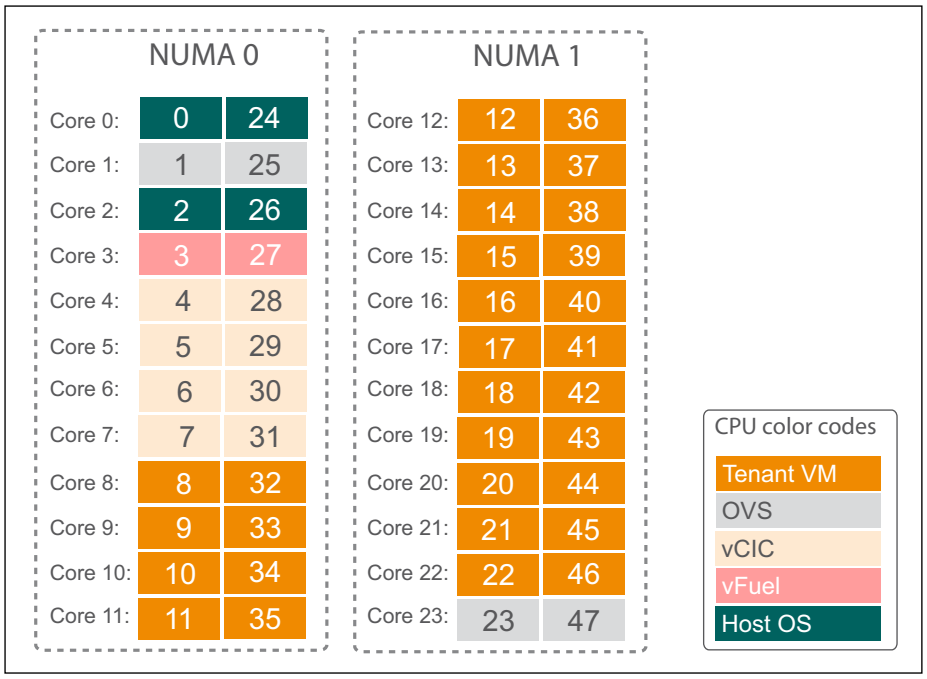

Figure 3 Automatic CPU Allocation of the Respective Resource Owner per NUMA Node on HP BL460c G9 with Intel Xeon E5-2680 v3 Processor for VMs, OVS, vCIC, and vFuel in a Compute Host

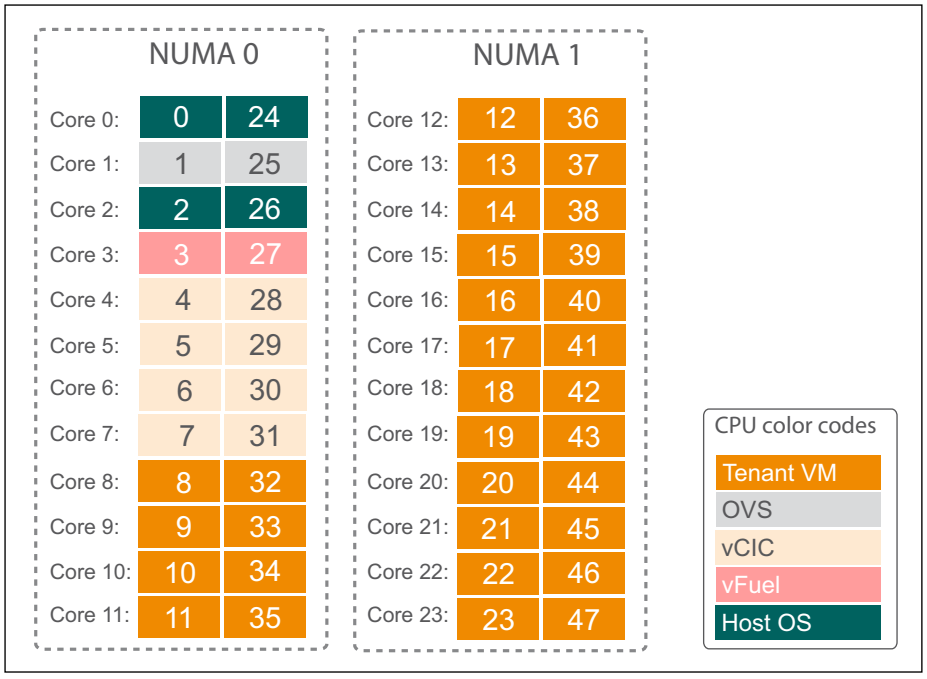

Figure 4 Manually Adjusted CPU Allocation of the Respective Resource Owner per NUMA Node on HP BL460c G9 with Intel Xeon E5-2680 v3 Processor for VMs, OVS, vCIC, and vFuel in a Compute Host

3.4.1.1 Allocating Single CPUs for OVS

To achieve a more predictable OVS performance, only a single CPU must be allocated to OVS from each CPU core reserved for OVS PMD threads. The other CPU on the same core, also called hyper-thread sibling, must be isolated, that is, not used by the host OS process scheduler, and is free from any extra load that can negatively influence the OVS performance.

To achieve the described allocation, the following settings must be performed:

- Refer to the *Configuration File Guide* for the following setting in the configuration file:

  The other CPU in the core must be reserved for the `idle` owner. By this reservation, the following results are achieved:

  - The idle CPUs are listed in the kernel's `isolcpus` boot parameter. It ensures that the process scheduler does not assign any process to these CPUs.
  - The idle CPUs are listed in the kernel's `nohz_full` boot parameter. It ensures that the kernel does not generate scheduling clock interrupts on these CPUs.

- Refer to *SW Installation in Multi-Server Deployment* for the following setting at each Compute host:

  The idle CPUs must be listed in the `noirqs` parameter of the `configure-interrupts` upstart task in order to avoid these CPUs from using them for serving interrupt requests.

3.4.1.2 Allocating One Core to the Host OS

However the recommended CPU allocation to the host OS is two cores for each Compute host, for some use-cases, where a single or very few VMs are used on each Compute host, it can be possible to reduce the host OS allocation to one core.
Having a single core allocated for the host OS can impact the characteristics of the system. For example, actions such as starting, stopping, and migrating VMs can become slower, and it can also impact the performance of other VMs on the same host. Before using a system with a single core allocated to the Host OS in a production environment, sufficient testing must be performed. Beside other checks, this testing must include the capacity and performance test of the VM during the time of performing life cycle management operations for other VMs running on the same host. While performing such test, the processor load for the host OS must be monitored carefully for deviation from a steady state.

3.4.2 ScaleIO Host

Table 6 shows the minimum number of cores to be allocated to the CPU owners in a ScaleIO host.

Table 6 Minimum Number of Allocated Cores in a ScaleIO Host

CPU Owner	Minimum Number of Allocated Cores
Host OS	2
Meta Data Manager/Tie-Breaker (MDM/TB)	1
SDS	2–4

3.5 RAM Configuration

This section describes the optimal RAM configuration for the CEE.

Refer to the *Configuration File Guide* for more information about the configuration procedure.

Each Neutron network created consumes RAM in the vCIC, and it influences the maximum number of virtual tenant networks. See Section 4.4.3 on page 28 for more information.

The minimal RAM size is 128 GiB.

Running vCIC, vFuel, or both on the server modifies the optimal memory allocation. The subsections below describe the possible cases for 128 GiB memory. If the system contains more than 128 GiB RAM, the extra memory goes to the tenant VMs.

See the following cases:

- Server without vCIC and vFuel, see Section 3.5.2 on page 13.
- Server with vCIC, see Section 3.5.3 on page 14.
- Server with vFuel, see Section 3.5.4 on page 15.
- Server with vCIC and vFuel, see Section 3.5.5 on page 16.

Section 3.5.1 on page 12 provides general information about RAM configuration in CEE.

3.5.1 Introduction

The default configuration for the Host OS on Compute nodes is 8 GiB. More memory can be reserved for the Host OS by setting the relevant configuration parameters. The needed amount of memory depends on the values of a number of other parameters as described below.

The memory used for the VMs is allocated to huge pages. This is the memory visible from the inside of the VMs. The 1 GiB huge pages are referred to

as **Tenant VM** in the RAM reservation tables of the upcoming sections. In addition to the 1 GiB huge pages, the VMs need memory allocated from the host OS. This memory is used, for example, to emulate devices used by the virtual machine. It is hard to predict the amount of host OS memory used by the emulator since, for example, it depends on the type and number of the used devices. A small VM consumes less than 100 MiB, while it can grow to several hundred MiB in specific cases. About 300 MiB host OS memory would be enough for each virtual machine but we must double it and calculate with 600 MiB as explained below.

In a system using the NUMA architecture, the NUMA location of VMs must be considered. The available memory, that is, the huge pages and the Host OS memory, are evenly distributed between the NUMA nodes. By design, OpenStack Nova allocates VMs on the first NUMA node that fits the VM. Apart from the VMs running on both NUMA nodes, the VMs allocate memory from the NUMA node on which they are running. In a worst case scenario where all VMs are allocated on the same NUMA node, all the memory for the VMs will be allocated from the same NUMA node. In such a scenario most of the memory on the other NUMA node will be unused, and half of the memory on the Compute node will be free. To be on the safe side in a dual socket system, the 300 MiB host OS memory per VM must be doubled to cover the case where all VMs are allocated on the same NUMA node.

The host OS memory usage for processes other than the VMs depends on the CPU reservations. The host OS uses the unreserved CPUs. By default, CEE allocates two cores to the host OS on NUMA node 0. This means that most of the memory used by the Host OS will be allocated from NUMA node 0. In some scenarios it is preferred to run the host OS on both NUMA nodes. It can be achieved by modifying the CPU reservation.

Note: In order to allocate as little memory for the host OS as possible, memory profiling of the host OS for the specific scenario is recommended.

3.5.2 Compute Server without vCIC and vFuel

Table 7 specifies the volumes allocated to the resource owners. Host Operating System (OS) is not included, so the table contains 120 GiB.

Table 8 contains the total memory sizes.

Table 7 Memory Allocation for System with 128 GiB RAM

RAM Resource Owner	Hugepage size (MiB)	Number of Hugepages (Count)	Total Size (GiB)
Tenant VM(1)	1024	118	118
OVS	2	1024	2

RAM Resource Owner	Hugepage size (MiB)	Number of Hugepages (Count)	Total Size (GiB)
vCIC	-	-	-
vFuel	-	-	-

(1) If Atlas is used, it runs at one of the hosts, and it uses 4 GiB of the RAM configured for the tenant VMs.

Table 8 Total Memory Sizes for System with 128 GiB RAM

RAM Resource Owner	Total Size (GiB)
Tenant VM(1)	118
OVS	2
vCIC	-
vFuel	-
Host OS(2)	8(2)
Total (with host OS)	128

(1) If Atlas is used, it runs at one of the hosts, and it uses 4 GiB of the RAM configured for the tenant VMs.
(2) If ScaleIO is installed, SDC uses 0.05 GiB of the RAM configured for the host OS.

3.5.3 Compute Server with vCIC

Table 9 specifies the volumes allocated to the resource owners. Host OS is not included, so the table contains 114 GiB.

Table 10 contains the total memory sizes.

Table 9 Memory Allocation for System with 128 GiB RAM

RAM Resource Owner	Hugepage size (MiB)	Number of Hugepages (Count)	Total Size (GiB)
Tenant VM(1)	1024	82	82
OVS	2	1024	2
vCIC	1024	30	30
vFuel	-	-	-

(1) If Atlas is used, it runs at one of the hosts, and it uses 4 GiB of the RAM configured for the tenant VMs.

Table 10 Total Memory Sizes for System with 128 GiB RAM

RAM Resource Owner	Total Size (GiB)
Tenant VM(1)	82
OVS	2
vCIC	30
vFuel	-
Host OS(2)	14(2)
Total (with host OS)	128

(1) If Atlas is used, it runs at one of the hosts, and it uses 4 GiB of the RAM configured for the tenant VMs.
(2) If ScaleIO is installed, SDC uses 0.05 GiB of the RAM configured for the host OS.

3.5.4 Compute Server with vFuel

Table 11 specifies the volumes allocated to the resource owners. Host OS is not included, so the table contains 120 GiB.

Table 12 contains the total memory sizes.

Table 11 Memory Allocation for System with 128 GiB RAM

RAM Resource Owner	Hugepage size (MiB)	Number of Hugepages (Count)	Total Size (GiB)
Tenant VM(1)	1024	115	115
OVS	2	1024	2
vCIC	-	-	-
vFuel	1024	3	3

(1) If Atlas is used, it runs at one of the hosts, and it uses 4 GiB of the RAM configured for the tenant VMs.

Table 12 Total Memory Sizes for System with 128 GiB RAM

RAM Resource Owner	Total Size (GiB)
Tenant VM(1)	115
OVS	2
vCIC	-
vFuel	3

RAM Resource Owner	Total Size (GiB)
Host OS(2)	8(2)
Total (with host OS)	128

(1) If Atlas is used, it runs at one of the hosts, and it uses 4 GiB of the RAM configured for the tenant VMs.
(2) If ScaleIO is installed, SDC uses 0.05 GiB of the RAM configured for the host OS.

3.5.5 Compute Server with vCIC and vFuel

Table 13 specifyspecifies the volumes allocated to the resource owners. Host OS is not included, so the table contains 114 GiB

Table 14 contains the total memory sizes.

Table 13 Memory Allocation for System with 128 GiB RAM

RAM Resource Owner	Hugepage size (MiB)	Number of Hugepages (Count)	Total Size (GiB)
Tenant VM(1)	1024	79	79
OVS	2	1024	2
vCIC	1024	30	30
vFuel	1024	3	3

(1) If Atlas is used, it runs at one of the hosts, and it uses 4 GiB of the RAM configured for the tenant VMs.

Table 14 Total Memory Sizes for System with 128 GiB RAM

RAM Resource Owner	Total Size (GiB)
Tenant VM(1)	79
OVS	2
vCIC	30
vFuel	3
Host OS(2)	14(2)
Total (with host OS)	128

(1) If Atlas is used, it runs at one of the hosts, and it uses 4 GiB of the RAM configured for the tenant VMs.
(2) If ScaleIO is installed, SDC uses 0.05 GiB of the RAM configured for the host OS.

3.5.6 ScaleIO Server with MDM/TB and SDS

Table 15 specifies the volumes allocated to the resource owners.

Table 16 contains the total memory sizes.

Table 15 Memory Allocation for System with 128 GiB RAM

RAM Resource Owner	Hugepage size (MiB)	Number of Hugepages (Count)	Total Size (GiB)
MDM/TB			0.5
SDS			0.5

Table 16 Total Memory Sizes for System with 128 GiB RAM

RAM Resource Owner	Total Size (GiB)
MDM/TB	0.5
SDS	0.5
Host OS	8
Total (with host OS)	9

3.6 Storage Configuration

This section describes centralized, local, and distributed storage implementations, and disk requirements.

3.6.1 Centralized Storage

Centralized Storage can be used as the back end for Cinder.

Table 17 HW Configuration, Centralized Storage

Aspect	Requirement
Storage System Enclosure	EMC² VNX series 2 model, for example, VNX 5400
Disk drive configuration (for example, SSDs or rotating disks)	Dimensioning needed
iSCSI IO Module Pairs	1–2

Storage system dimensioning needs separate design work, for more information refer to the EMC homepage, Reference [2].

3.6.2 Local Storage Disk Space

in Table 18 and Table 19.

This section lists requirements on disk space in Table 18 and Table 19.

Table 18 vCIC Host and Compute Host Disk Allocation

Use	vCIC Host	Compute Host	Note
Root partition (host OS)	50 GiB	50 GiB	
Logs and core/crash dumps	40 GiB	40 GiB	When 40 GiB is allocated to log partition, 10 GiB is allocated to logs and 30 GiB is allocated to core/crash dumps.
vCIC total	240–320 GiB + additional space for Glance	-	vCIC backup is not included. See Table 3.
vCIC backup	52 GiB	-	
Host total without vFuel and Atlas	**382–462 GiB**	**90 GiB**	
vFuel	50 GiB	50 GiB	vFuel can be run on any Compute host and it uses disk from the ephemeral storage. **Limitation:** Only two of the following can be run on the same Compute host: • vCIC • vFuel • Atlas

Use	vCIC Host	Compute Host	Note
Atlas	130 GiB	130 GiB	Atlas can be run on any Compute host and it uses disk from the ephemeral storage. The disk for Atlas is allocated by the CEE. **Limitation:** Only two of the following can be run on the same Compute host: • vCIC • vFuel • Atlas
Remaining storage is for ephemeral storage for tenant VMs.	Dimensioned depending on application need.	Dimensioned depending on application need.	Valid for boot from image. Calculated from total disk reduced by used space.

Table 19 vCIC Disk Allocation

Use	vCIC	Note
Root partition (host OS)	50 GiB	
Logs and crash dumps	40 GiB	

Use	vCIC	Note
Database for OpenStack and Zabbix (MySQL)	40–120	The size needs to be adjusted to the storage needs of Zabbix. The database size depends on several factors, for example, the number of blades and the number of vNICs. For most installations, allow 1 to 1.5 GiB per host, plus at least 10 GiB margin to the total. For example, allow at least 60 GiB for a 47 blade deployment and between 90 and 130 GiB for an 80 blade deployment.
Database for Ceilometer (MongoDB)	70 GiB	The size needs to be adjusted to the storage needs for Ceilometer.
Glance repository in Swift	> 40 GiB(1)	The size needs to be adjusted depending on the amount and size of images stored in Glance.
vCIC sum	240–320 GiB + additional space for Glance	

(1) Dimensioned depending on application need.

3.6.3 Disk Requirements for Atlas

When Virtual Machine (VM) images are loaded to Atlas as part of an `.ova` file, the image is temporarily stored in ephemeral storage in Atlas. To support loading of large images, the recommendation is to use 120 GiB for the Atlas ephemeral storage.

In case the local disk is used as ephemeral storage (no centralized storage), the Atlas VM occupies 120 GiB of the local disk on the compute node where it is running.

To reduce the disk allocated to Atlas, the size of the ephemeral disk can be reduced from 120 GiB to a minimum of 10 GiB. Since 30% of the ephemeral disk in Altas is used as temporary storage for .ova files, the size of the ephemeral disk needs to be adjusted according to the size of .ova files to be loaded. Using a reduced disk size of 10 GiB implies that it can be impossible to load .ova files that contain images larger than 3 GiB.

3.6.4 Disk Requirements for Nova Snapshots

Nova snapshots are stored in the `/var/lib/glance` partition of CIC nodes.

There are certain disk requirements for the Nova snapshots to work. Depending on the requirements and frequency on Nova snapshots, the system must be dimensioned with free disk space, according to the following guidelines:

- Disk partition `/var/lib/nova` in the compute host where the VM is hosted, must have **at least** double the space of the snapshot/VM size, for a successful Nova snapshot. The reason is that the snapshot will be first extracted locally in the compute node before it is uploaded to the Glance/Swift store.
- The disk space needed in the `/var/lib/nova` partition of the compute disk must have free space **at least** twice the size of VMs root disk. The reason is that during the extraction of the snapshot, first the delta of the VM disk will be extracted, after which the complete disk will be extracted.
- Disk partition `/var/lib/glance` in each CIC node must have free space **at least** equal to the root disk size of the VM, in order to accommodate the snapshot.

3.6.5 Disk Requirements for Distributed Storage (ScaleIO)

Distributed Storage is optional. It can be used as the back end for Cinder.

The following requirements must be fulfilled:

- Needed disk space for ScaleIO component on server: 1 GB
- Minimum disk space to be added as device to one SDS: 100 GB (this must be a physical disk)
- Minimum number of SDSs: 3

4 Characteristics

This section describes the system characteristics of CEE.

4.1 General System Limits

For the list of system limits, see Table 20.

Table 20 General System Limits

Slogan	Limit
RAM used by the infrastructure[1] See Section 3.5 on page 12 for more information.	Compute Host: • Host OS: 8 GiB • OVS: 2 GiB • vFuel (if present): 3 GiB
	vCIC Host: • Host OS: 14 GiB • OVS: 2 GiB • vCIC: 30 GiB • vFuel (if present): 3 GiB

Slogan	Limit
Number of hosts (blades)	The maximum blade configuration within a CEE region is 80 blades. CEE supports 3–80 blades.
Number of cores occupied by infrastructure	CEE infrastructure uses the following on all Compute hosts (including vCIC and vFuel hosts): • 2 cores occupied by Host OS(2) • 1–4 cores occupied by OVS, see Section 4.4.1 on page 24. In addition to allocation to the Compute host running vCIC and vFUel, one vCIC and one vFuel instance can be combined on one Compute host configured for such usage: • vCIC uses 4 cores in the certified configuration. Three instances of vCIC are needed, and each of these must run on a separate Compute host that is configured as a vCIC host.(3) • vFUel uses 1 core. Two instances of vFuel must be allocated.

(1) If Atlas is used, it runs at one of the hosts, and it uses 4 GiB of the RAM configured for the tenant VMs.
(2) The certified configuration use 2 cores for the host OS. Depending on the load generated and the characteristics required by the VNF, this can be reduced to 1 core. Careful load measurements are needed for such tuning, and the instructions for performing such measurements are outside the scope of this document.
(3) The capacity and characteristics for a vCIC using 4 cores have not been verified. Depending on the characteristics needed by the CEE infrastructure and the size of the CEE region, the vCIC can require more resources (cores).

4.2 Orchestration Interface

The system limits for orchestration are listed in Table 21.

Table 21 Orchestration Limits

Slogan	Limits
Number of tenants	The maximum number of supported tenants is 50.

4.3 Tenant Execution Environment

This section describes the tenant-related limits on the environment.

4.3.1 Performance

Performance limits are listed in Table 22.

Table 22 Tenant Execution Performance

Slogan	Limits
Oversubscription	CPU, memory, and disk overcommit are not supported.

4.3.2 Resiliency

Resiliency-related tenant limits are listed in Table 23.

Table 23 Tenant Execution Resiliency

Slogan	Limits
Execution environment resiliency	The execution environment resiliency is relying on VM evacuation. States not conserved in storage are lost. Each hypervisor instance is not redundant, and, apart from attached storage, assumed to be a knock-out unit.

4.4 Network

This section lists the limits on the network.

Neutron with VLAN segmentation is used.

4.4.1 Performance

The virtual switch (vSwitch) performance is measured by the packet rate (packets per second). The packet size has a very limited impact on the packet rate.

Note: Therefore, the forwarded amount of data (bit per second) increases if the packet size is increased.

The CSS executes a configurable number of threads running in endless loops, called PMD threads. Each PMD thread polls interfaces that are automatically assigned to it, processes the incoming packets, and puts them into a queue to be transmitted. The VM interfaces are polled by PMD threads located on the same NUMA node as their OVS control thread.

If the VM and the PMD thread polling the VM are located on different NUMA nodes, the maximum performance (packets per second) decreases since the packets must cross the NUMA border, and it increases the time for accessing the memory. A similar traffic capacity drop occurs if the interfaces are located on different NUMA nodes, since all the traffic must cross the NUMA border.

Table 24 shows the throughput for PHY to VM traffic cases, and Table 25 for VM to VM cases. Table 26 provides dimensioning guidelines for specifying the amount of capacity that is safe to use. Table 27 describes vSwitch capacity for bandwidth-based scheduling.

Table 24 Measured Per Host Forwarding Capacity, 64 Byte Frames, PHY to VM

Slogan	Limits
Bidirectional traffic from PHY on NUMA 0 to VM on NUMA 1	One PMD core is allocated to CSS. One HT is used by CSS, the other is idle. Value = 3.35 Mpps
	One PMD core is allocated to CSS. Both HTs are used by CSS. This allocation is used if `auto` is set as value for core allocation for OVS. Value = 4.28 Mpps. This is the capacity expected from the auto configuration.
	Two PMD cores allocated to CSS. One HT in each core is used by CSS, the other is idle. Value = 5.65 Mpps
	Two PMD cores are allocated to CSS. CSS uses both HTs in each core. Value = 4.11 Mpps
	Four cores are allocated to CSS. One HT in each core is used by CSS, the other is idle. Value = 5.69 Mpps

Table 25 Measured Guest VM Delivery Forwarding Capacity, 64 Byte Frames, VM to VM Intrahost Traffic

Slogan	Limits
Bidirectional traffic from VM on NUMA 0 to VM on NUMA 1	One PMD core is allocated to CSS on each NUMA node. One HT in each core is used by CSS, the other is idle. Value = 2.35 Mpps
	One PMD core is allocated to CSS on NUMA node 0, and one core is allocated on NUMA node 1. CSS uses both HTs on NUMA node 0, and one HT on NUMA node 1, the other HT on NUMA node 1 is idle. Value = 3.70 Mpps
	Two PMD cores are allocated to CSS on NUMA node 0, and one core is allocated on NUMA node 1. One HT in each core is used by CSS, the other is idle. Value = 4.27 Mpps
	Two PMD cores are allocated to CSS on NUMA node 0, and one core is allocated on NUMA node 1. CSS uses both HTs of each core on NUMA node 0, and one HT on NUMA node 1, the other HT on NUMA node 1 is idle. One vNIC per VM. Value = 3.52 Mpps
	Four cores are allocated to CSS on NUMA node 0, and one core is allocated on NUMA node 1. One HT in each core is used by CSS, the other is idle. One vNIC per VM. Value = 5.69 Mpps

Table 26 Dimensioning Capacity (Bidirectional Traffic)

Slogan	Limits
Total vSwitch capacity (bidirectional traffic)	The total vSwitch capacity to be used for dimensioning is 80% of the per host forwarding value above for the number of allocated cores for OVS PMD threads on NUMA node 0. It is different from the value of the ‘‘Per Host Forwarding Capacity’’, in order to take into account external effects impacting the deterministic behavior of the vSwitch. The user can use a different value tuned for a specific system configuration, preferably based on measurements.
Per interface vSwitch capacity (bidirectional traffic)	The maximum dimensioning limit per interface is 80% of the per host forwarding value for one PMD core allocated to OVS, when the HT functionality is not used. If HT is used on the cores hosting OVS PMD threads, the value is 50% of the per host forwarding value. If the number of interfaces is not bigger than the number of PMD threads, the 2 core values can be used as base, but this is not a likely scenario. If more interfaces are configured than OVS-assigned PMD threads, the maximum dimensioning limit per interface is reduced by an additional factor: the number of interfaces divided by the number of PMD threads, rounding any fraction to the next higher integer. Two examples: 7 interfaces and 3 PMD threads => 7/3 = 2.33, round up => dividend is 3, which means that the capacity figure should be divided by 3; 9 interfaces and 3 PMD threads => 9/3 = 3, there is no fraction so no round up => dividend is 3, which means that the capacity figure should be divided by 3. It is not recommended to change the per interface limit, even if measurements indicate that it is possible, as the behavior is highly dependent on the automatic distribution of the interfaces over the PMD threads.

Table 27 Virtual Switch Capacity for Bandwidth-Based Scheduling

Slogan	Limits
vSwitch capacity for bandwidth-based scheduling	The value is used as maximum threshold for the vSwitch bidirectional throughput per host. It is used to configure bandwidth-based scheduling per host when installing CEE. It must be below or, at most, equal to ‘‘Total vSwitch capacity (bidirectional traffic)’’ detailed in Table 26.

4.4.2 Resiliency

Network resiliency is listed in Table 28.

Table 28 Network Resiliency

Slogan	Limits
Self-healing network	The network solution is self-healing, including network fault detection and automated failover.

4.4.3 Tenant Network Limitations

Limitations of the tenant network are listed in Table 29.

Table 29 Tenant Network Limitations

Slogan	Limits
Number of virtual networks	The theoretical aggregated maximum number of virtual tenant networks per CEE region is 4050. Since each Neutron network created consumes RAM in the vCIC, this theoretical maximum cannot be reached. The default configuration of RAM for vCIC allows 1000 networks. Additional memory is needed if more Neutron Networks are created. For rough estimations, consider that 100 Neutron networks with 1 subnet and 1 port for each cost about 2 GiB memory.
Number of virtual L3 routing instances	The aggregated maximum number of virtual tenant L3 routing instances per CEE Region is 190.
Number of IPv4 routes per CEE region	The maximum aggregated number of IPv4 per CEE region is 16K.
Number of routes per ECMP group	CEE supports up to 32 routes per ECMP group.

Slogan	Limits
Number of vNICs per guest VM	The maximum number of vNICs per guest VM is 10 (+ 1 Trunk vNIC).
Number of Trunk vNIC attached vLANs	The number of Trunk vNIC attached vLANs is limited to 100.
Number of vNICs per blade	CSS supports up to 128 vNICs per Compute host.
L2 Packet MTU	The L2 packet MTU size is 2140 bytes. Setting the L2 packet MTU size to a value larger than 2140 bytes for forwarding is unsupported.
Supported routing types	Static routing is supported. Dynamic routing is not supported.
Number of static routes	The default maximum number of static routes is 1000. It is specified by the `ext_max_routes` value in the `neutron.conf` file. • For setting the value at CEE installation, refer to the *Neutron Configuration Options* in the *Configuration File Guide*. • For setting the value in a running system, refer to the *Static Routes* section in the *Runtime Configuration Guide*.

4.5 Storage

This section describes CEE characteristics on storage.

4.5.1 Limitations When Using Local Storage

For tenants, ephemeral storage (non-persistent block storage) is supported on local disks of the compute hosts. Persistent block storage is supported on centralized storage only.

There is no support for any shared file system in CEE. For distributed storage, see separate sections.

The CEE storage architecture supports persistent block storage through OpenStack Cinder and object storage through OpenStack Swift. Swift uses the Local Storage. Object storage is only used for the CEE infrastructure.

Management of VM images is supported by the OpenStack image service.

The following deployment options are available:

Option 1 Local Storage only

Option 2 Local Storage and Centralized Storage with local Swift Storage

Option 3 Local Storage and Distributed Storage with local Swift Storage

For Local Storage, only ‘‘boot from image’’ is supported.
With additional Centralized Storage, ‘‘boot from image’’ and ‘‘boot from volume’’ are supported.

Table 30 shows where data is stored for both deployment options.

Table 30 Storage Locations

Data	Storage Location
CEE infrastructure backups (incl. Fuel backups)	On local disks of vCIC hosts(1)
Ephemeral storage	On local disks of Compute hosts(1)
vCIC storage (OpenStack infrastructure)	On local disks of vCIC hosts(1)
Core/crash dumps, logs	On local disks of all hosts

(1) Local storage is limited by the local, non-scalable disk capacity.

If data is stored on a local disk, it is erased in case of disk failure or rollback from a failed update, meaning that the VM disappears. The application must be designed accordingly.

4.5.2 VM Migration with NoMigration Policy Set

When a VM has `No Migration` policy set and is booted from local storage, it will not be started again after a rollback since the `/var/lib/nova` partition is not preserved.

The ephemeral disk for the VM is stored on local disk of the compute node. If the compute node must be replaced (because of, for instance, HW failure), any change in the VM is lost.

Boot from Volume is only available when centralized storage is used (Cinder volume stored on centralized storage).

4.5.3 Resiliency

Storage resiliency characteristics are listed in Table 31.

Table 31 Storage Resiliency

Slogan	Characteristics
Centralized storage resiliency	All components of the centralized storage array are redundant.

4.5.4 Centralized Storage Limits

Storage resiliency characteristics are listed in Table 32.

Table 32 Centralized Storage Limits

Slogan	Characteristics
Max number of LUNs (Pool)	The total maximum number of LUNs is 1000.

4.5.5 Distributed Storage, ScaleIO

Distributed Storage Limits are listed in Table 33.

Table 33 Distributed Storage Limits

Slogan	Characteristics
Maximum device size	8 TB
Maximum capacity per SDS	64 TB
Maximum number of devices per SDS	64

4.6 In-Service Performance

This section lists the characteristics on in-service performance.

Table 34 In-Service Performance

Slogan	Characteristics
Guest execution retainability	Guest execution is not interrupted at a virtual Infrastructure management cluster restart or update.

Slogan	Characteristics
Update availability	When the update is running, OpenStack API is unavailable for about a minute for each CIC node. During rollback, negative response is occasionally returned.
Restart availability	It is not possible to connect to the API during the restart. The applications are designed to handle this and will not time out during restart.

5 System Limitations

This section describes the system limitations in R6.

5.1 OpenStack Deviations

The major deviations from the OpenStack SW are:

- Floating IP
- Object Storage
- Live Migration
- Security Groups

See relevant API descriptions for more information about limitations.

Limitations, Listed in API Documents

See the following API documents for more information on limitations:

- *Atlas OVFT API*
- *In Service Performance Northbound API*
- *Fault Management Northbound API*
- *Performance Management Northbound API*
- *Preconfigured Key Performance Indicators*
- *OpenStack API Complete Reference*

5.2 SW Configurations and Options

This section describes SW configurations and options.

5.2.1 Allocation of Memory

CEE supports flavors that allocate VM memory aligned to n × 1 GiB memory when hugepages are enabled. There is a Nova patch that makes Nova aware of this. The hugepages are in chunks of 1 GiB memory, all of which is reserved to the VM even if less memory is asked for.

Each Neutron network created consumes RAM in the vCIC, and it influences the maximum number of virtual tenant networks. See Section 4.4.3 on page 28 for more information.

5.2.2 Allocation of vCPU

The vCPU is limited to even number of vCPUs, see Section 4.1 on page 21.

5.2.3 Collocation of vCIC, vFuel, and Atlas

Only two of the following can be run on the same Compute host: vCIC, vFuel, Atlas.

5.2.4 Number of Parallel Root Volume Operations

Nova in CEE supports about 500 parallel stop/detach root volume operations.

5.3 Not Supported

This section describes functionalities that are not supported. These are included here because they are not related to any specific configurations.

5.3.1 Dashboard Does Not Support Internet Explorer

The Dashboard does not support Internet Explorer because of the Ericsson Graphical User Interface (GUI) Software Development Kit (SDK).

5.4 Limitations and Workarounds

CM-HA operates in active-passive mode. If a Compute host containing a vCIC that runs the active CM-HA restars, the VM evacuation will not start within one minute. The evacuation only starts when the CM-HA process is moved

to another vCIC by Corosync, and the Compute unavailability is detected by the CM-HA.

5.5 Update Limitations

OpenStack API is not always available during update and rollback. When the update is running, the OpenStack API is unavailable for about a minute for each CIC node. During rollback, a negative response is occasionally returned. For more information, see Section 4.6 on page 31.

Reference List

[1] *Juniper Networks homepage, www.juniper.net*

[2] *EMC homepage, www.emc.com*