Cloudera Data Platform on Oracle Private Cloud Appliance and Oracle Compute Cloud@Customer

Joint Reference Architecture, Installation, Patching, and Design

Oracle Private Cloud Appliance (PCA)
Oracle Compute Cloud@Customer (C3)
Cloudera Private Cloud Base (CDP Base 7.1.x)
Cloudera Private Cloud Data Services (CDP DS 1.5.x)

Kevin Talbert, Senior Solutions Engineer, Cloudera
ktalbert@cloudera.com

Scott Ledbetter, Principal Software Engineer, Oracle
scott.ledbetter@oracle.com

21 April 2024

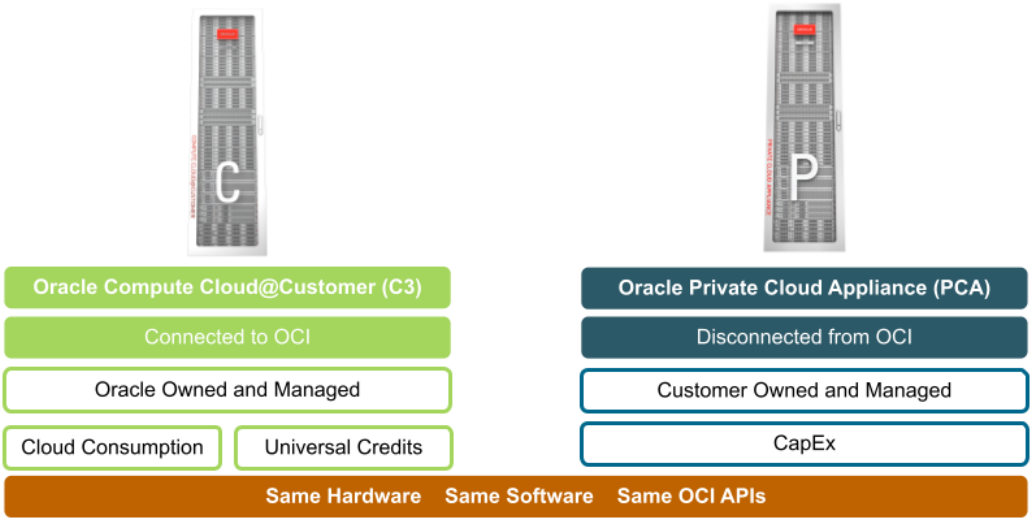
Table of Contents

Executive Summary

Oracle Private Cloud Appliance (PCA) and Oracle Compute Cloud@Customer (C3)

Oracle Private Cloud Appliance (PCA) and Oracle Compute Cloud@Customer (C3) are integrated infrastructure systems engineered to enable rapid deployment of converged compute, network and storage technologies for hosting applications or workloads on a guest OS. PCA and C3 bring infrastructure and architectures that are compatible with Oracle Cloud Infrastructure (OCI) to the enterprise datacenter enabling customers to utilize the same infrastructure, skill sets, tooling, and related services for deployments in both public and private clouds.

Oracle Compute Cloud@Customer combines on-premise hardware with Oracle-provided end-to-end management and support via Oracle Cloud Infrastructure connectivity. Customers pay for C3 as an operational expense rather than capital outlays.



PCA/C3 supports VMs with up to 128 vCPUs and running a wide variety of guest OS, including Linux, Oracle Solaris, and Microsoft Windows. In particular, Cloudera recommends Oracle Linux or Red Hat Enterprise Linux as the guest VM for its services. See the most up to date certification matrix for major/minor versioning, including your choices in Operating System from Cloudera's Support Matrix: https://supportmatrix.cloudera.com/

Oracle ZFS Storage Appliance

The core storage in PCA/C3 is provided by a dedicated Oracle ZFS Storage Appliance (ZFSSA). ZFSSA is an enterprise-grade, unified storage system providing block, file, and object storage to the PCA/C3 infrastructure and guest VM instances. It provides features such as data mirroring, snapshots, copy-on-write clones, and continuous integrity checking to prevent data corruption. Being that ZFS is the only physical storage layer backing CDP's file storage solutions (like HDFS, Ozone, etc), such capabilities create a secondary layer of data resilience.

Cloudera Private Cloud Base

Cloudera Data Platform (CDP) Private Cloud Base lays the foundation of Cloudera's modern, on-premises data and analytics platform by offering faster analytics, improved hardware utilization, and increased storage density. Strengthened platform security and simplified governance for regulatory compliance helps organizations manage enterprise readiness.

CDP Private Cloud Base consists of a variety of components from which you can select any combination of services to create clusters that address your business requirements and workloads. Several pre-configured packages of services are also available for common workloads.

CDP Private Cloud Base includes Cloudera Manager for managing, monitoring, and configuring your clusters and services using its Admin Console web application or the Cloudera Manager API.

Cloudera Private Cloud Data Services

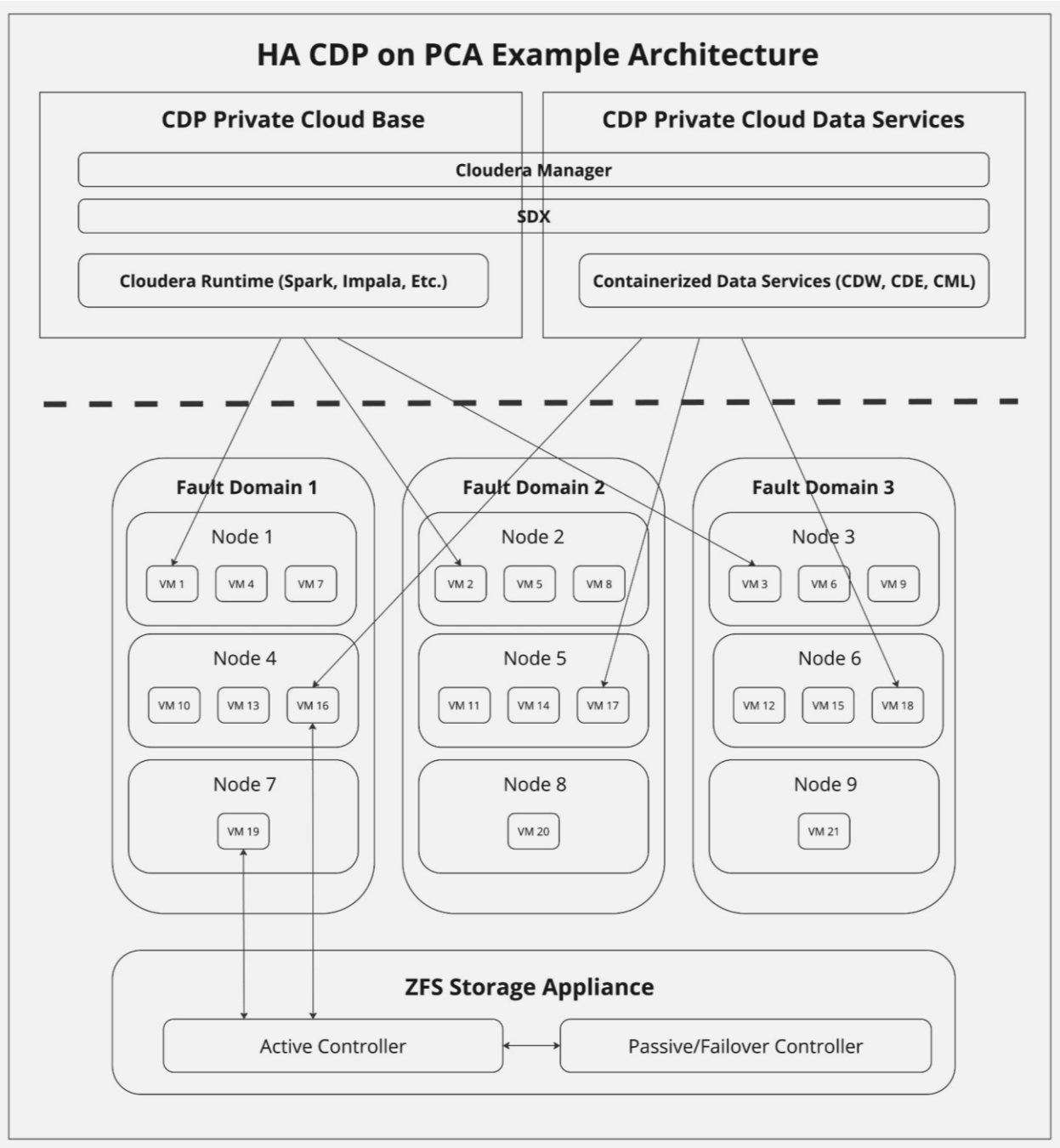
CDP Private Cloud Data Services is an on-premises offering of CDP that brings many of the benefits of the public cloud to your data center. It is the framework on top of CDP Private Cloud Base that lets you deploy and use the collection of Cloudera data services such as Cloudera Data Warehouse (CDW), Cloudera Machine Learning (CML), and Cloudera Data Engineering (CDE). These data services can cater to your data-lifecycle goals.

Installing CDP Private Cloud Data Services requires a compatible version of Cloudera Manager and Cloudera Private Cloud Base which can be found using the compatibility matrix at Cloudera's website: https://supportmatrix.cloudera.com/

Component Architecture

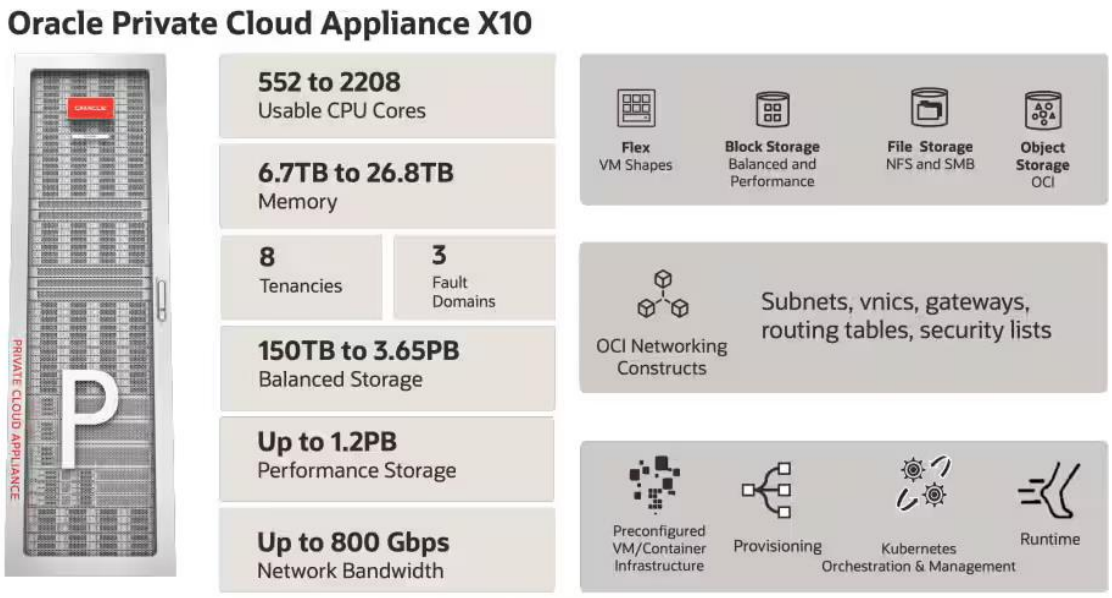
The below diagram illustrates a logical design of the interaction between CDP services and Oracle components. Practically speaking, CDP services are never made 'aware' of the PCA or ZFSSA infrastructure, but the System Architect will need to be aware of these services in

designing the configuration of the VMs on the backend. The theme of PCA is to bring "cloud-like" infrastructure on-prem.



Designing CDP on PCA/C3

Oracle Private Cloud Appliance (PCA) and Oracle Compute Cloud@Customer (C3)



A key differentiator of PCA/C3 over standard racked clusters is that by design, it provides an Oracle Cloud Infrastructure-like deployment for its users, while providing compute and storage resiliency via redundant management infrastructure, switches, and the internal, dedicated Oracle ZFS Storage Appliance (ZFSSA). ZFSSA is an enterprise network attached storage ("NAS") appliance engineered for high availability and performance. Understand more about the role of ZFS in this architecture by reading, "Oracle ZFS Storage Appliance," in the following section.

Migrating Legacy Systems such as Oracle Big Data Appliance (BDA)

Oracle and Cloudera have a long-standing relationship and we have worked closely to develop enterprise class solutions that enable customers to quickly manage Big Data workloads. Oracle Big Data Appliance (BDA) is an engineered system optimized for acquiring, organizing, and loading unstructured data into an Oracle Database. The Oracle Big Data Appliance includes Cloudera Distributed Hadoop (CDH) software which is a distribution of Apache Hadoop and related projects. This combined solution has been successfully deployed and utilized by hundreds of customers worldwide to address Big Data needs. Both Oracle BDA and Cloudera CDH products are reaching end-of-life. The replacement combination for these two products is Cloudera's Private Cloud Data

Platform (CDP) and PCA or C3. Cloudera and Oracle Professional Services offer migration assistance moving legacy BDA customers to CDP Base to PCA and/or C3. The support model for CDP on PCA/OC3 also differs from BDA in that Cloudera will now support the CDP components and Oracle the PCA/OC3 components.

Maintenance and Patching of PCA/C3

Understanding how maintenance and patching effects the cluster design

Minimally, PCA/C3 will require 3 physical compute nodes; however, if the architect wishes to ensure Cloudera services are not shutdown during patching, at least 6 compute nodes should be used, and ideally 9.

PCA/C3 are architected using the same model as Oracle Cloud Infrastructure (OCI). Each PCA/C3 is an Availability Domain (AD) in OCI terms. OCI uses a concept called Fault Domains (FD) to allow a form of anti-affinity. Each PCA/C3 has three FDs. There must be three Fault Domains and it is assumed that each FD contains the same number of Compute Nodes (CNs), with a minimum of one CN per FD. A Compute Node is a physical server that runs the PCA/C3 hypervisor and which hosts Compute Instances (VMs).

There is no way to control distribution of Instances across Compute Instances within a FD. PCA/C3 software determines which Compute Node is best suited to run a new Instance. However, Instances assigned to different FDs are guaranteed not to be running on the same physical Compute Node.

In a high availability environment, it is important to distribute the Cloudera nodes in such a way that if one compute node should fail, the cluster remains running. Ensure that nodes and their associated Cloudera roles are not on the same PCA/C3 compute node by assigning them to different FDs. For example, when using two CDP Manager nodes, ensure that each CDP Manager node resides on a different FD. Other critical redundant functions such as HDFS Name Nodes , and journaling nodes should also be distributed across the three FDs. Worker VMs too may be evenly distributed on different physical nodes (eg with anti-affinity rules) to avoid a significant loss of processing capacity in case of single node maintenance.

This theme is in alignment with the principles of High Availability (HA). The manner in which the VMs are then sized and divided over this physical structure will then come down to the targeted services and maintenance needs. For example, to patch PCA compute nodes without shutting down Cloudera services, unutilized overhead should be given to each fault domain where the amount of space is in alignment with the ability to cycle VMs intra-fault domain, instead of between fault domains, where proper segregation of Cloudera roles may be compromised.

Designing Highly Available PCA systems where CPU/Memory resources are minimal

Shutting down VMs and the services on them should be done in a way where no more than ONE (1) of a highly available service is allowed to be shut down. VMs (and therefore nodes) with non-Highly available services should *not* be a candidate for being the first node to be shut down during a patch/maintenance event. Read "A Special Note About Live Migrations During Patching" below to understand in detail why we design this way.

Oracle ZFS Storage Appliance

Oracle's ZFS (https://docs.oracle.com/en/storage/index.html) uses a combination of standard enterprise-grade hardware and a storage-optimized operating system based on the Oracle Solaris kernel to provide low latency compute power. PCA contains a dedicated, dual-controller ZFSSA X9-2 with a customer-configured mix of up to 3.65 petabytes of high-capacity storage and/or up to 1.1 petabytes of high performance all-flash storage.

Like all data center components, the PCA internal ZFS Storage Appliance requires occasional firmware upgrades. Because all volumes of all VMs in PCA are served by this Storage Appliance, and from a single ZFSSA Storage-Pool" running on a single ZFSSA controller at a time, special care is necessary during upgrade cycles to allow the internal ZFS Storage Appliance to perform takeover and failback operations, which in some cases may halt I/O for 60 seconds or more. It may be necessary to manage active Cloudera workloads during these operations. It is recommended to shut down Cloudera services during this time.

Maintenance and Patching of ZFS

Understanding how maintenance of ZFS affects Cloudera services

A dual-controller system is the recommended approach for ZFS. It is also recommended to shut down Cloudera services prior to PCA/ZFS patching. This, coupled with proper saving of the HDFS namespace prior to shutdown, does not have to imply a long outage. However, if the organization desires to leave services alive during patching, they should ensure the time is less than 1 minute for ZFS failover. It is HIGHLY recommended to stop user operations with the storage layer. This may include but is not limited to, suspending all jobs that require interfacing with the ZFS storage layer, and changing queues to not temporarily not accept jobs/queries. A great way to lower the outage time is to perform an HDFS checkpoint (in normal operation this is performed once an hour) just before taking the HDFS service offline. This will ensure a faster time to live when the service is brought back online.

Oracle recommends enough extra compute nodes be installed in each Fault domain such that there is enough spare capacity to hold one entire compute node worth of VM instances. In other words, if a customer was going to run three compute nodes per Fault

domain, if they have to have each one close to being fully utilized and there is nowhere to migrate instances when one compute node needs to be evacuated, the recommendation would be for four compute nodes per fault domain since unbalanced configurations are not allowed (e.g. each fault domain must have the same number of compute nodes).

Lowering YARN pools to a low percentage (e.g. 5-10%) can lower IO operations to ZFS and lower the total time of a ZFS Failover. This will allow jobs to still continue graceful queueing while lowering the total volume of IO running through ZFS from CDP. Kafka is an outlier for this procedure as it would sit outside of YARN but the remaining Cloudera services would respond to this change.

PCA Compute Instances (VMs) and Operating System

The foundation of PCA is built on a highly customized and optimized hypervisor that is controlled using a browser, the Oracle Cloud Infrastructure CLI, a custom infrastructure administration API, or a REST API.

Oracle Linux is Oracle's own distribution of Linux for enterprise and cloud-native workloads. It is available as one of different guest OS options in the PCA software stack. Cloudera software can be installed on top of Oracle Linux or Red Hat Enterprise Linux in accordance with the Cloudera OS support matrix: https://supportmatrix.cloudera.com/

Cloudera Data Platform Private Cloud Base

This is the core of the Cloudera Platform. It represents both the compute and storage layer of the Cloudera ecosystem and is functionally managed by Cloudera Manager. The orchestration between all of the Cloudera services and their interconnectivity is carried out by Cloudera's Management Service which includes both the CM Agent, running on each of the compute/data nodes, and the CM Server, running the Cloudera Manager service. Installation of CDP follows the standard process documented at https://docs.cloudera.com/cdp-private-cloud-base/7.1.9/installation/topics/cdpdc-installation.html with the important distinctions noted in this document.

One of the core tenants of the Cloudera Data Platform is resiliency. The notion of high availability is baked into most of the services within the platform and should be employed in any design built on PCA. Building a fault tolerant CDP Base installation requires consideration for how services are to be positioned with logical separation considered at the service levels. For example, an HDFS Journal node should be distributed across three fault domains; those who are familiar with rack awareness should understand this concept quite well.

In large part, the considerations for Private Cloud on PCA are derived not from the installation but rather the ongoing maintenance of the system. Actually installing the

services is quite standard; the process is reminiscent to any virtual machine-based installation and can be conducted accordingly.

The five main considerations are:

1) Replication Factor
2) Distributed placement of the Cloudera services to ensure high availability (HA)
3) VM sizing
4) Rack design / Oracle Fault domains
5) Block Volume / Block Size for IO traffic

Properly addressing these three considerations will ensure a smooth patch cycle and maintenance of the Oracle system components.

1. Replication Factor

Cloudera recommends using an HDFS replication factor of 2x instead of the default 3x. This is because combining the replication factor of 2x induced by the ZFS controllers ultimately leads to 2x2 = 4x total replication.

2. Distributing the Cloudera Services

There are four key node types which should be understood when positioning services as part of the CDP Base deployment. They are Master Hosts, Utility Hosts, Gateway Hosts, and Worker Hosts. In the case of worker hosts without the need for high availability, the Utility and Gateway can be combined. Review the design patterns below to determine a resilient distribution pattern for the Cloudera runtime components you plan to deploy. In the context of a CDP deployment on PCA, the term “host” refer to a VM.

*Worker Hosts **without** High Availability*

Note that designs which follow this architecture **must be shut down during maintenance and patching windows**. Even taking a VM-by-VM or node-by-node approach will not work for this architecture. For example, if the NameNode service is down on Master Host 1, the cluster is effectively useless since the HDFS service is considered dead. CDP installations without high availability are generally only used for non-production and/or non-business critical workloads.

*Worker Hosts **with** High Availability*

This is always the **recommended** approach to deploying Cloudera runtime components. For the most part, the services running in this design can sustain isolated outages. For example, a Master Host VM is suspended during a platinum patch window that is migrated across a fault domain. Assuming this is the only Master Host which has been suspended, the business impact is actually only performance related. The same is true for services like Hue where a load balancer sits in front of the Hue Servers and can route traffic from a

'dead' node and to a 'live' Hue Server; however, this depends on more than one Hue Server having been configured. The following considers 3 fault domains in use by the Private Cloud Appliance. Fault domains are a crucial part of the Highly Available setup in that they provide a similar concept to rack affinity.

Master Hosts	Utility Hosts	Gateway Hosts	Worker Hosts
Master Host 1, Fault Domain 1: • NameNode • JournalNode • FailoverController • YARN ResourceManager • ZooKeeper • JobHistory Server • Kudu master • HBase master • Schema Registry **Master Host 2, Fault Domain 2:** • NameNode • JournalNode • FailoverController • YARN ResourceManager • ZooKeeper • Kudu master • HBase master • Schema Registry **Master Host 3, Fault Domain 3:** • Kudu master (Kudu requires an odd number of masters for HA.) • Spark History Server • JournalNode (requires dedicated disk) • ZooKeeper	**Utility Host 1, Fault Domain 1:** • Cloudera Manager • Cloudera Manager Management Service • Cruise Control • Hive Metastore • Impala Catalog Server • Impala StateStore • Oozie • Ranger Admin, Tagsync, Usersync servers • Atlas server • Solr server (CDP-INFRA-SOLR instance to support Atlas) • Streams Messaging Manager • Streams Replication Manager Service **Utility Host 2, Fault Domain 2:** • Hive Metastore • Ranger Admin server • Atlas server • Solr server (CDP-INFRA-SOLR instance to support Atlas)	**One or more Gateway Hosts, Distributed equally across Fault Domains 1-3:** • Hue • HiveServer2 • Gateway configuration	**3 - 20 Worker Hosts, Distributed equally across Fault Domains 1-3:** • DataNode • NodeManager • Impalad • Kudu tablet server • Kafka Broker (Recommend 3 brokers minimum) • Kafka Connect • HBase RegionServer • Solr server (For Cloudera Search, recommend 3 servers minimum) • Streams Replication Manager Driver

A Special Note About Live Migrations During Patching

High availability and service layout design is a crucial consideration for patching and VM live migrations. In the case where VMs are live migrated off of PCA /C3 Compute Nodes(CNs), the VMs must have a place to go to.

Evacuating a Compute Node for patching is accomplished by first setting a provisioning lock on the Compute Node to prevent any new resource allocation on that CN. Then, a "`migratevm`" command is issued against the Compute Node, which initiates a process to live migrate all VMs from that Compute Node.

In live migrations, the PCA /C3 will attempt to move the VM to a CN in the same Fault Domain (FD); however, in the event the Fault Domain does not have space to fit the VM(s), PCA can migrate the VM(s) across Fault Domains. The StrictFD setting in PCA/C3 determines whether live migrations across Fault Domains will be allowed. If StrictFD is enabled during a `migratevm` operation, a VM will *not* migrate to a different Fault Domain during live migration. If there is no room in the current Fault Domain for the VM, it will remain running on its current Compute Node, and the `migratevm` command will be marked as failed. Other VMs that could fit in the current FD may have migrated, but those that cannot fit will remain running on their current CN. If StrictFD is not enabled during the `migratevm` command activity, the PCA/C3 will first try to fit the VM into the current FD, but if there are not enough resources to fit the VM, it will be moved to a different FD. If there are not enough resources in another FD, the VM will remain running on its current Compute Node and the `migratevm` command will be marked as failed.

By recommendation, during PCA patching events we recommend using StrictFD set to *Enabled*. This ensures that the live migration will gracefully fail if the VMs being relocated cannot be moved within the same FD. Remember, moving a VM across Fault domains can cause service disruption since the VM is seen as offline to CDP. Also, in the PCA x9 and , ***the destination FD for VMs selected for evacuation can not be specified.*** This means moving a VM across Fault Domains may create circumstances where the CDP services residing on a VM become "doubled" up for a particular Compute Node. This is particularly concerning if say two VMs, both housing Journal nodes (3 Journal nodes total in the cluster) become co-located on a single server, and the server is then shutdown. Quorum (> 50% availability) is essential for services to remain healthy. Because of this, before the patching event, it is important to map the VMs to the Compute Nodes they reside on. This process can be done several ways, most notably using the OCI Designer Toolkit (https://github.com/oracle/oci-designer-toolkit), a health check tool which can be obtained from Oracle support, or the PCA-ADMIN command "`getrunninginstances`". If for some reason the *only* way a VM can be evacuated is across Fault domains (usually seen in heavily utilized environments), this process is paramount. In the case of the Journal node example, without this logical mapping, shifting two Journal nodes (albeit residing on one PCA compute node), across Fault domains, could cause HDFS to become unhealthy

and lose quorum. The only way to ensure two VMs do not run on the same compute node is to put them in separate fault domains.

For more information about the StrictFD setting and other PCA/C3 compute service settings, refer to the "Compute Service Configuration Commands" section of the PCA/C3 Administrator Guide.

Larger Clusters with High Availability

Follow the documentation set forth by Cloudera for the proportional size cluster you intend to build. For example, larger clusters with node counts in excess of 20 nodes should consider distributing services in a more dedicated manner as opposed to running many services on the same utility or worker nodes. Refer to the documentation: https://docs.cloudera.com/cdp-private-cloud-base/7.1.9/installation/topics/cdpdc-runtime-cluster-hosts-role-assignments.html

3. Sizing Virtual Machines

While the general practices of sizing Oracle VMs within PCA is broadly the same as any like-for-like VMWare deployment, the key differentiator here is architecting for resilience during patching. Oracle's patching process involves evicting all VMs from a particular node (an order which can and *should* be selected by the organization based on the priority of the services running) and then patching the node. These VMs need a place to go within the fault domain so if the VMs are sized too large (e.g. 96vCPU of an available 128vCPUs per node) they will not have a place to go within the same fault domain. Note, if a VM shifts across a fault domain, it will be suspended and Cloudera services will likely consider the services on that VM temporarily dead. This can, of course, be avoided by N+1 setup within a fault domain during patching, but architecting around this may just be as simple as lowering the total size of the VM and raising the number of VMs. This, of course, is also a balancing act as Cloudera runtime components generally prefer fewer, beefier compute nodes.

The sizing of the number of OCPUs for each Cloudera node will vary depending on the exact CDP roles installed. The Cloudera Manager dashboard can help fine tune the CPU and RAM allocations, as it has dashboard elements showing overall CPU utilization for a cluster, CPU utilization for individual nodes, and RAM utilization. PCA/C3 have predefined Instance shapes, as well as Flex Instance shapes, which allow the administrator to fine tune the number of OCPUs and the amount of RAM given to a Compute Instance.

It is important to understand that in PCA/C3, the number of OCPUs assigned to a VM Instance also determine the maximum network bandwidth that the Instance will be allowed to consume. An Instance is network-throttled to be limited to a maximum of 24Gbps of bandwidth if it is assigned 24 or fewer OCPUs. This is the aggregate bandwidth across all vNICs assigned to the Instance. Instances with more than 24 OCPUs are

allowed more network bandwidth. If it appears that network throttling is impacting Cloudera performance, increasing the number of OCPUs assigned to node Instances may be necessary even if CPU utilization is not near the limit. Refer to the PCA/C3 Concepts Guide, in the section entitled "Compute Shapes", for charts showing maximum network bandwidth for various Compute Shapes. https://docs.oracle.com/en/engineered-systems/private-cloud-appliance/3.0-latest/concept/EN-PCA-3-0-LATEST-CONCEPT.pdf

Cloudera VM instances do not have to be the exclusive workloads on the PCA/C3 Compute Nodes, but the user must maintain awareness of **all** workloads running in the PCA/C3 when resolving performance issues. The Cloudera performance dashboards will not see non-Cloudera workloads running on shared PCA/C3 Compute Nodes, storage, or networks . On PCA, Grafana dashboards can be used to help see the overall resource utilization on a PCA. On C3, dashboards in the OCI control plane will show resource utilization. Grouping all CDP instances for a cluster into one PCA/C3 compartment can help in managing and measuring the infrastructure resources.

For RAM allocation to Cloudera nodes, use Cloudera sizing recommendations as a starting point. The Cloudera Manager dashboards can be used to see how much RAM is actually being used by a node, and adjustments can be made accordingly. A Compute Instance running on PCA/C3 can be defined with Flex shapes, allowing complete control over the number of OCPUs and the amount of RAM allocated to each Instance, but an Instance must be in a stopped state to change RAM or OCPUs.

4. Fault Domains / Rack Awareness

The Oracle Private Cloud Appliance aligns with Oracle Cloud Infrastructure (OCI) architecture which includes the concept of an Availability Domain (AD). Every Oracle Private Cloud Appliance is an Availability Domain (AD), which consists of one or more server racks, independently managed from other PCA systems. A Fault Domain is a grouping of infrastructure components within an Availability Domain. A PCA, and therefore an Availability Domain, contains three Fault Domains, each consisting of one or more physical compute nodes. Compute nodes are evenly distributed over the three Fault Domains. Fault Domains are used to isolate downtime events due to failures or maintenance, and to ensure that resources in other Fault Domains are not affected. Fault domains provide anti-affinity, which lets you distribute your instances so that they are not on the same physical hardware. A hardware failure or compute hardware maintenance event that affects one fault domain does not affect instances in other fault domains.

Bringing this concept into the "Cloudera-world" involves dispersing the same "highly available" services across fault domains. For example, an HDFS Journal node should have at least one instance per fault domain to ensure this principle is applied. An example is provided in a previous section titled "Worker Hosts with High Availability".

5. Block Volumes / Block Size for IO Traffic

PCA/C3 block storage is provisioned from the internal ZFSSA. The default blocksize of a PCA/C3 block volume is 8k, and it has been shown that 128k is a better block size for Cloudera. There are also cache parameters that can help with throughput. The block size and cache parameters are specified at the time the block volume is created.

Changing the blocksize to a larger value than the default of 8K is desirable for HDFS volumes due to the sequential characteristics of the IO written to and read from HDFS volumes. A larger blocksize reduces the number of IOPs necessary to transfer large sequential streams from and to the HDFS files.

When creating block volumes for Cloudera, volumes containing databases should use default creation parameters, but volumes dedicated to HDFS should be created with a block size of 128K, and the "logbias" setting should be set to "Throughput". These parameters are passed using Tags at creation time. Refer to the Block Volume section of the PCA/C3 User Guide for more information.

Here is an example of a OCI CLI command to create a block volume for HDFS on PCA/C3:

```
oci bv volume create --display-name cloudera-hdd-1 --size-in-gbs 500 -
-vpus-per-gb 10 --compartment-id
ocid1.compartment.AKnnnnn.b52ok4v7jicfdol5ta6k4iujt6jsxcx6p0kvlp720bzk
i0uywal5r0sr38l2 --availability-domain AD-1 --freeform-tags
'{"PCA_blocksize":"131072"}' --defined-tags
'{"OraclePCA":{"logBias":"THROUGHPUT"}}'
```

Cloudera Data Platform Private Cloud Data Services

Cloudera's Private Cloud Data Services platform represents a unique departure from its usual VM-like deployment strategy, allowing organizations to build highly scalable, container-based compute platforms on the Open Shift Container Platform (OCP) or Rancher Kubernetes' Engine dubbed the Elastic Container Service (ECS). The service contains three experiences: Cloudera Data Warehouse, Cloudera Data Engineering, and Cloudera Machine Learning. These three services are facilitated with an installation via Cloudera Manager, and management via the installed Data Services Control Plane. In the case of the Data Services installation being considered here, we will describe an experience orchestrated with the ECS flavor of the product.

Installation Prerequisites

As of the writing of this, the software support matrix for Cloudera's Private Cloud Data Services can be found at the following: https://docs.cloudera.com/cdp-private-cloud-data-services/1.5.4/installation-ecs/topics/cdppvc-installation-software-support-matrix-base-and-ds-ecs.html

This also contains the detailed breakdown of the software, hardware, and services requirements necessary to successfully install a CDP Data Services cluster.

Starting with CDP Private Cloud Data Services 1.5.4 release, Oracle Linux (RHCK Kernel only) 8.7, 8.8, 8.9, 9.1, 9.2, and 9.3 are certified to run Data Services; however, using Red Hat Enterprise Linux (RHEL) 8.8 is certified currently and continues to be the first class guest OS recommended by Cloudera. Typically the decision on which OS to choose is driven by the business needs and security team.

CDP Data Services 1.5.4 Release Notes: https://docs.cloudera.com/cdp-private-cloud-data-services/1.5.4/release-notes/cdppvc-release-notes.pdf

Hardware Considerations

To run Cloudera's Data Services on Oracle PCA, the same general principles of installing Data Services all apply. This means a minimal Data Services cluster requires 1 control plane node (VMs) for non-HA, and 3 control plane nodes (VMs) for HA.

For services which require high performance such as those listed in Cloudera's Public Documentation as "NVMe", this corresponds to Oracle's High Performance Disk Tray. These will functionally achieve the same low latency performance needed by the Data Services for caching.

The following table enumerates these requirements explicitly, and a node listed in the table can be provisioned as VM within a compute node with recommended configuration on PCA:

Note the term "nodes" and "VMs" may be used interchangeably here since these sizing recommendations apply to the way Cloudera services 'see' infrastructure and not PCA.

COMPONENT	MINIMUM	RECOMMENDED
NODE COUNT	1 (Non-HA)	3 (HA)
CPU	16 cores	32 cores (per node)
MEMORY	32 GB	64 GB (per node)
STORAGE	300 GB	1 TB (per node)
NETWORK BANDWIDTH	1GB/s to all nodes and base cluster	1GB/s to all nodes and base cluster

The three services, Cloudera Data Warehouse, Cloudera Data Engineering, and Cloudera Machine Learning, all have their own minimal requirements as well:

1. *Cloudera Data Warehouse*

COMPONENT	MINIMUM	RECOMMENDED
NODE COUNT	4	10
CPU PER WORKER	16 cores	32+ cores
MEMORY PER WORKER	128 GB per node	384 GB per node
STORAGE	1.2 TB SAS3, SSD per host	1.2 TB/SSD per host*
NETWORK BANDWIDTH	1 GB/s guaranteed bandwidth to every CDP Private Cloud Base node	10 GB/s guaranteed bandwidth to every CDP Private Cloud Base node

** This maps to Oracle's High Performance Disk Trays to achieve the same speeds desired by a worker on CDW.*

2. *Cloudera Data Engineering*

COMPONENT	MINIMUM	RECOMMENDED
NODE COUNT	2	4
CPU	16 cores for CDE workspace (base and virtual cluster) and 8 cores for workload	16 cores for CDE workspace (base and virtual cluster) and 32 cores (you can extend this depending upon the workload size)
MEMORY	64 GB for CDE workspace (base and virtual cluster) and 32 GB (you can extend this depending upon the workload size)	64 GB for CDE workspace (base and virtual cluster) and 64 GB (you can extend this depending upon the workload size)
STORAGE	200 GB blob storage and 500 GB NFS storage	200 GB blob storage and 500 GB NFS storage
NETWORK BANDWIDTH	1 GB/s to all nodes and base cluster	10 GB/s to all nodes and base cluster

3. *Cloudera Machine Learning*

COMPONENT	MINIMUM	RECOMMENDED
NODE COUNT	1	1 per workspace + additional nodes depending on expected user workloads

CPU	32 Cores Per Workspace+ additional Cores depending on expected user workloads	32 Cores Per workspace + additional Cores depending on expected user workloads
MEMORY	128 GB minimum + additional for heavier workloads	256 GB per workspace + additional for heavier workloads
STORAGE	600 GB Block storage + 1000 GB NFS storage (Block if internal and NFS if external/PCA) PCA supports NFS organically.	4500 GB Block storage + External/PCA NFS with 1000 GB NFS storage minimum + additional storage based on sizing of project files
NETWORK BANDWIDTH	1GB/s to all nodes and base cluster	1GB/s to all nodes and base cluster

In most cases these VMs will sit coresiding with the CDP Base cluster from a PCA perspective, however ***CDP Data Services should be turned off if any of the VMs its running on are being considered for evacuation across Fault domains***.

Example PCA Deployment of CDP Data Services

An example 6 node PCA installation with 3 fault domains, 11 used VMs, and 7 empty VMs (or consumed by other non-CDP services), may therefore look something like:

PCA Fault Domain		*VM*	*VM*	*VM*
FD 1	*pca_cn_1*	**ECS Master 1 (Control Plane)**		ECS Compute
	pca_cn_4 …	ECS Compute	ECS Compute	
FD 2	*pca_cn_2*		**ECS Master 2 (Control Plane)**	ECS Compute
	pca_cn_5 …	ECS Compute		
FD 3	*pca_cn_3*		ECS Compute	**ECS Master 3 (Control Plane)**
	pca_cn_6 …	ECS Compute	ECS Compute	

An Important Note on GPUs

Oracle PCA does not currently support GPUs at the time of this document writing. This is an important consideration for many modern generative AI applications.

Installing Data Services

For installation, use the guide "Installation using the Embedded Container Service (ECS)" available at Cloudera's documentation website: https://docs.cloudera.com/cdp-private-cloud-data-services/1.5.3/installation-ecs/topics/cdppvc-ecs-install.html

Maintenance and Patching of Oracle Systems running Data Services

It is recommended to shut down Cloudera's Data Services platform before beginning any maintenance activities on Oracle's systems. If the organization chooses to leave these services online during a patching event, it is critical that the node running a Data Services VM not go into a suspended or shutdown state as the Docker engine may not gracefully recover from these states. This means the same notion of fault domains should be considered here as well with respect to sizing the VMs housing Data Services. Perhaps even more crucially here is the balancing act of sizing large enough VMs for the Data Services without going too large that there are no available nodes within a fault domain to house VMs being evicted.

For the same reason recommended to use `STRICT_FD` in CDP Base patch events, the same is true for CDP Data Services. `STRICT_FD` is essential given the VMs cannot be told which compute node to evacuate to. This guarantees not only high availability for the CDP Data Services but also for the fact we do not want any ECS Masters co-located on the same PCA compute node.

Additional Considerations for Sizing VMs running Data Services

While Private Cloud Base nodes require very minimal overhead for essential CDP services, Private Cloud Data Services nodes require a much larger safety value of 8 vCPU and 8gb Memory (RAM). This means that when sizing the Data Services nodes, it's especially important to consider a larger VM size over the quantity of VMs, to maximize the available CPU and RAM to the business users at a given time.

Enabling Data Services High Availability within the ECS Control Plane

Most Data Services deployments will be Highly Available post installation. It is important to follow the documentation Cloudera provides on setting up ECS high availability as this is not the default when installing Data Services. Check out the following documentation and

its sub-trees: https://docs.cloudera.com/cdp-private-cloud-data-services/1.5.3/installation-ecs/topics/cdppvc-install-ecsha.html

It is recommended to have an odd number of members in a cluster. An odd-size cluster tolerates the same number of failures as an even-size cluster but with fewer nodes. The difference can be seen by comparing even and odd sized clusters:

Cluster Size	Majority	Failure Tolerance
1	1	0
2	2	0
3	2	1
4	3	1
5	3	2
6	4	2
7	4	3
8	5	3
9	5	4

Adding a member to bring the size of cluster up to an even number doesn't buy additional fault tolerance. Likewise, during a network partition, an odd number of members guarantees that there will always be a majority partition that can continue to operate and be the source of truth when the partition ends.

Conclusion

Cloudera Data Platform running on Oracle Private Cloud Appliance or Oracle Compute Cloud@Customer provide a compelling software and hardware stack, delivering faster and easier data management and data analytics, with optimal performance, scalability, and security. Careful pre-installation planning will ensure a successful and satisfying solution to large data challenges.