zadara

Zadara Cloud Services - Introduction

Release 24.03

CONTENTS

Zadara Cloud Services is a full stack of cloud services: Compute, Networking and Storage. It is protected by Zadara security controls and provisioned as managed service.

Zadara cloud services gives any MSP the ability to have an AWS-compatible on-premise private or public cloud within its own data center. It gives enterprise customers the ability to consume above services in the exact same way they consume the public cloud using AWS-compatible API.

Zadara cloud services are developed, deployed and maintained by Zadara on behalf of its partners and customers.

Zadara cloud services always run on Zadara owned equipment (either in public or on premise deployments), and delivered as a service with a "per consumption" business model.

At a high level:

- Zadara compute cluster is added to Zadara storage cloud to provide the full range of Zadara Cloud Services.

- Zadara Software that runs on the above cloud configures the servers to be either compute or storage nodes, thus providing the experience of a private cloud region.

The Zadara Cloud Service includes the following:

- A **fully-featured compute layer** that allows for creation, management, and manipulation of x86-based virtual machines. Zadara provides intelligent load balancing. If it finds an overloaded node, it automatically moves workloads away from that node to a node with more capacity - typically moving a VM from source to destination in about half a second.

- The ability to use **any type of storage** - Block, File and Object Storage, all provided by Zadara VPSA.

- A **complete virtual networking layer** allowing for the creation of objects ranging from internal private networks, external NAT communication networks, fixed IPs for specific VMs, and security groups to control what may flow across those networking objects and paths.

- A **fully distributed automation and control architecture** that exists across the cloud. Management processes and state are replicated to every node in the region, This make the usage of management resources much more efficient, and much more reliable in case of a planned or unplanned outages.

- An **API** and **CLI** for scripting, automation, and control.

- A fully-featured, browser-based **GUI** to handle all day-to-day operations without having to drop to the command line.

CONTENTS

ZADARA CLOUD SERVICES CATALOG

1.1 Cloud Services Administration

Each customer that orders Cloud Services from the Zadara Provisioning Portal, gets an account on a specific cloud in the requested location. The account contains one Project by default, but additional projects can be created. A Project is similar to an AWS project or VMware folder, and is basically a container of resources (VPCs, VMs, ELB, etc.). Definition of different projects allows separation of resources to different application groups or environments. The Tenant Admin can assign roles and policies to users per project.

1.2 Users and Roles

On Zadara IaaS platform, users are defined with one of the following roles:

1.2.1 Account Member

This role allows the user to use the console, policies and APIs for creating, viewing, modifying and deleting virtual resources (e.g. VMs, volumes, etc.) belonging to projects to which the user has been assigned. This is the standard role for most users.

1.2.2 Tenant Admin (Account Admin)

In addition to allowing the use all functions which are granted to a Member, the Tenant Admin role also allows the user to use policies and APIs for creating and managing new projects and users within a specific account. The user that is registered on the Zadara Provisioning Portal for Compute services, has the role of Tenant Admin and is responsible to manage the account on behalf of his organization. The Tenant Admin can create additional users with this role.

The following services are available for regular users (account members):

1.3 Compute
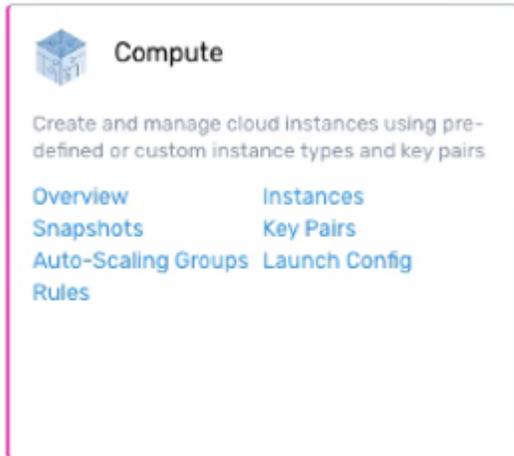
1.3.1 Compute Instances (VMs)

Virtual Machines of any type (built-in or user defined) can be created based on OS image with the use of a cloud-optimized hypervisor (Enhanced KVM). This service has AWS EC2 compatible API that offers standard AWS API for application integration.

1.3.2 Snapshots

A snapshot is a copy of the volumes of a virtual machine at a given point in time. Snapshots provide a change log for the volume and are used to restore a VM to a particular point in time when a failure or system error occurs. You can create snapshots of the VMs that are in the system, and use these snapshots to restore any VM if needed.

1.3.3 Key Pairs (security keys settings)

Public–key cryptography is used to encrypt and decrypt login information for Linux instances. (Instances created from a Windows image are accessed by a password instead.) A public key is used to encrypt a piece of data, and then the recipient uses the private key to decrypt the data. These two keys are known as a key pair. Key pairs enable you to securely access your instances using a private key instead of a password.

1.3.4 Auto Scaling

The system monitors applications and automatically adjusts allocated resources to meet evolving requirements. It simplifies scaling operations to a minimal effort based on a simple and powerful interface for planning and defining launch configurations and scaling policies. You define the policy, and the system ensures performance and availability while keeping resource utilization optimized to reduce costs.

1.4 Networking

1.4.1 VPC

The Virtual Public Cloud (VPC) is your own private data center within the cloud infrastructure. You get to select the network addresses that you will use throughout your infrastructure. Since this is your network, you can decide to slice it up any way you prefer. The VPC is a networking resource with a logical router at its core. When you create a VPC, you specify a CIDR block. All subnets that you will create in the VPC will be carved out from this CIDR block, and the router will ensure IP connectivity between them.

1.4.2 Subnets

Subnets are used for configuration of networking within a VPC. As compared to using one big network, use of subnets in your VPC allows segregation between private and public facing networks, definition of availability zones, and other similar advantages.

1.4.3 Network Interfaces

A **virtual network interface** (VIF) is an abstract virtualized representation of a computer network interface. Each virtual machine (VM) may have one or more virtual network interfaces that act as virtual NICs.

1.4.4 Route Tables

The basic tenet of networking is that everything inside your subnet stays inside your subnet – and if you want to go outside of your subnet, you need to go through the default gateway through which traffic is routed to an external destination network. The route tables are associated with each subnet to allow the flow of traffic according to the VPC policies and configured options.

1.4.5 Internet Gateways (IGW)

Your connection to the outside world is the Internet Gateway. Having instances running in a cloud is great, but if you cannot get to them from the outside world they might be useless. Without an IGW, manageability would be very challenging – if not completely impossible.

1.4.6 Elastic IP (EIP)

Elastic IPs are used to expose instances outside of the Zadara compute cloud. An EIP will be used in the network address operation (NAT) of all traffic to/from the virtual network interface with which it is associated.

1.4.7 VPC Peering (in the same cluster)

VPC peering lets you create direct IP connectivity between any two VPCs. Direct connectivity between VPCs means that servers in a VPC can be reached from the other VPC without the need for elastic IPs or traffic flowing through external routing. VPC peering is simply L3 connectivity realized using routing tables and IP connectivity.

1.4.8 NAT Gateways (NGW)

There will be cases where you don't want instances to be exposed to the outside world and you don't want them to have public IP addresses. However, these instances may still need access to the outside world to get updates or to send information. By using a NGW, you can allow outbound access to the internet while limiting inbound access, thus providing an additional layer of abstraction and protection for your workloads.

1.4.9 DNS

The DNS Service provides an easy-to-manage DNS solution for your private network. It offers high availability and a cost-effective way to connect applications and services and make them available to users. The DNS service is seamlessly integrated with the resources and services of the cloud. You can map domain names to your compute instances and other services.

1.4.10 Security Groups

Security groups are essentially whitelists applied to the virtual network interfaces to control the inbound and outbound traffic. Traffic that does not match any rule in the security group will be discarded.

For each security group, you include one set of rules that controls the inbound traffic to the instances, and a separate set of rules that controls the outbound traffic from the instances.

1.5 Storage
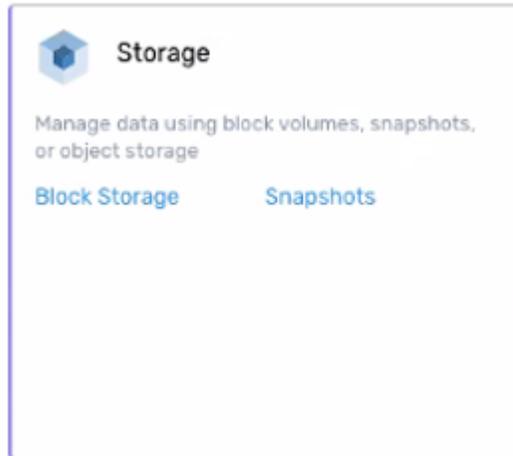
1.5.1 Block Storage (EBS)

EBS (Elastic Block Store) is native block storage, designed to be extremely scalable, agile and flexible, while being fully compatible with the requirements of cloud computing. The Block Storage service offers comprehensive block storage capabilities, enabling you to reduce latency and overhead and to increase throughput. The service is based on Zadara VPSA block services and fully compatible with AWS EBS API for application integration.

Different capabilities are available, based on Volume types.

1.5.2 Volume types

From version 23.08, zCompute simplifies storage management with volume types that provide users with a range of options to meet their storage requirements, balancing factors such as performance, cost, and specific workload demands.

✔ **Note:** The cloud administrator needs to explicitly activate the volume types capability for specific edge clouds.

To check if this capability is configured, go to **Storage > Block Storage** and see whether the **Volume Type** column is displayed instead of the **Storage Pool** column.

By abstracting the underlying storage infrastructure, volume types simplify storage management and allow users to focus on selecting the appropriate type for their zCompute needs.

Different volume types offer varying levels of input/output operations per second (IOPS) and throughput, which determine the storage performance. Higher performance volumes are typically associated with faster data transfer rates and lower latency, enabling applications to handle demanding workloads or perform intensive operations.

1.5.3 File Shares

NAS service is available to provide scalable files shares via NFS or SMB protocols. The service is based upon the Zadara VPSA NAS offering that allows file access to all associated VM instances. (This service is currently provisioned via the Zadara VPSA interface, and is not compatible with AWS EFS API).

1.6 Load Balancing
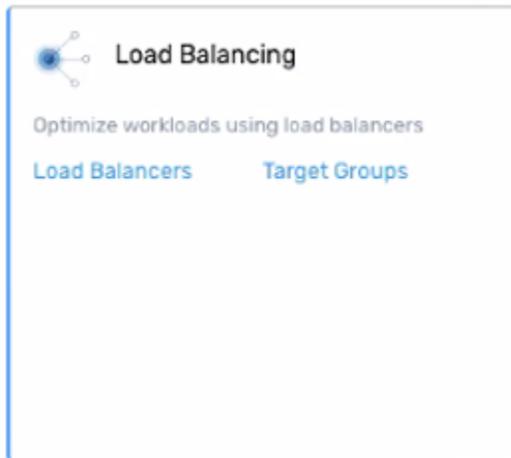
1.6.1 Load Balancers

The Managed Load Balancer service offers the ability to spin up, customize and scale load balancers to support fault tolerance and ensure high availability and application scalability over time. To simplify operations even further, this service supports AWS ELB APIs. It routes traffic according to application or network considerations and provides the required amount of load balancing capacity needed and distributes it to meet high availability and network performance requirements.

1.6.2 Target Groups

A target group is a group of instances to which a load balancer directs application traffic. The instances in this group collectively do the processing work that the application requires.

1.7 Machine Images
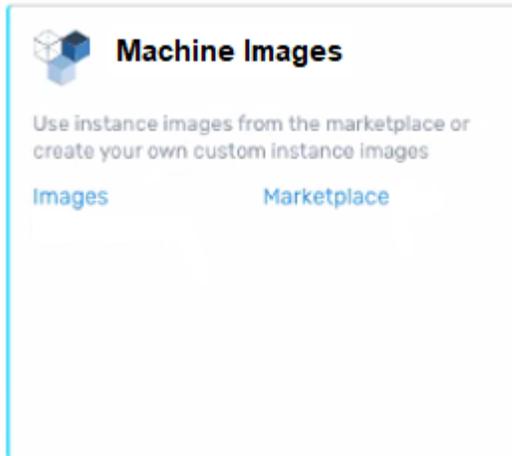
1.7.1 Images

The system comes with a number of VM images ready to use. In addition you can create your own images as needed. Two image formats are supported - VMware OVA image and the KVM compatible image in RAW or QCOW2 formats. The latter means that most OpenStack KVM images can be uploaded easily. All of the above are converted into the RAW format and preserved in the image repository in the cluster.

1.7.2 Images Marketplace

Machine images of the different OSs can be downloaded from the Images Marketplace, and be used to create Instances. For OSs that require activation (and vendor fees) you will need to get the activation codes directly from the vendor.

1.8 Monitoring
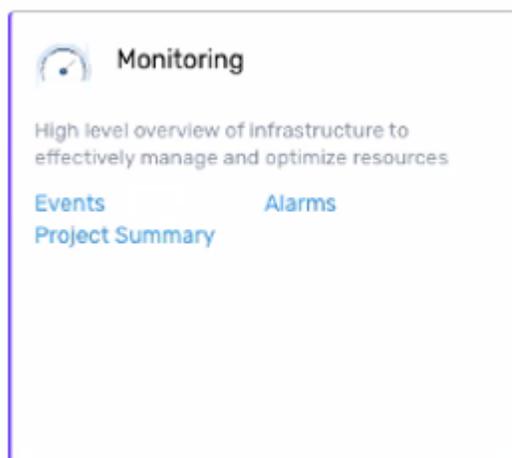
1.8.1 Events

Zadara compute cloud provides information about various events in your system. Events are any actual changes or attempts at changes in the state of the services, whether initiated manually or automatically by the system. This service provides an events viewer with the ability to filter by date, severity, time, and other relevant parameters.

The system offers visualized metrics, logs and events of all resources, including CPU utilization, memory, storage and network usage. Insights are available on a broad system view, as well as on a per-instance and per-volume view, to quickly and effectively resolve issues by analyzing diagnostics to understand the root cause.

You can configure the system events to be sent to a remote syslogs/logstash server located at some external endpoint.

1.8.2 Alerts

The cloud's automated monitoring helps you focus on proactively solving problems, rather than wasting time to identify them. You can set alarms on all relevant metrics and reduce the response time for facilitating automation. You can easily customize a set of metrics-based alarms with multiple conditions to notify you on threshold crossing. You will receive your alerts in real-time.

1.9 Certificates

1.9.1 Certificates

To secure your cloud domain name you need to procure an SSL Certificate from a trusted SSL Certificate Authority and install it on your domain. Once you get the certificate file from the CA, this service allows you to apply the certificate to the system.
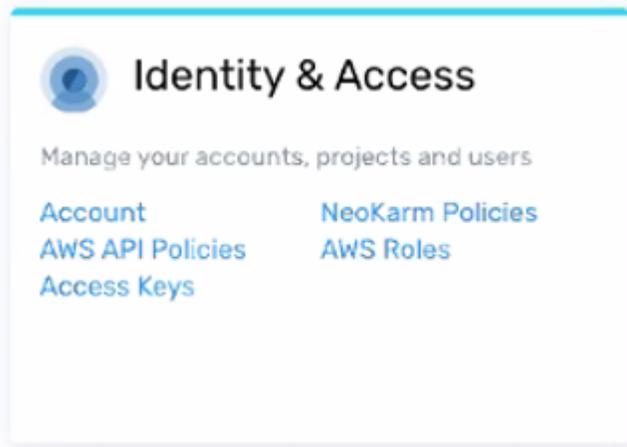
1.10 Protection

1.10.1 Groups (policies)

The Protection Group basically consists of a schedule for automatic local backup of a project. It is a backup policy that defines the backup window, recurring schedule, and its retention period.

In addition to the member services above, the following services are available for Tenant Admins:

1.11 Identity and Access

1.11.1 Account

Tenant Admins have the added ability to monitor and modify the account state, create and manage projects within the account, and member user's access rights.

1.11.2 Access Keys

Tenant Admins can manage account-specific credentials that authenticate users before they can access the system. For example, to access the GUI or Management console, users sign in with a username and password. Similarly, for programmatic access, users provide an access key and a secret key. As the API's are AWS compatible, the authentication mechanism is also identical.

1.11.3 AWS Roles

AWS Roles are policy-based tokens with temporary credentials allowing a user temporary access to AWS services and actions which the user is normally not permitted to access. These users may be from different projects or even different accounts. These roles can also be embedded into specific instances allowing these instances access to the necessary actions.

The AWS roles are independent of the Zadara user roles.

1.11.4 AWS API Policies

Usage of all AWS compatible services and actions are governed by their corresponding AWS-managed policies. These policies can be assigned per project to Users, Groups of users and STS Roles. System usage is governed by Zadara Cloud Services policies together with the Zadara roles.

1.11.5 Zadara Cloud Services API Policies

✔ **Note:** This section is under development.

TWO

MAPPING ZADARA CLOUD SERVICES TO AWS SERVICES

Zadara Cloud Services deliver a software-defined data center platform that enables true IaaS, PaaS and CaaS in data centers and edge locations. Zadara Cloud Services runs on any hardware and is combined with cloud management features such as centralized user access management, self-service portals, integrated metering for showback / chargeback, and more.

In addition, Zadara Cloud Services delivers a suite of managed open source platforms for developers to accelerate application development and delivery. By offering AWS compatible APIs, Zadara Cloud Services enables multi-cloud and hybrid applications, and supports advanced DevOps and Infrastructure-as-Code in enterprise environments.

The following table provides a high-level mapping of Zadara Cloud Services products and AWS services.

2.1 IaaS Services

SERVICE	AWS	ZADARA CLOUD SERVICES	ZADARA-CS-SUPPORTED AWS APIs
Compute	Amazon Elastic Compute Cloud (EC2)	Zadara Compute Service (zCompute)	EC2
	Amazon EC2 Auto Scaling	Zadara Auto Scaling	EC2 Auto Scaling
Virtual Networks	Amazon Virtual Private Cloud (VPC)	Zadara Networking	VPC
Load Balancer	Elastic Load Balancer (ELB)	Zadara Load Balancers	ELB
Object Storage	Amazon Simple Storage Service (S3)	Zadara VPSA Object Storage	S3
Block Storage	Amazon Elastic Block Storage (EBS)	Zadara VPSA Storage Array	EBS
File Storage	Amazon Elastic File System (EFS)	Zadara VPSA Storage Array	

2.2 Management Services

SERVICE	AWS	ZADARA CLOUD SERVICES	ZADARA-CS-SUPPORTED AWS APIs
Monitoring	Amazon CloudWatch	Zadara Monitoring Service	CloudWatch
Security, Identity & Compliance	Amazon Identity & Access Management (IAM)	Zadara Identity Management	IAM
	AWS Security Token Service (STS)		STS
	AWS Certificate Manager (ACM)	Zadara Certificate Manager	ACM
Application & Service Catalog	AWS Service Catalog	Zadara CS Application Catalog	

THREE

GLOSSARY

Affinity Rules

Affinity Rules allow you to control the placement of virtual machines on hosts within a cluster. For example, rules can be defined to prevent virtual machines to be placed on the same host in order to enable robustness in case of a host failure.

API

Application program interface (API) is a set of routines, protocols, and tools for building software applications. An API specifies how software components should interact and can be used when programming graphical user interface (GUI) components. API makes it easier to develop a program by providing all the basic building blocks to a programmer who then integrates them to provide a complete application.

AWS

Amazon Web Services (AWS) is a secure Cloud services platform, offering compute power, database storage, content delivery and other functionality to businesses.

Cloud

The cloud refers to the use of generic resources (e.g. data centers) for storage or computing, in contrast to the use of local, dedicated computers. The resources may be available as a private cloud for a single enterprise or entity, or may be available as a public cloud over the internet with services provided to multiple entities.

Cloud Computing

A computing capability that provides an abstraction between the computing resource and its underlying technical architecture (e.g., servers, storage, networks), enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction. This definition states that clouds have five essential characteristics: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. Narrowly speaking, cloud computing is client-server computing that abstracts the details of the server; one requests a service (resource), not a specific server (machine). Cloud Computing enables Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Cloud computing means that infrastructure, applications, and business processes can be delivered to you as a service, over the Internet (or your own network).

Clusters

A cluster is a collection of nodes. The Zadara system runs over a cluster and uses it as a logical unit to provide:

- Node Management – the cluster is used for suspending and removing nodes, as well as monitoring the node status and activity.

- High Availability – The cluster is responsible for automatically detecting hardware failures, and for handling them by moving workloads from failed nodes to active ones. By performing High Availability procedures, Zadara assures data loss prevention and that the overall cluster operation remains intact.

- Cluster health and operational status – the cluster has pre-defined events and alarms that reflects the state of the system. The events provide information about monitored actions such as the creation and stopping of a workload,

or monitored resource status such as disk failure. Alarms are notifications that are activated in response to an event, or in response to a state of a certain object in the system. They include the event severity and may indicate a major issue in the system.

Data Center

A data center is a centralized repository, either physical or virtual, for the storage, management, and dissemination of data and information organized around a particular body of knowledge or pertaining to a particular business.

Infrastructure as a service (IaaS)

Cloud infrastructure services or "Infrastructure as a Service (IaaS)" delivers computer infrastructure, typically a platform virtualization environment, as a service. Rather than purchasing servers, software, data center space or network equipment, clients instead buy those resources as a fully outsourced service. The service is typically billed on a utility computing basis and amount of resources consumed (and therefore the cost) will typically reflect the level of activity. It is an evolution of web hosting and virtual private server offerings.

Interfaces

You can communicate with your environment by using one of the three existing Zadara interfaces: REST API - the main programmable interface into the system, the CLI, or the web-based graphical User Interface.

Hyperconverged

Hyperconverged is a type of infrastructure system with a software-centric architecture that tightly integrates compute, storage, networking and virtualization resources onto the same underlying hardware.

Hypervisor

A hypervisor or virtual machine monitor (VMM) is a piece of computer software, firmware or hardware that creates and runs virtual machines. A computer on which a hypervisor is running one or more virtual machines is defined as a host machine.

Each virtual machine is called a guest machine.

Monitoring Agent

A Monitoring Agent exists on each node, and it communicates with the Monitoring Service. The Agent provides the Monitoring Service with information, such as how much CPU is being used, the amount of allocated and used disk space on a given node, and more. Based on this information, the Monitoring Service is able to keep track of the system status across all nodes, and initiate changes when resources are running low.

Multi Tenancy

Multi Tenancy refers to a software architecture in which a single instance of software runs on a server and serves multiple tenants. A tenant is a group of users who share a common access with specific privileges to the software instance.

Nodes

Nodes are the physical machines that provide the infrastructure for storage, compute, and memory, and on which Zadara Cloud Services platform is installed. The minimal Zadara configuration requires 3 nodes. Nodes can be added, removed, or put in a maintenance mode after the initial installation and configuration. Virtual Machines are running on the nodes of the cluster. The placement of the Virtual Machines on the nodes is determined by a Zadara management sub-system, which also migrates Virtual Machines from one node to another according to the changing state of the Virtual Machines, nodes, and the cluster as a whole.

Platform as a service (PaaS)

Cloud platform services, whereby the computing platform (operating system and associated services) is delivered as a service over the Internet by the provider.

The PaaS layer offers black-box services with which developers can build applications on top of the compute infrastructure. This might include developer tools that are offered as a service to build services, or data access and database services, or billing services.

Private Cloud

Private cloud is a type of cloud computing that delivers similar advantages to public cloud, including scalability and self-service, but through a proprietary architecture.

Unlike public clouds, which deliver services to multiple organizations, a private cloud is dedicated to a single organization.

Provisioning Services

Zadara Provisioning Services are responsible for:

- Placement – determines where to place resources.

- Monitoring – monitors node resources and workload demands.

- Live Migration – migrates workloads between nodes according to the node loads, the resources required by the workloads, and the priority and profile of the workloads.

Self Service

A feature that allows customers to provision, manage, and terminate services themselves, without involving the service provider, via a Web interface or programmatic calls to service APIs.

Server Sprawl

Server sprawl is a situation in which multiple, under-utilized servers take up more space and consume more resources than can be justified by their workload.

Software as a service (Saas)

Cloud application services, whereby applications are delivered over the Internet by the provider, so that the applications don't have to be purchased, installed, and run on the customer's computers. SaaS providers were previously referred to as ASP (application service providers). In the SaaS layer, the service provider hosts the software so you don't need to install it, manage it, or buy hardware for it. All you have to do is connect and use it. SaaS examples include customer relationship management as a service.

Software Defined Everything

With Software Defined Everything, the computing infrastructure is virtualized and delivered as a service. In a Software-Defined Everything environment, management and control of the networking, storage and/or data center infrastructure is automated by intelligent software typically through API calls, rather than by the hardware components of the infrastructure.

Storage

Zadara Cloud Services uses its award-winning VPSA Storage Array in order to provide an enterprise-grade storage solution for the system backend. System users can leverage file services & object storage services as well.

Tenant

A Tenant is a group of users that has access to certain logical resources in a cluster. The logical resources that a user can access, are the resources that were created by all of the users who belong to the same tenant. Users in the same tenant can have different roles.

Each role defines the type of privileges the user will have. Every user in the system must belong to at least one tenant, and each tenant can include one or more users.

Vendor lock

Dependency on the particular cloud vendor and difficulty moving from one cloud vendor to another due to lack of standardized protocols, APIs, data structures (schema), and service models.

Workloads

Workloads refer to the capacity available to or used by applications on a cloud infrastructure. For example, the available workload which can be supported by a given set of resources represents the total CPU, storage, and networking capacity

available on a given data center. The used workload refers to the actual amount of those resources needed to support a specific service or application.