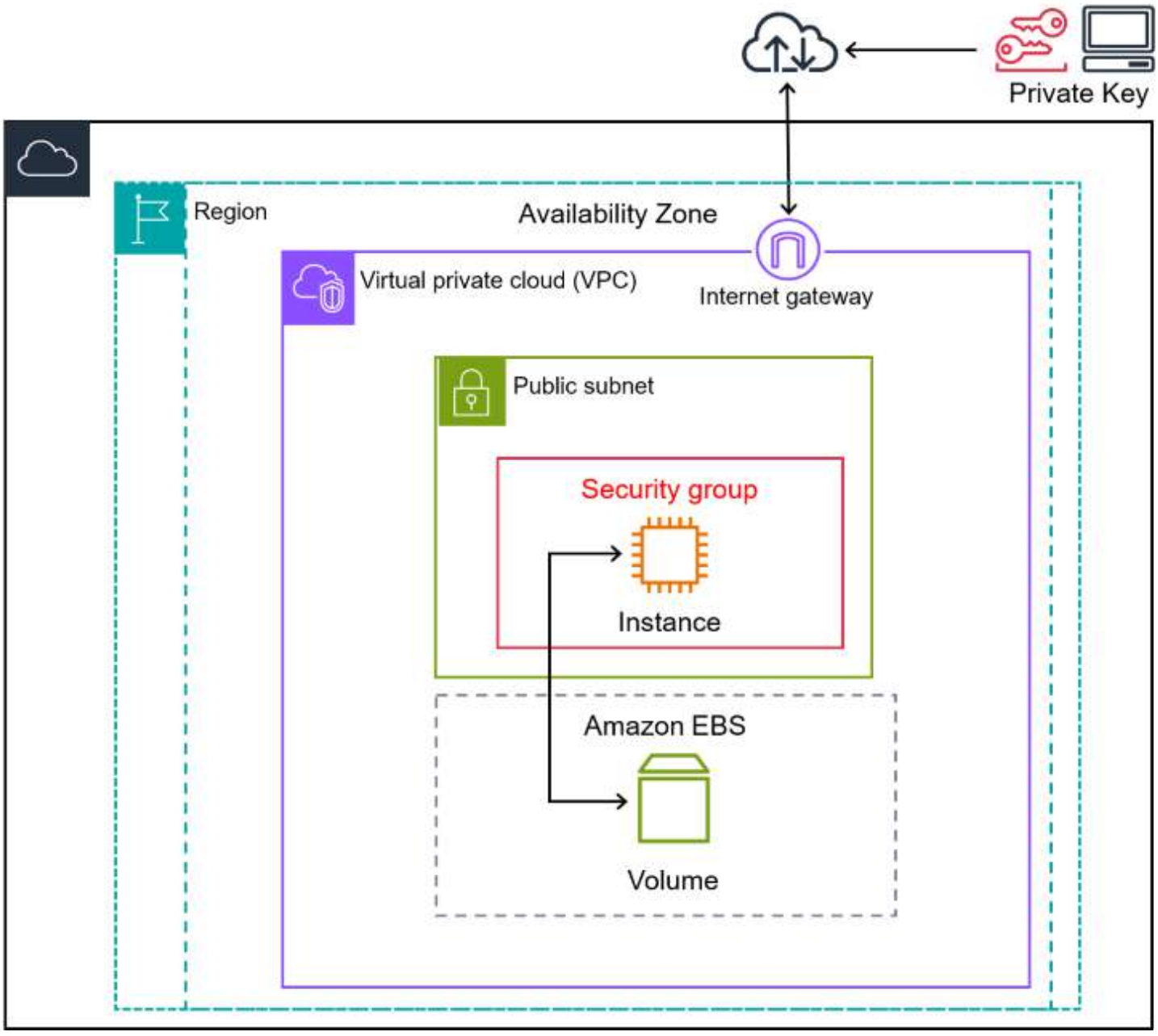
Private Key
Region
Availability Zone
Virtual private cloud (VPC)
Internet gateway
Public subnet
Security group
Instance
Amazon EBS
Volume

Amazon Elastic Compute Cloud (Amazon EC2)

Amazon EC2 offers scalable computing capacity in the AWS cloud, allowing you to easily adjust your compute resources based on your needs. It enables you to scale up or down to handle changes in traffic requirements, reducing the need to forecast accurately.

With Amazon EC2, you can launch as many or as few virtual servers as required, configure security settings, manage networking, and handle storage. EC2 provides two storage options: Elastic Block Store (EBS) and instance store.

Preconfigured templates known as Amazon Machine Images (AMIs) are available for easy setup.

By default, when you create an EC2 account, you are initially limited to a maximum number of instances per EC2 Region. Additionally, there are two default high I/O instances.

Types of EC2 Instances:

General Purpose Instances: These instances offer a balanced combination of compute, memory, and networking resources. They are well-suited for various workloads and include the following series: A, M, and T series.

Compute Optimized Instances: These instances provide more CPU power than RAM and are optimized for CPU-intensive workloads. Examples include the C series.

Memory Optimized Instances: These instances offer more RAM compared to CPU power, making them ideal for memory-intensive tasks. Examples include the R and X series.

Accelerated Computing / GPU Instances: These instances are optimized for tasks that require hardware acceleration, such as GPU-intensive workloads and graphics processing.

Storage Optimized Instances: Designed for low-latency and high-throughput storage, these instances are suitable for workloads that demand efficient access to large datasets. Examples include the I, D, and H series.

High Memory Optimized Instances: These instances are equipped with a high amount of RAM, making them suitable for memory-intensive applications. They leverage the Nitro system for enhanced performance.

General Purpose Instances:

General-purpose instances offer a balanced combination of compute, memory, and networking resources. They include the T series, M series, and A1 instances. T series instances, for instance, provide a baseline level of performance while having the ability to burst to higher levels when needed. These instances are often used in containerized microservices and are well-suited for various workloads within the Arm ecosystem.

Compute Optimized Instances:

These instances are well-suited for compute-bound applications that benefit from high-performance processors. Notable examples include the C4, C5, and C6g instance types, which find use cases in web servers, MMO gaming, and video encoding. C5 instances support a maximum of 25 EBS volumes and utilize a high-speed Elastic Network Adapter (ENA) for network performance. The C5 instances also employ a new EC2 hypervisor to enhance their capabilities.

Memory Optimized Instances:

Memory optimized instances are designed to deliver fast performance for workloads that require processing large datasets in memory. This category includes R series, X

series, and Z series instances, which are optimized for memory-intensive tasks such as in-memory databases and data analytics.

Storage Optimized Instances:

These instances are tailored for workloads that demand high, sequential read and write access to extensive datasets stored on local storage. The I series, D series, and H series fall under this category, offering optimal performance for applications requiring intensive storage operations.

Accelerated Computing Instances: Accelerated computing instances leverage hardware accelerators or co-processors to enhance functions like floating-point number calculations, graphics processing, or pattern matching. Notable instances within this category include P series, G series, and F series instances, which are well-suited for tasks such as high-performance computing and graphics-intensive applications.

High Memory Instances:

High memory instances are purpose-built for running large in-memory databases, including production deployments of SAP HANA in the cloud. These instances are bare metal, meaning they do not run on a hypervisor. The operating system is directly installed in the hardware. They are available only through the Dedicated Host purchasing category and come with direct access to the underlying hardware.

EC2 Instance Purchasing Options:

Amazon EC2 offers various purchasing options to match different usage scenarios. These options include On-Demand Instances, which provide flexibility without long-term commitments; Dedicated Instances, which run on hardware dedicated to a single customer; Dedicated Hosts, which offer dedicated physical servers; Spot Instances, which offer savings for unused capacity; Scheduled Instances, which allow you to purchase capacity on a recurring schedule; and Reserved Instances, which offer cost savings for committed usage.

VPC Endpoint:

A VPC endpoint provides a secure and private connection between your VPC and AWS services, eliminating the need for public IPs. It ensures that resources within your VPC can directly communicate with specific AWS service instances, enhancing security and reducing exposure to the public internet. This direct connection offers efficient and cost-effective data transfer while simplifying networking configurations and maintaining isolation from external networks.