Virtualization

Tejas Parikh (t.parikh@northeastern.edu)

CSYE 6225

Northeastern University

Virtualization

Understanding Virtualization

In computing, virtualization refers to the act of creating a virtual (rather than actual) version of something, including virtual computer hardware platforms, storage devices, and computer network resources.

Limitations of Bare-metal (physical) Server

Servers are designed to run one Operating System and application at a time.

Benefits of Virtualization

- Effective way to reduce IT expenses while boosting efficiency and agility for all size businesses
- Reduce capital and operating costs.
- Minimize or eliminate downtime.
- Increase IT productivity, efficiency, agility and responsiveness.
- Provision applications and resources faster.
- Enable business continuity and disaster recovery.
- Simplify data center management.
- Build a true Software-Defined Data Center.

How Virtualization Works

- Virtualization uses software to simulate the existence of hardware and create a virtual computer system.
- Doing this allows businesses to run more than one virtual system and multiple operating systems and applications on a single server.
- This can provide economies of scale and greater efficiency.

The Virtual Machine

- A virtual computer system is known as a “virtual machine” (VM): a tightly isolated software container with an operating system and application inside.
- Each self-contained VM is completely independent.
- Putting multiple VMs on a single computer enables several operating systems and applications to run on just one physical server, or “host”.

Hypervisor

A thin layer of software called a **hypervisor** decouples the virtual machines from the host and dynamically allocates computing resources to each virtual machine as needed.

Key Properties of Virtual Machines

- Partitioning
  - Run multiple operating systems on one physical machine
  - Divide system resources between virtual machines
- Isolation
  - Provide fault and security isolation at the hardware level
  - Preserve performance with advanced resource controls
- Encapsulation
  - Save the entire state of a virtual machine to files
  - Move and copy virtual machines as easily as moving and copying files
- Hardware Independence
  - Provision or migrate any virtual machine to any physical server

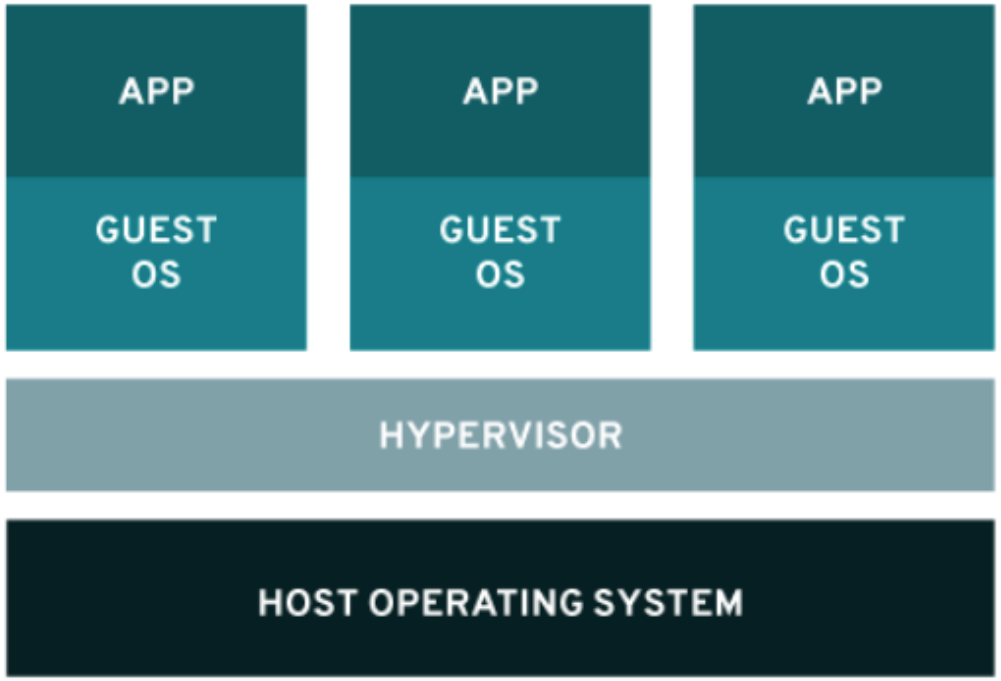
Server Consolidation

Using server virtualization, a company can maximize the use of its server resources and reduce the number of servers required. The result is server consolidation, which improves efficiency and cuts costs.

Virtualization Is Not Cloud Computing

- Cloud computing is not the same thing as virtualization; rather, it's something you can do using virtualization.
- Cloud computing describes the delivery of shared computing resources (software and/or data) on demand through the Internet.
- Whether or not you are in the cloud, you can start by virtualizing your servers and then move to cloud computing for even more agility and increased self-service.

VIRTUALIZATION



Amazon Elastic Compute Cloud (Amazon EC2)

Amazon EC2

- Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides secure, resizable compute capacity in the cloud. It is designed to make web-scale cloud computing easier for developers.
- Amazon EC2's simple web service interface allows you to obtain and configure capacity with minimal friction.
- Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change.
- Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use.
- Amazon EC2 provides developers the tools to build failure resilient applications and isolate them from common failure scenarios.
- **Amazon EC2 is a virtual machine in the cloud.**

Benefits of Amazon EC2

- Elastic Computing
- Complete Control
- Flexibility
- Reliable
- Secure
- Pay-as-you-go
- Integrated*
- Easy to Start*

Amazon EC2 Instance/Machine Types

- Amazon EC2 provides a wide selection of instance types optimized to fit different use cases.
- Instance types comprise varying combinations of CPU, memory, storage, and networking capacity and give you the flexibility to choose the appropriate mix of resources for your applications.
- Each instance type includes one or more instance sizes, allowing you to scale your resources to the requirements of your target workload.

Amazon EC2 Instance Type Categories

- General Purpose
- Compute Optimized
- Memory Optimized
- Accelerated Computing (GPU)
- Storage Optimized

Amazon EC2 Pricing

Various pricing models

- On-Demand
- Spot Instances
- Reserved Instances
- Dedicated Hosts
- Dedicated Instances

On-Demand Instances

- With On-Demand instances, you pay for compute capacity by the hour with no long-term commitments or upfront payments. You can increase or decrease your compute capacity depending on the demands of your application and only pay the specified hourly rate for the instances you use.
- On-Demand instances are recommended for:
  - Users that prefer the low cost and flexibility of Amazon EC2 without any up-front payment or long-term commitment
  - Applications with short-term, spiky, or unpredictable workloads that cannot be interrupted
  - Applications being developed or tested on Amazon EC2 for the first time

Reserved Instances

- Reserved Instances provide you with a significant discount (up to 75%) compared to On-Demand instance pricing.
- In addition, when Reserved Instances are assigned to a specific Availability Zone, they provide a capacity reservation, giving you additional confidence in your ability to launch instances when you need them.
- For applications that have steady state or predictable usage, Reserved Instances can provide significant savings compared to using On-Demand instances.
- Reserved Instances are recommended for:
  - Applications with steady state usage
  - Applications that may require reserved capacity
  - Customers that can commit to using EC2 over a 1 or 3 year term to reduce their total computing costs

Spot Instances

- Amazon EC2 Spot instances allow you to bid on spare Amazon EC2 computing capacity for up to 90% off the On-Demand price.
- Spot instances are recommended for:
  - Applications that have flexible start and end times
  - Applications that are only feasible at very low compute prices
  - Users with urgent computing needs for large amounts of additional capacity

Dedicated Hosts

- A Dedicated Host is a physical EC2 server dedicated for your use.
- Dedicated Hosts can help you reduce costs by allowing you to use your existing server-bound software licenses, including Windows Server, SQL Server, and SUSE Linux Enterprise Server (subject to your license terms), and can also help you meet compliance requirements

Dedicated Instances

- Dedicated Instances are Amazon EC2 instances that run in a VPC on hardware that's dedicated to a single customer.
- Your Dedicated instances are physically isolated at the host hardware level from instances that belong to other AWS accounts.
- Dedicated instances may share hardware with other instances from the same AWS account that are not Dedicated instances.

Dedicated Hosts vs. Dedicated Instances

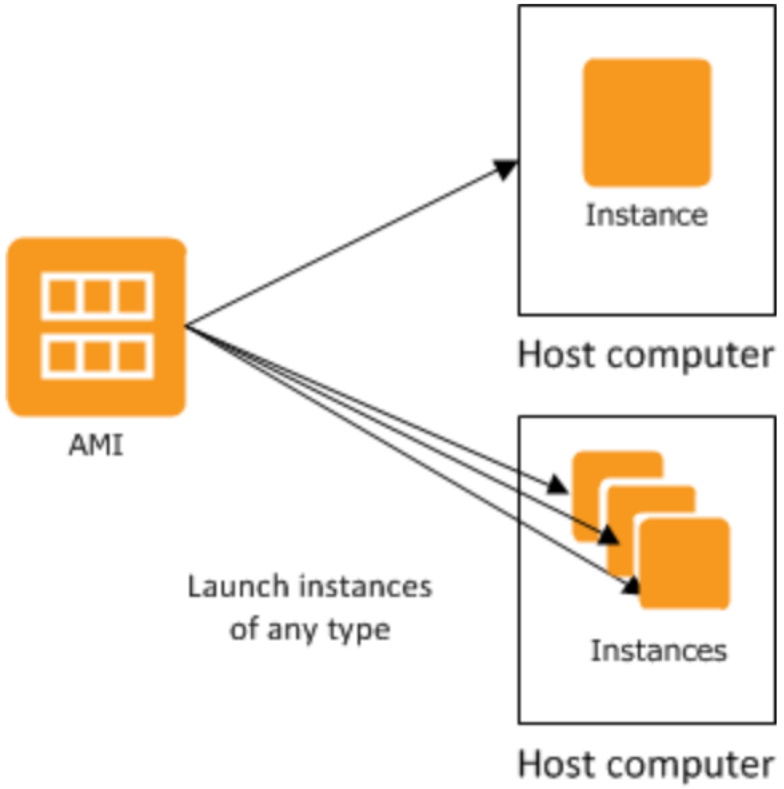
Characteristic	Dedicated Instances	Dedicated Hosts
Enables the use of dedicated physical servers	X	X
Per instance billing (subject to a $2 per region fee)	X	
Per host billing		X
Visibility of sockets, cores, host ID		X
Affinity between a host and instance		X
Targeted instance placement		X
Automatic instance placement	X	X
Add capacity using an allocation request		X

Amazon EC2 Features

- Secure login using key pairs
- Temporary & Permanent Storage Volumes
- Security Groups (firewalls)
- Static IPv4 address (Elastic IP address)
- Metadata, known as tags
- Subnets (virtual networks)

Amazon Machine Image (AMI) & Instances

- An *Amazon Machine Image (AMI)* is a template that contains a software configuration (for example, an operating system, an application server, and applications).
- From an AMI, you launch an *instance*, which is a copy of the AMI running as a virtual server in the cloud.
- You can launch multiple instances of an AMI, as shown in the following figure.
- Your instances keep running until you stop or terminate them, or until they fail.
- If an instance fails, you can launch a new one from the AMI.

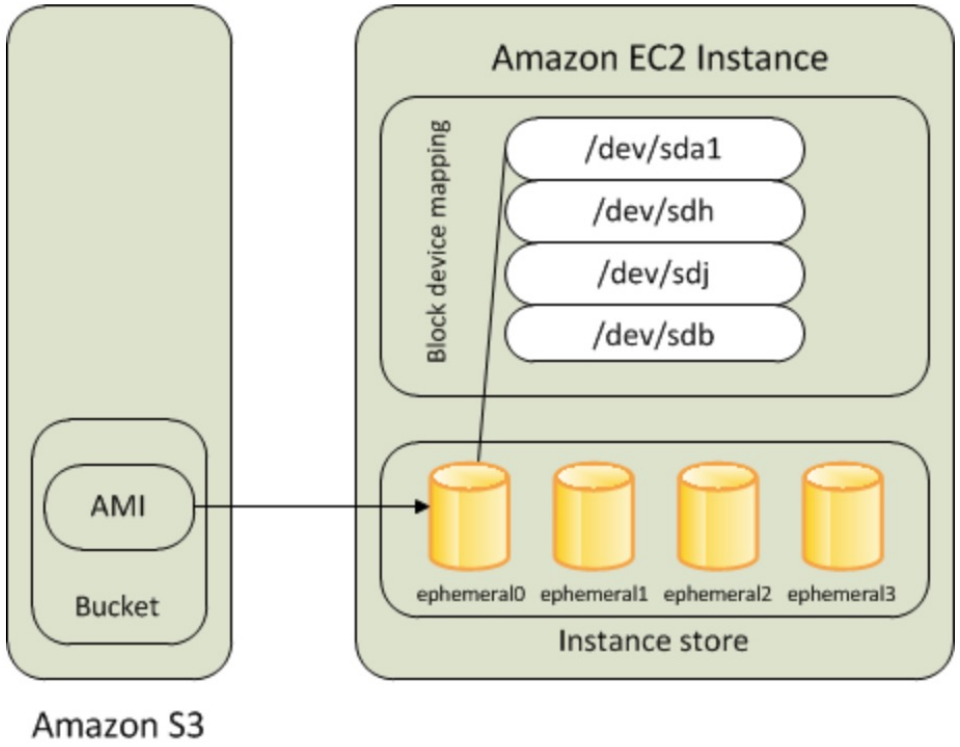


Amazon EC2 Root Device Volume

- When you launch an instance, the *root device volume* contains the image used to boot the instance.
- Instance can be launched with either
  - AMIs backed by Amazon EC2 instance store
  - AMIs backed by Amazon EBS

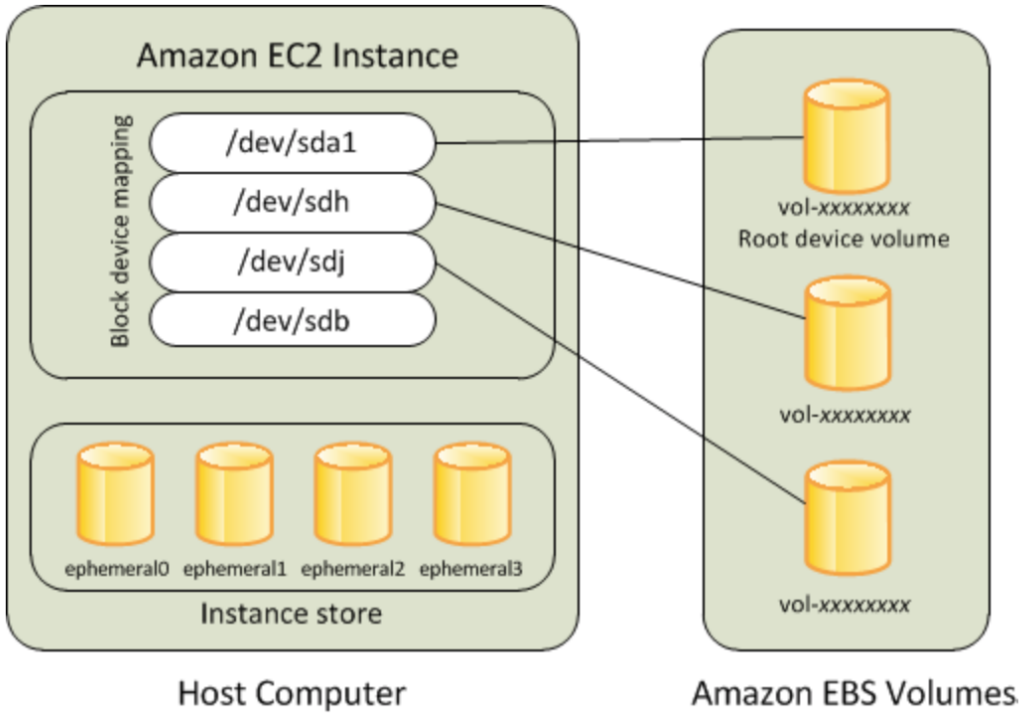
AMIs backed by Amazon EC2 Instance Store

- Instances that use instance stores for the root device automatically have one or more instance store volumes available, with one volume serving as the root device volume.
- When an instance is launched, the image that is used to boot the instance is copied to the root volume.
- Any data on the instance store volumes persists as long as the instance is running, but this data is deleted when the instance is terminated (instance store-backed instances do not support the Stop action) or if it fails (such as if an underlying drive has issues)



Amazon EBS-backed Instances

- Instances that use Amazon EBS for the root device automatically have an Amazon EBS volume attached.
- When you launch an Amazon EBS-backed instance, an Amazon EBS volume is created for each Amazon EBS snapshot referenced by the AMI used.
- An Amazon EBS-backed instance can be stopped and later restarted without affecting data stored in the attached volumes.



Amazon Elastic Block Store (Amazon EBS)

- Amazon Elastic Block Store (Amazon EBS) provides block level storage volumes for use with EC2 instances.
- EBS volumes are highly available and reliable storage volumes that can be attached to any running instance that is in the same Availability Zone.
- EBS volumes that are attached to an EC2 instance are exposed as storage volumes that persist independently from the life of the instance.
- Amazon EBS is recommended when data must be quickly accessible and requires long-term persistence.

	Solid-State Drives (SSD)		Hard disk Drives (HDD)	
Volume Type	General Purpose SSD (gp2)*	Provisioned IOPS SSD (io1)	Throughput Optimized HDD (st1)	Cold HDD (sc1)
Description	General purpose SSD volume that balances price and performance for a wide variety of transactional workloads	Highest-performance SSD volume designed for mission-critical applications	Low cost HDD volume designed for frequently accessed, throughput-intensive workloads	Lowest cost HDD volume designed for less frequently accessed workloads

Use Cases	• Recommended for most workloads
	• System boot volumes
	• Virtual desktops
	• Low-latency interactive apps
	• Development and test environments
		• Critical business applications that require sustained IOPS performance, or more than 10,000 IOPS or 160 MiB/s of throughput per volume
		• Large database workloads, such as:
		  ○ MongoDB
		  ○ Cassandra
		  ○ Microsoft SQL Server
		  ○ MySQL
		  ○ PostgreSQL
		  ○ Oracle
			• Streaming workloads requiring consistent, fast throughput at a low price
			• Big data
			• Data warehouses
			• Log processing
			• Cannot be a boot volume
				• Throughput-oriented storage for large volumes of data that is infrequently accessed
				• Scenarios where the lowest storage cost is important
				• Cannot be a boot volume

API Name	gp2	io1	st1	sc1
Volume Size	1 GiB - 16 TiB	4 GiB - 16 TiB	500 GiB - 16 TiB	500 GiB - 16 TiB
Max. IOPS/Volume**	10,000	20,000	500	250
Max. Throughput/Volume†	160 MiB/s	320 MiB/s	500 MiB/s	250 MiB/s
Max. IOPS/Instance	75,000	75,000	75,000	75,000
Max. Throughput/Instance	1,750 MB/s	1,750 MB/s	1,750 MB/s	1,750 MB/s
Dominant Performance Attribute	IOPS	IOPS	MiB/s	MiB/s

Input/output Operations Per Second (IOPS)

- IOPS are a unit of measure representing input/output operations per second. The operations are measured in KiB, and the underlying drive technology determines the maximum amount of data that a volume type counts as a single I/O.
- I/O size is capped at 256 KiB for SSD volumes and 1,024 KiB for HDD volumes because SSD volumes handle small or random I/O much more efficiently than HDD volumes.
- Example: A single 1,024 KiB I/O operation counts as 4 operations (1,024÷256=4), while 8 contiguous I/O operations at 32 KiB each count as 1operation (8×32=256). However, 8 random I/O operations at 32 KiB each count as 8 operations. Each I/O operation under 32 KiB counts as 1 operation.

Security Group

- A *security group* acts as a virtual firewall that controls the traffic for one or more instances.
- When you launch an instance, you associate one or more security groups with the instance.
- You add rules to each security group that allow traffic to or from its associated instances.
- You can modify the rules for a security group at any time; the new rules are automatically applied to all instances that are associated with the security group.
- All the rules from all the security groups that are associated with the instance are evaluated to decide if traffic should be allowed or not.

Instance Metadata and User Data

- *Instance metadata* is data about your instance that you can use to configure or manage the running instance.
- You can also use instance metadata to access *user data* that you specified when launching your instance.
- For example, you can specify parameters for configuring your instance, or attach a simple script.

Additional Resources

See Lecture Page