**Notes :** A hypervisor is a software that you can use to run multiple virtual machines on a single physical machine. Every virtual machine has its own operating system and applications. The hypervisor allocates the underlying physical computing resources such as CPU and memory to individual virtual machines as required. Thus, it supports the optimal use of physical IT infrastructure.

**Hypervisor:**

- Software for multiple VMs on one physical machine.

- Allocates CPU, memory for VMs.

- Enables optimal hardware use.

**Importance:**

- Basis for virtualization.

- Creates multiple VMs on a single host.

- Enables efficient cloud scaling.

**Benefits:**

1. Hardware Independence:

- Abstracts hardware.

2. Efficiency:

- Simplifies server OS setup.

3. scalability:

- Maximizes resource use.

4. Portability:

- Shifts resources easily.

**Use Cases:**

1. Desktop Virtualization:

- Remote workstation access.

2. Resource Optimization:

- Consolidates functions.

3. Failure Recovery:

- Snapshots for restoration.

4. Legacy System Continuity:

- Virtualizes hardware.

**How it Works:**

- Installed on servers.

- Intermediary for resource allocation.

**Types of Hypervisors:**

1. Type 1 (Bare-Metal):

- Efficient, direct resource management.

- Secure and stable.

2. Type 2 (Hosted):

- Installed on host OS.

- Introduces latency, OS-dependent.

**Cloud Hypervisor:**

- Cloud data center virtualization.

- Allows multi-tenancy, logical independence.

**Hypervisors vs. Containers:**

- Hypervisor: Hardware abstraction.

- Containers: OS abstraction.

**Security Considerations:**

- Hypervisor critical for VM security.

- Built-in safeguards essential.

**AWS Nitro System:**

- AWS lightweight hypervisor.

- Delegates tasks to specific hardware.

- Monitors resources, enhances performance.

- Compatible with existing workloads.

**# Instance Purchasing Options**

**1.On-Demand Instances:**

- Pay by the second for launched instances.

- **Savings Plans:**

- Commit to consistent usage for 1 or 3 years.

- Reduce costs with a USD per hour commitment

- **Reserved Instances:**

- Commit to consistent instance configuration.

- Term: 1 or 3 years.

- **Spot Instances:**

- Request unused instances.

- Significant cost reduction.

- **Dedicated Hosts:**

- Pay for a dedicated physical host.

- Use existing software licenses for cost reduction.

**- Dedicated Instances:**

- Pay by the hour for single-tenant hardware.

**1.1 On-Demand Instances: Working Methods**

**1. Launch:**

- Create and start an On-Demand Instance.

**2. Connect:**

- Access your Linux instance.

**3. Stop and Start:**

- Temporarily halt and resume your instance.

**4. Hibernate:**

- Preserve state for On-Demand or Spot Instances.

**5. Reboot:**

- Restart your instance.

**6. Instance Retirement:**

- Handle instances reaching retirement.

**7. Terminate:**

- Permanently end an instance.

**8. Recover:**

- Restore or retrieve a terminated instance.

9. Configure:

- Customize settings for Amazon Linux instances.

10. Identify:

- Recognize and manage EC2 Linux instances.

**2. Reserved Instances (RIs): -** AWS Reserved Instances (RIs): Up to 75% discount on On-Demand Instances with advance purchase for one or three years.

* Three Payment Options:
* All Up-front Reserved Instances (AURI)
* Partial Up-front Reserved Instances (PURI)
* No Up-front Reserved Instances (NURI)
* Booking Computing Power: organizations "book" computing power in advance, ensuring cost predictability.
* Duration Options: RIs available for one or three-year terms.
* Savings: Significant cost savings compared to On-Demand Instances.
* Payment Flexibility: Different payment options (AURI, PURI, NURI) for managing upfront costs.
* Advance Purchase: RIs require upfront payments for reserved capacity.

**- Cost Savings:**

- Significantly reduce Amazon EC2 costs compared to On-Demand pricing.

**- Nature:**

- Not physical instances.

- Billing discount applied to On-Demand Instances.

- **Attributes for Discount:**

- On-Demand Instances must match specific attributes.

- Example: Instance type, Region.

**2.1 Reserved Instances Scope:**

**- Regional:**

- Purchased for a Region.

- Referred to as regional Reserved Instances.

**- Zonal:**

- Purchased for a specific Availability Zone.

- Referred to as zonal Reserved Instances

**2.2 Types of Reserved Instances:**

- **Offering Classes:**

- Standard or Convertible.

- **Standard Reserved Instance:**

- Higher discount.

- Non-exchangeable.

- Modifiable.

- **Convertible Reserved Instance:**

- Lower discount.

- Exchangeable.

- Modifiable.

**2.3 Application of Reserved Instances:**

**- Nature:**

- Billing discount, not physical instances.

**- Applied To:**

- Running On-Demand Instances.

- **Matching Specifications:**

- On-Demand Instances must match Reserved Instance specs.

- **Automatic Application:**

- Billing discount applies immediately.

- No need to restart instances.

- **Procedure**:

- If no eligible running On-Demand Instance, launch one with matching specs.

**Differences between Standard and Convertible Reserved Instances**

**Standard Reserved Instance**:

* Modify Reserved Instances: Some attributes can be modified.
* Exchange Reserved Instances: Can't be exchanged.
* Sell in the Reserved Instance Marketplace: Can be sold in the Reserved Instance Marketplace.

**Convertible Reserved Instance:**

* Modify Reserved Instances: Some attributes can be modified.
* Exchange Reserved Instances: Can be exchanged during the term for another Convertible Reserved Instance with new attributes, including instance family, instance type, platform, scope, or tenancy.
* Sell in the Reserved Instance Marketplace: Can't be sold in the Reserved Instance Marketplace.

**3. Spot Instances:**

**- Definition:**

- Uses spare EC2 capacity.

- Available for less than On-Demand price.

**- Cost Benefits:**

- Steep discounts on EC2 costs.

- Hourly price known as Spot price.

**- Pricing Mechanism:**

- Set by Amazon EC2.

- Adjusted based on supply and demand.

**- Availability:**

- Runs whenever capacity is available.

**- Suitability:**

- Cost-effective for flexible, interruptible applications.

**- Use Cases:**

- Data analysis, batch jobs, background processing, optional tasks.

**3.1 Spot Instance Interruptions:**

**- Launch on Spare EC2 Capacity:**

- Offers steep discounts but may be interrupted.

**- Spot Instance Interruption:**

- Event when EC2 reclaims Spot Instance.

**- Interrupt Action:**

- Termination, stopping, or hibernation based on user's specification.

**- Demand and Availability:**

- Varies significantly, subject to EC2 capacity.

**- Possibility of Interruption:**

- Spot Instances may be interrupted.

**- Fleet Instances:**

- On-Demand Instances in EC2 Fleet or Spot Fleet cannot be interrupted.

**3.2 - Reasons for Spot Instance Interruption:**

**1. Capacity:**

- EC2 reclaims Spot Instances for repurposing capacity.

- Host maintenance or hardware decommission.

**2. Price:**

- Spot price exceeds your specified maximum price.

- Specifying a maximum price may increase interruptions.

**3. Constraints:**

- Termination as a group if constraints (launch group, AZ group) can't be met.

**3.3 - Specifying Interruption Behavior:**

**- Default Behavior:**

- EC2 terminates Spot Instances by default.

**- Launch Instance Wizard:**

- Advanced details > Request Spot Instances > Customize > Interruption behavior.

- Option for hibernation (Stop - Hibernate behavior).

**- run-instances CLI:**

- (--instance-market-options) > Instance Interruption Behavior.

- Hibernation: --hibernation-options Configured=true.

**- Launch Template:**

- Advanced details > Request Spot Instances > Customize > Interruption behavior.

**- Spot Console:**

- Maintain target capacity > Interruption behavior.

**- create-fleet CLI:**

- Instance Interruption Behavior in request configuration.

- **request-spot-fleet CLI:**

- Instance Interruption Behavior in request configuration.

**- request-spot-instances CLI:**

- --instance-interruption-behavior for specifying behavior.

**3.4 - Hibernate for Interrupted Spot Instances:**

**- Specification:**

- Option to hibernate Spot Instances during interruption.

**- Hibernation Experience:**

- Similar to On-Demand Instances.

**- Enhancements:**

- More supported AMIs.

- Expanded support for instance families.

- User-initiated hibernation now supported.

**4. - Burstable Performance Instances (T Types):**

**- Nature:**

- T instance types are burstable performance instances.

**- Launch Mode Recommendation:**

- For immediate, short-duration usage, launch in Standard mode to avoid higher costs.

- **Unlimited Mode Consideration:**

- Suitable if instance runs long enough to accrue CPU credits for bursting.

- Otherwise, more expensive than other instances due to surplus credit charges.

**- Launch Credits:**

- T2 instances in Standard mode receive launch credits for initial launch experience.

- Repeated launches for new credits are not allowed.

**- Sustained CPU Usage:**

- For sustained CPU usage, use Unlimited mode for T2 Spot Instances or dedicated CPU instance types.

**4.1 - Dedicated Hosts:**

**- Definition:**

- Physical server fully dedicated to your EC2 instance capacity.

- **Purpose:**

- Allows use of existing per-socket, per-core, or per-VM software licenses.

**- License Compatibility:**

- Supports licenses for Windows Server, Microsoft SQL Server, SUSE, and Linux Enterprise Server.

**4.2 - Burstable T3 Instances on Dedicated Hosts:**

**- Support:**

- Dedicated Hosts support T3 instances.

**- Cost Efficiency:**

- Cost-efficient BYOL license use on dedicated hardware.

**- vCPU Footprint:**

- Smaller vCPU footprint for workload consolidation.

**- Suitable For:**

- Low to moderate CPU utilization workloads.

- Small to medium databases, virtual desktops, development and test environments, code repositories, and product prototypes.

**- Not Recommended For:**

- Sustained high CPU utilization or correlated CPU bursts.

**- Credit Model:**

- Standard mode, no unlimited credit mode support.

**- Performance Characteristics:**

- Baseline CPU with burst capability using accrued credits.

- Credits spent during bursts; lowered CPU when credits are depleted.

**- T3 Dedicated Host Features:**

- Supports EC2 Dedicated Host features (multiple sizes, Host resource groups, BYOL).

**- Supported Sizes/Configurations:**

- General-purpose burstable T3 instances on Dedicated Hosts.

- 48 cores per T3 Dedicated Host, supporting up to 192 instances.

- Automatic calculation of supported instances and sizes.

- Multiple T3 instance types supported on the same host, up to CPU limit.

**4.3 - Shared Dedicated Hosts**

**- Sharing Capability:**

- Allows Dedicated Host owners to share with other AWS accounts or within an AWS organization.

**- Centralized Management:**

- Enables central creation and management of Dedicated Hosts.

**- Model Overview:**

- Owner (AWS account) shares with consumers (other AWS accounts).

- Consumers can launch instances on shared Dedicated Hosts as if they were in their own account.

**- Responsibilities:**

- Owner manages Dedicated Host and instances launched on it.

- Consumers manage instances launched on Dedicated Hosts shared with them.

**- Owner Actions:**

- Can't modify instances launched by consumers on shared Dedicated Hosts.

**- Consumer Actions:**

- Can't modify Dedicated Hosts shared with them or instances owned by others.

**- Sharing Options:**

- Share with specific AWS accounts (inside/outside organization).

- Share with organizational units inside the AWS organization.

- Share with the entire AWS organization.

**4.4 - Host Recovery:**

**- Auto Recovery:**

- Restarts instances on a new host when specific issues detected.

**- Purpose:**

- Reduces manual intervention and operational burden during unexpected host failures.

**- Conditions for Recovery:**

- Detected issues related to system power or network connectivity.

**- Manual Intervention:**

- Required for other Dedicated Host issues.

**4.5 - Host Maintenance Basics:**

**- Degradation Detection:**

- Triggered by hardware issues or specific problematic conditions.

**- Recovery Process:**

- New Dedicated Host allocated; instances on degraded host automatically rebooted.

**- Attributes Retained:**

- Auto placement settings, Availability Zone, Reservation, Host affinity, Host maintenance settings, Host recovery settings, Instance type, Tags.

**- Release of Degraded Host:**

- Occurs after instances are rebooted or stopped on the new Dedicated Host.

**- Access to Degraded Host:**

- Instances accessible before scheduled maintenance but no new launches supported.

**- Use of New Host:**

- Can launch new instances on the new Dedicated Host before maintenance.

- **Capacity Reservation:**

- Some capacity reserved for rebooting instances from the degraded host.

**- Availability:**

- Host maintenance available in all AWS Regions for supported Dedicated Hosts.

**5. Dedicated Instances:**

**- Default Tenancy:**

- EC2 instances run on shared hardware by default.

**- Dedicated Instances:**

- Run on hardware dedicated to a single customer.

- Physically isolated for different AWS accounts, even if linked to a common payer account.

- Might share hardware with non-Dedicated Instances from the same AWS account.

**- Dedicated Host:**

- Physical server dedicated for customer use.

- Provides visibility and control over instance placement on the server.

**5.1 - Dedicated Instances vs. Dedicated Hosts:**

**- Common Features:**

- Used for launching EC2 instances on dedicated physical servers.

- No performance, security, or physical differences.

**- Differences:**

**- Billing:**

- Dedicated Host: Per-host billing.

- Dedicated Instance: Per-instance billing.

**- Visibility:**

- Dedicated Host: Provides visibility of sockets, cores, and host ID.

- Dedicated Instance: No visibility.

**- Host and Instance Affinity:**

- Dedicated Host: Allows consistent deployment to the same physical server over time.

- Dedicated Instance: Not supported.

**- Targeted Instance Placement:**

- Dedicated Host: Provides additional visibility and control over instance placement.

- Dedicated Instance: Not supported.

**- Automatic Instance Recovery:**

- Dedicated Host: Supported (Host recovery).

- Dedicated Instance: Supported.

**- Bring Your Own License (BYOL):**

- Dedicated Host: Supported.

- Dedicated Instance: Partial support.

**- Capacity Reservations:**

- Dedicated Host: Not supported.

- Dedicated Instance: Supported.

**6. - Capacity Reservations Overview:**

**- Definition:**

- Reserve compute capacity for EC2 instances in a specific Availability Zone.

**- Types:**

**- 1. On-Demand Capacity Reservations:**

- Use Cases:

- Scaling events.

- Regulatory requirements and disaster recovery.

**- When to Use:**

- Strict capacity requirements.

- Business-critical workloads.

**- 2. Capacity Blocks for ML:**

**- Use Cases:**

- ML model training and fine-tuning.

- ML experiments and prototypes.

**- When to Use:**

- Ensure uninterrupted access to GPU instances for a defined period.

- Ideal for ML workloads, experiments, and short-term surges in inference demand.

**6.1 - On-Demand Capacity Reservations:**

**- Purpose:**

- Reserve compute capacity for EC2 instances in a specific Availability Zone.

- Mitigate the risk of being unable to get On-Demand capacity during constraints.

**- Recommendation:**

- Suitable for strict capacity requirements and business-critical workloads.

**- Flexibility:**

- Reserve capacity for any duration without a long-term commitment.

- Start and stop capacity assurance as needed.

**- Billing and Discounts:**

- Billing starts as soon as the Reservation is provisioned.

- Utilize billing discounts from Savings Plans and Regional Reserved Instances.

**- Reservation Configuration:**

- Specify:

- Availability Zone.

- Number of reserved instances.

- Instance attributes (type, tenancy, platform/OS).

**- Usage and Matching:**

- Automatically used by running instances with matching attributes.

- Unused if no running instances match the specified attributes.

- Attributes include instance type, tenancy, and platform/OS.

**6.2 - Pricing and Billing:**

**- Payment Options:**

- On-Demand Dedicated Hosts

- Dedicated Host Reservations

- Savings Plans

**- On-Demand Dedicated Hosts:**

- Billing activated upon host allocation.

- Price varies by family and Region.

- Pay per second, minimum 60 seconds.

- Release anytime to stop charges.

**- Dedicated Host Reservations:**

- Provide billing discount compared to On-Demand.

- Three payment options: No Upfront, Partial Upfront, All Upfront.

- Cover one or more hosts in a single Availability Zone.

- Reservation associated hosts can't be released until term ends.

**- Savings Plans:**

- Flexible pricing model for significant savings.

- Commitment to consistent usage, USD per hour, for one or three years.

- Offers flexibility to use Dedicated Hosts as needed.

- Not supported with specific Dedicated Host types.

**- Pricing for Windows Server on Dedicated Hosts:**

- Bring existing licenses with no additional charge.

- Use Windows Server AMIs for latest versions on Dedicated Hosts.

- Supported on current generation instance types.

**6.3 -Capacity Reservation Pricing and Billing:**

**- Pricing:**

- Charged at On-Demand rate, whether used or not.

- Unused reservations appear on the bill.

- No upfront or additional charges.

- Billing discounts for Savings Plans and Regional Reserved Instances apply.

**- Billing:**

- Billing starts upon reservation provisioning.

- Billed at per-second granularity.

- Charged for partial hours.

- Example: Reservation for m4.large Linux instance, billed based on usage hours.

- Billing discounts apply to matching attributes.

- Discounts applied preferentially to instance usage before covering unused reservations.

**- Billing Discounts:**

- Savings Plans and Regional Reserved Instances discounts apply.

- Applied to Capacity Reservations with matching attributes.

- Automatically applied to used instances.

- Discounts preferentially applied to instance usage.

- Zonal Reserved Instances discounts do not apply to Capacity Reservations.

**# SAVING PLANS**

- Flexible pricing model

- EC2, Fargate, Lambda usage

- Low prices

- Commitment to consistent usage

- 1 or 3-year term

- $/hour measurement

- Flexibility in compute options

- Automatic cost savings

- No need for exchanges or modifications

- Discounted Savings Plans price

- Easily reduce your bill

- Commitment to compute usage

- Compute Savings Plans

**1. Compute Savings Plans**

* Flexible
* Reduce costs by up to 66%
* Automatically apply to:
* EC2 instance usage (regardless of instance characteristics)
* Fargate usage
* Lambda usage (Duration, Provisioned Concurrency, Provisioned Duration)

**2. EC2 Instance Savings Plans**

* Significant Discounts
* Lowest prices
* Savings up to 72%
* Commitment to usage of individual instance families in a region
* e.g., M5 usage in N. Virginia
* Automatic cost reduction on selected instance family
* Regardless of AZ, size, OS, or tenancy
* Flexibility to change usage between instances within a family in that region

**Rest notes of AMIs in Note book ….**

**Notes**: By default, EBS root volumes have the DeleteOnTermination flag set to true. For information about how to change this flag so that the volume persists after termination

**# Amazon S3**

**Amazon S3 Overview:**

* Object storage service by Amazon Web Services (AWS).
* Scalable, durable, and versatile for various use cases.

**Key Features:**

* Stores data as objects in buckets.
* Highly scalable and supports virtually unlimited data.
* High data availability and durability across geographically dispersed data centers.

**Security:**

* Access control lists (ACLs) and bucket policies for data access control.
* Supports server-side encryption for enhanced security.

**Performance:**

* Low-latency and high-throughput performance.
* Multipart uploads for efficient handling of large objects.

**Use Cases:**

* Data lakes, websites, mobile applications, backup and restore, archive, enterprise applications, IoT devices, and big data analytics.

**Management Features:**

* Versioning, lifecycle policies, logging, and event notifications.

**Pay-as-You-Go Model:**

* + Billing based on actual usage of storage and data transfer.

**Conclusion:**

* + Reliable and scalable storage solution for diverse storage needs in cloud computing.

1. **Amazon S3 Features**

* **Storage Classes:**
  + Different classes for varied use cases (e.g., S3 Standard, S3 Glacier).
  + S3 Express One Zone for high-performance, low-latency storage.
* **Storage Management:**
  + S3 Lifecycle for cost-effective object management.
  + S3 Object Lock for data protection.
  + S3 Replication for object duplication.
  + S3 Batch Operations for massive-scale object management.
* **Access Management and Security:**
  + S3 Block Public Access for control.
  + IAM for access control.
  + Bucket policies for permissions.
  + S3 Object Ownership for simplified access.
* **Data Processing:**
  + S3 Object Lambda for custom code execution.
  + Event notifications for workflow triggers.
* **Storage Logging and Monitoring:**
  + CloudWatch metrics for health.
  + CloudTrail for detailed API tracking.
  + Server access logging for request records.
  + Trusted Advisor for best practices.
* **Analytics and Insights:**
  + S3 Storage Lens for storage analysis.
  + Storage Class Analysis for access pattern analysis.
  + S3 Inventory for auditing and reporting.
* **Strong Consistency:**
  + Read-after-write consistency for PUT and DELETE.
  + Consistency for S3 Select, ACLs, Object Tags, and metadata.

**2.** **How Amazon S3 Works:**

* Object Storage Model:
  + Amazon S3 stores data as objects within buckets.
  + An object includes the file and associated metadata.
  + A bucket serves as a container for objects.
* Data Upload Process:
  + To store data, create a bucket with a unique name and specify an AWS Region.
  + Upload data as objects to the created bucket.
  + Each object has a unique key within the bucket.
* Object Configuration:
  + Configure features based on use case needs.
  + S3 Versioning maintains multiple versions of an object for recovery.
* Access Control:
  + Buckets and objects are private by default.
  + Access is granted explicitly through permissions.
  + Use bucket policies, IAM policies, ACLs, and S3 Access Points for access control.
* Additional Topics:
  + Buckets: Containers for objects.
  + Objects: Files and their associated metadata.
  + Keys: Unique identifiers for objects within buckets.
  + S3 Versioning: Keeps track of multiple versions of objects.
  + Version ID: Unique identifier for each version of an object.
  + Bucket Policy: Defines access rules for a bucket.
  + S3 Access Points: Named network endpoints for data access.
  + ACLs: Access control lists for granular permissions.
  + Regions: Specify AWS Region for bucket location.

**3. Amazon S3 Components Key Points:**

* **Buckets:**
  + Containers for objects in S3.
  + Up to 100 buckets/account.
  + Manage access with policies, ACLs, and Access Points.
* **Objects:**
  + Fundamental entities with data and metadata.
  + Identified by a unique key within a bucket.
* **Keys:**
  + Unique identifier for an object in a bucket.
* **S3 Versioning:**
  + Preserves, retrieves, and restores object versions.
  + Helps recover from unintended actions.
* **Version ID:**
  + Generated for each object with versioning.
  + Unique identifier for object versions.
* **Bucket Policy:**
  + IAM policy for access permissions.
  + Limited to 20 KB in size.
  + Uses JSON-based language.
* **S3 Access Points:**
  + Named endpoints for access policies.
  + Simplifies data access management.
  + Each has its access point policy.
* **ACLs (Access Control Lists):**
  + Grants read/write permissions to users.
  + Attached to buckets and objects.
* **Regions:**
  + Geographical location for bucket storage.
  + Choice based on latency, costs, or regulations.
  + Objects stay within the region unless explicitly moved.

**# STORAGE [EBS & INSTANCE STORE]**

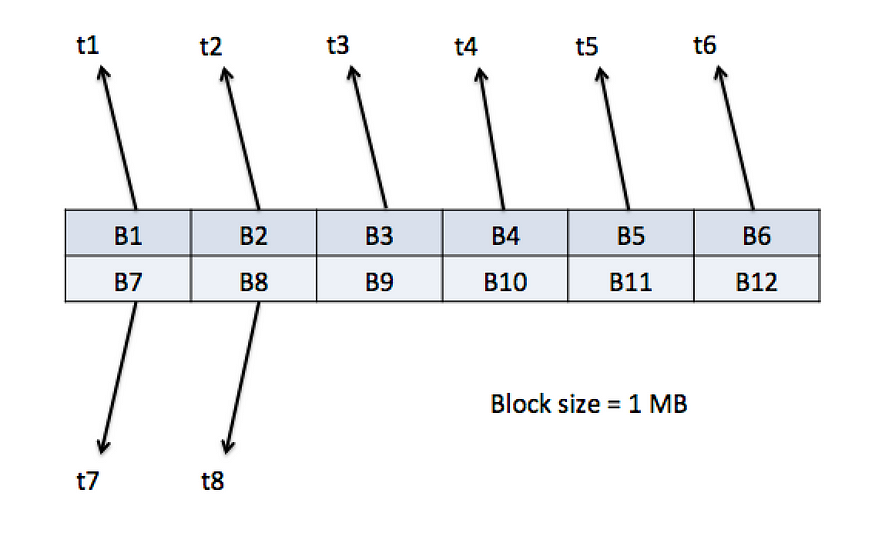
**NOTES:**

**What is an IO operation?**

IO operation stands for Input Output operation. It is a common unit of measurement for storage and system performance. Here, Input is a write operation and Output is a read operation that the disk can handle. Read or write of a single block is considered as an IO operation. Based on the above configuration, if an application wants to read say 5 records from the disk and if those 5 records span in 3 blocks, then the number of IO operations required to read 5 records is 3. Similarly, if each of the 5 records are in 5 different blocks, then the number of IO operations required are 5.

# Throughput

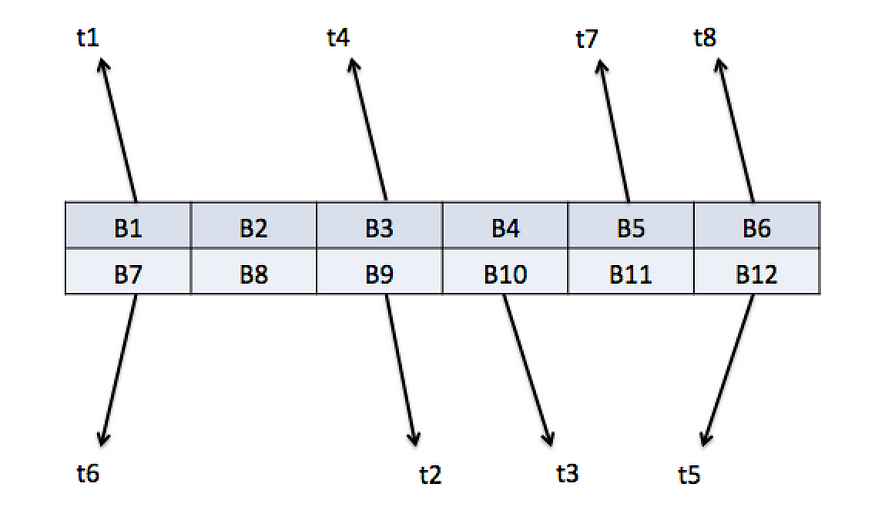
Throughput is the sequential transfer rate which the disk maintains continuously. Throughput deals with only sequential reads or writes. It is simply a measure of IO operations performed on sequential blocks one after the other on the disk. It is represented as total blocks size read/written in a second.



Above is a internal representation of a disk with each small box as a block numbered from B1 to B12. When the application reads 8 blocks in sequence one after the other say block1, block2, block3 …….. block8 in a second (each block is read at sub second interval t<n>) then throughput is measured as 8 MB/S (each block being 1 MB of size)HDDs are designed for sequential IO operations. They give better throughput compared to SSDs.

# IOPS

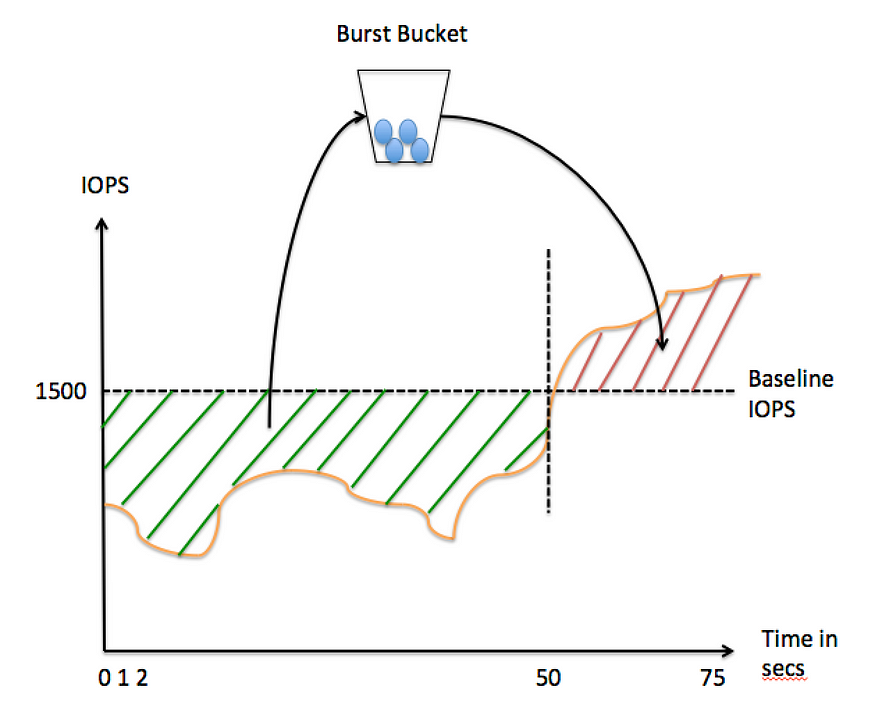
IOPS stands for Input Output operations Per Second. It is a integer value measure of number of IO operations that a disk can handle per second with each block being read from or written to a random location on the disk.



In the above representation, 8 blocks are read from random locations on the disk in a second making IOPS as 8.SSDs are better suited for random IO operations.

# Burst IOPS

In AWS EBS volume types, we often encounter something like **Burst performance IOPS**. AWS volumes provide baseline IOPS for each volume type and when the application do not consume all of these baseline performance IOPS, then the unused IOPS are stored in a burst bucket. These unused IOPS can be drained by the application when it needs more performance.



* **Amazon EBS Overview:**
  + Block level storage for EC2 instances.
* **Volume Behavior:**
  + Behaves like raw, unformatted block devices.
* **Mounting Volumes:**
  + Volumes can be mounted as devices on EC2 instances.
* **Persistence:**
  + Volumes persist independently of the instance's lifecycle.
* **File System and Usage:**
  + Supports file system creation on volumes.
  + Can be used similarly to a block device.
* **Configuration Changes:**
  + Configuration of attached volumes can be dynamically changed.
* **Recommended Use:**
  + Recommended for quickly accessible and persistently stored data.
* **Suitable Applications:**
  + Well-suited for primary storage in file systems, databases, and applications needing fine granular updates.
* **Access to Raw Storage:**
  + Useful for applications requiring access to raw, unformatted, block-level storage.
* **Versatility:**
  + Suited for both database-style applications with random reads and writes, and throughput-intensive applications with long, continuous reads and writes.

**Features of EBS**

* **Volume Creation:**
* Create EBS volumes in a specific Availability Zone and attach to an instance.
* **Snapshot and Restore:**
  + Use snapshots for cross-Availability Zone and cross-Region availability.
* **Volume Types:**
  + General Purpose SSD (gp2 and gp3)
  + Provisioned IOPS SSD (io1 and io2)
  + Throughput Optimized HDD (st1)
  + Cold HDD (sc1)
* **Performance and Use Cases:**
  + General Purpose SSD: Balanced for various workloads.
  + Provisioned IOPS SSD: Consistent IOPS for intensive workloads.
  + Throughput Optimized HDD: Ideal for large, sequential workloads.
  + Cold HDD: Cost-effective for sequential, cold-data workloads.
* **Encryption:**
  + Create encrypted volumes for data-at-rest security.
* **Snapshot Features:**
  + Point-in-time snapshots persisted to Amazon S3.
  + Serve as starting points for new volumes.
* **Cross-Region Snapshot Copy:**
  + Snapshots can be copied across AWS Regions.
* **Performance Monitoring:**
  + Use CloudWatch for metrics (bandwidth, throughput, latency) to optimize resource usage.

**Benefits of Using EBS Volume**

* **Data Availability:**
  + EBS volumes are replicated in the same Availability Zone to prevent data loss.
  + Attached volumes appear as native block devices.
* **Data Persistence:**
  + Volumes persist independently of the instance's lifecycle.
  + Can be detached with data intact for reattachment to new instances.
* **Data Encryption:**
  + Encrypted EBS volumes offer simplified data security.
  + Utilizes 256-bit AES encryption and AWS Key Management Service (AWS KMS).
* **Data Security:**
  + EBS volumes are presented as raw, unformatted block devices.
  + Supports secure data erasure procedures.
* **Snapshots:**
  + Snapshots are backups stored redundantly in Amazon S3.
  + Used for creating new volumes, moving volumes across Availability Zones, and incremental backups.
* **Flexibility:**
  + Supports live configuration changes in production.
  + Allows modification of volume type, size, and IOPS capacity without service interruptions.

**EBS Volume Types**

**Volume types**

* [Solid state drive (SSD) volumes](https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ebs-volume-types.html#vol-type-ssd)
* [Hard disk drive (HDD) volumes](https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ebs-volume-types.html#vol-type-hdd)
* [Previous generation volumes](https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ebs-volume-types.html#vol-type-prev) [magnetic]

**1.Previous-Gen EBS Volume (Magnetic):**

* Type: Standard
* Use Cases: Infrequently accessed data
* Size: 1 GiB - 1 TiB
* Max IOPS: 40 - 200
* Max Throughput: 40 - 90 MiB/s
* Boot Volume: Supported

1. **General Purpose SSD (gp2 and gp3) Volumes:**
   * **Durability:**
   * gp2: 99.8% - 99.9% (0.1% - 0.2% annual failure rate)
   * gp3: 99.8% - 99.9% (0.1% - 0.2% annual failure rate)

* **Use Cases:**
  + Transactional workloads, virtual desktops, medium-sized databases, low-latency apps, boot volumes, development, and test environments
  + **Volume Size:**
  + gp2: 4 GiB - 16 TiB
  + gp3: 1 GiB - 16 TiB
  + **Max IOPS per volume:**
  + gp2: 16,000 (16 KiB I/O)
  + gp3: 16,000 (64 KiB I/O)
  + **Max Throughput per volume:**
  + gp2: 250 MiB/s
  + gp3: 1,000 MiB/s
  + **Amazon EBS Multi-attach:**
  + Not supported for gp2 and gp3.
  + **NVMe Reservations:**
  + Supported for gp2, not supported for gp3.
  + NVMe (nonvolatile memory express) is a new storage access and transport protocol for flash and next-generation solid-state drives (SSDs) that delivers the highest throughput and fastest response times yet for all types of enterprise workloads.
  + **Boot Volume:**
  + Supported for both gp2 and gp3.

**3.Provisioned IOPS SSD (io1 and io2) Volumes:**

* + **Durability:**
  + io1: 99.8% - 99.9% (0.1% - 0.2% annual failure rate)
  + io2: 99.999% (0.001% annual failure rate)
  + **Use Cases:**
  + io1: Workloads requiring sustained IOPS performance or more than 16,000 IOPS, I/O-intensive database workloads.
  + io2: Workloads requiring sustained IOPS performance or more than 64,000 IOPS, I/O-intensive database workloads.
  + **Volume Size:**
  + io1: 4 GiB - 16 TiB
  + io2: 4 GiB - 64 TiB
  + **Max IOPS per volume:**
  + io1: 64,000 (16 KiB I/O)
  + io2: 256,000 (16 KiB I/O)
  + **Max Throughput per volume:**
  + io1: 1,000 MiB/s
  + io2: 4,000 MiB/s
  + **Amazon EBS Multi-attach:**
  + Supported for io1, not supported for io2.
  + **NVMe Reservations:**
  + Supported for io1, not supported for io2.
  + **Boot Volume:**
  + Not supported for both io1 and io2.

**NOTES:**

* + gp2 Volume Throughput:
  + Range: 128 MiB/s to 250 MiB/s (depends on volume size).
  + Pre-Dec 3, 2018: Older volumes may need modification for full performance.
* Max Throughput for Nitro Instances:
  + Achieve 1,000 MiB/s with:
  + 64,000 IOPS.
  + Attachment to Nitro System-built instances.
  + Pre-Dec 6, 2017: Older volumes may need modification for full performance.
* io2 Volume Evolution:
  + All io2 volumes after Nov 21, 2023, are io2 Block Express.
  + Conversion: Modify IOPS or size for io2 volumes before Nov 21, 2023.
* Volume Size Limits:
  + Nitro System Instances: Attach up to 64 TiB volumes.
  + Other Instances: Attach up to 16 TiB volumes.
* IOPS Limits:
  + Nitro System Instances: Attach up to 256,000 IOPS.
  + Other Instances: Attach up to 64,000 IOPS, achieving max 32,000 IOPS.

**4.Hard Disk Drive (HDD) Volumes:**

* **Volume Types:**
  + Throughput Optimized HDD (st1)
  + Cold HDD (sc1)
* **Durability:**
  + Both types have 99.8% - 99.9% durability with a 0.1% - 0.2% annual failure rate.
* **Use Cases:**
  + Throughput Optimized HDD (st1):
  + Big data, data warehouses, log processing.
  + Throughput-oriented storage for infrequently accessed data.
* **Cold HDD (sc1):**
  + Scenarios prioritizing low storage cost.
* **Volume Size:**
  + Both types support volumes ranging from 125 GiB to 16 TiB
  + Max IOPS per Volume (1 MiB I/O):
  + Throughput Optimized HDD (st1):500 IOPS
  + Cold HDD (sc1): 250 IOPS
* **Max Throughput per Volume:**
  + Throughput Optimized HDD (st1): 500 MiB/s
  + Cold HDD (sc1): 250 MiB/s
* **Amazon EBS Multi-attach:**
  + Not supported for both types.
* **Boot Volume:**
  + Not supported for both types.

# 5. General Purpose SSD volumes

# Volume Types:

# General Purpose SSD (gp3) volumes

# General Purpose SSD (gp2) volumes

# Backed by SSDs:

# General Purpose SSD (gp2 and gp3) volumes use solid-state drives.

# Balanced Price and Performance:

# Optimized for a variety of transactional workloads.

# Recommended Workloads:

# Virtual desktops, medium-sized single-instance databases, latency-sensitive interactive applications, development and test environments, and boot volumes.

# Amazon EBS Recommendation:

# Recommended for most workloads.

# 6. gp3 Volume Performance:

# - Burst Performance:

# Not applicable; gp3 volumes do not use burst performance.

# Sustain full provisioned IOPS and throughput indefinitely.

# - IOPS Performance:

# Consistent baseline of 3,000 IOPS included with storage cost.

# Additional IOPS can be provisioned, up to a max of 16,000, at a ratio of 500 IOPS per GiB of volume size.

# Maximum IOPS for volumes 32 GiB or larger.

# - Throughput Performance:

# Consistent baseline throughput of 125 MiB/s included with storage cost.

# Additional throughput can be provisioned, up to a max of 1,000 MiB/s, at a ratio of 0.25 MiB/s per provisioned IOPS.

# Maximum throughput for volumes 8 GiB or larger, provisioned at 4,000 IOPS or higher.

# - gp3 Volume Size:

# Range: 1 GiB to 16 TiB.

# 7. Migrate to gp3 from gp2:

# -Method: Use Amazon EBS Elastic Volumes for migration.

# -Operations: Modify type, IOPS, and throughput without instance interruption.

# Console Defaults:

# When creating volume or AMI from snapshot:

# Default is General Purpose SSD gp3.

# Otherwise:

# Default is gp2.

# Selection Option: Manually select gp3 when gp2 is default.

# 8. General Purpose SSD (gp2) Volumes:

# -Characteristics:

# Cost-effective storage for various transactional workloads.

# Performance scales with volume size.

# Tip:

# gp3 volumes are the latest generation with more predictable scaling and up to 20% lower prices than gp2.

# Estimate savings with the Amazon EBS gp2 to gp3 migration cost savings calculator.

# Performance Metrics:

# Single-digit millisecond latency.

# Volume durability: 99.8% to 99.9% with an AFR of no higher than 0.2%.

# Designed to deliver provisioned performance 99% of the time.

# 9. gp2 Volume Performance:

# IOPS Performance:

# Baseline IOPS scales linearly: 100 to 16,000 (3 IOPS per GiB).

# Provisioning:

# Volumes ≤ 33.33 GiB: 100 IOPS.

# - Volumes > 33.33 GiB and ≤ 5,334 GiB: 3 IOPS per GiB (up to 16,000 IOPS).

# Volumes > 5,334 GiB: 16,000 IOPS.

# Burst Capability (for volumes < 1 TiB and < 3,000 IOPS):

# Bursts to 3,000 IOPS, governed by I/O credits.

# Burst Duration Formula:

* + - * + (I/O credit balance)
    - Burst duration = -----------------------------
      * + (Burst IOPS) - (Baseline IOPS)

# I/O credit accrual at 3 credits per GiB per second when demand drops.

# I/O credit accrual limit: 5.4 million, sustaining max burst for at least 30 minutes.

**10. gp2 Volume Characteristics:**

* **-Burst Capability:**
* Burst duration and refill time based on volume sizes, baseline performance, and burst duration at 3,000 IOPS.
* Example sizes: 1 to 33.33 GiB, 334 GiB, 750 GiB, 1,000 GiB, 5,334 GiB, and larger.
* Monitor I/O credit balance with "Amazon EBS BurstBalance" metric in CloudWatch.
* **Throughput Performance:**
* Varies between 128 MiB/s and 250 MiB/s based on volume size.
* Up to 128 MiB/s for volumes ≤ 170 GiB.
* Up to 250 MiB/s for volumes > 170 GiB but < 334 GiB.
* 250 MiB/s for volumes ≥ 334 GiB.
* **Throughput Calculation Formula:**
* Throughput in MiB/s = IOPS performance × I/O size in KiB / 1,024
* **-Volume Size:** Ranges from 1 GiB to 16 TiB, with linear performance scaling with volume size.

**11. Provisioned IOPS SSD Volumes:**

**Types:**

* io2 Block Express volumes
* io1 volumes

**io2 Block Express Volumes:**

* Designed for demanding I/O applications.
* Uses Block Express for durability, low latency, high performance.
* Communicates with Nitro System via SRD protocol.
* 99.999% durability, sub-ms latency, high IOPS, throughput.
* Supports Nitro instances, specific regions, conversion from io2.
* Up to 64 TiB, 256,000 IOPS, 4,000 MiB/s throughput.

**Performance:**

* Sub-ms latency, 64 TiB storage.
* Up to 256,000 IOPS, 4,000 MiB/s throughput.
* Achievable on Nitro; up to 32,000 IOPS on other instances.
* Throughput scales with 0.256 MiB/s per IOPS.
* Max throughput at 16,000 IOPS or higher.

**Considerations:**

* Convert io2 volumes to io2 Block Express.
* Attachment limits based on instance type.
* Creation of encrypted volumes >16 TiB, IOPS >64,000 requires specific steps.

**NOTES:**

* The Nitro System provides enhanced security that continuously monitors, protects, and verifies the instance hardware and firmware. Virtualization resources are offloaded to dedicated hardware and software minimizing the attack surface.
* A Xen hypervisor allows multiple instances to share a single hardware platform. A hardware on which a hypervisor is running is called a Host machine and the virtual machine is called a Guest machine.