Table of Contents

[Cloud computing 1](#_Toc33274002)

[Infrastructure 1](#_Toc33274003)

[Pricing 1](#_Toc33274004)

[Support 3](#_Toc33274005)

[Compute 3](#_Toc33274006)

[Networking 5](#_Toc33274007)

[Deploying 5](#_Toc33274008)

[Load balancing/scaling 5](#_Toc33274009)

[Basic Storage 7](#_Toc33274010)

[Advanced Storage/Data 7](#_Toc33274011)

[“Simple” services 9](#_Toc33274012)

# 

# Cloud computing

# Infrastructure

|  |  |
| --- | --- |
| Region | * Physical location/geographic area with 2+ availability zones. * Minimize latency by deploying to 2+ regions |
| AZ (Availability Zone) | * Physically/logically isolated data centers * Data provisioned across AZs * Not all zones offer all services |
| Data Center | * 1+ per AZ |
| Edge Location | * Host Cloudfront (Content delivery network) for faster delivery of static content with low latency/high transfer speeds * More edge locations than AZs * Caches data |

# Pricing

|  |  |
| --- | --- |
| Basics | * Usually no charge for inbound data or data within AWS region * Pay for CPU, data storage, outbound data transfer * The more you use, the less it costs |
| On demand | * Pay as you go * Most services pay per second of use * Good for short term, spiky or unpredictable use |
| Reservations | * Up to 75% less * 1-3 year commitment * Pay none/partial/all up front * Costs less if pay more up front * Good for steady state usage |
| Spot | * Up to 90% less * Pay for unused capacity * Unpredictable when runs * Ends when complete or price goes above bid |
| Dedicated instance | * Pay set hourly price * Dedicated hardware for VPC * Can use existing software licenses |
| Free tier | * Some services free forever – VPC, Elastic Beanstalk, , CloudFormation, IAM, , Autoscaling, Opsworks, DynamoDB, Glacier, Lambda, Glue, Cognito, SNS, SES, SQS, SWF, Cloudwatch, Xray, Storage Gateway, etc * Some services free 12 months – EC2, S3, RDS, CloudFront |

# Support

|  |  |
| --- | --- |
| Basic | * 7 trusted advisor checks, personal health dashboard, docs/support forms |
| Developer | * Basic + email support * 1 contact * Response time 24 hours for general, 12 hours for impaired system |
| Business | * Developer + full trusted advisor checks, phone support * Unlimited contacts * Response time 1 hour for prod down |
| Enterprise | * Business + senior cloud support engineers * Response time 15 minutes for business critical systems * Includes Well Architected Review by AWS Solution Architects, self packed labs, concierge support team, dedicated technical account   manager |
| Support forms for | * Encountering Abuse (sent to Abuse team) * Increasing limits beyond a point * Penetration testing |
| Acceptable Use Policy | * What you’d expect; don’t do bad things |

# Compute

|  |  |
| --- | --- |
| EC2 (Elastic Compute Cloud) | * Virtual server * Proper name is EC2 instances * Pay as you go. Pay for time running * Maintain control * Don’t have to provision/maintain server * Assigned both public/private IP * Has instance metadata * Responsible for patching OS |
| VPC (Virtual Private Cloud) | * Isolate compute resources * Control network config, access, what expose, etc * Can span AZs |
| ECS (Elastic Container  Service) | * Supports Docker containers |
| AMI | * Amazon Machine Image * Can use variety of preconfigured ones or create own * Specifies type of hardware * Bootable |
| Lambda | * Serverless * Pay only for compute by fraction of millisecond |

|  |  |
| --- | --- |
|  | * Ideal for variable/intermittent workloads * Auto-scales * Supports many programming languages * Limited disk space/memory * Must run less than 5 minutes |

# Networking

|  |  |
| --- | --- |
| IGW (Internet Gateway) | * Allows access to internet from VPC |
| Subnet | * Divides VPC * Public subnets can access internet * Private subnets cannot (by default) * VPC can have multiple subnets |
| Route tables | * Register traffic leaving subnet |
| NAT Gateway | * Allows private subnet to access internet |
| CIDR (classless  interdomain routing) | * Internal IP address look like 10.0.0.0/16 |
| Direct Connect | * On premises to VPC connectivity or VPC to VPC connectivity |
| PrivateLink | * Connects to VPCs through endpoints |
| VPC Peering | * Connect to VPCs privately |
| Route 53 | * DNS * Geolocation routing * Latency based routing * Defaults to up to 50 domain names * Global service |
| Elastic IP | * Static IPv4 address * Up to 5 per region * Pay if have more than one and not associated with running instance |

# Deploying

|  |  |
| --- | --- |
| Elastic Beanstalk | * PaaS application server * Supplies all infrastructure so can just deploy app |
| CloudFormation | * Manage/provision collections of servers |

# Load balancing/scaling

|  |  |
| --- | --- |
| Application Load Balancer | * HTTP/HTTPS level * Includes HTTPs and WebSockets * Can route by path or hosts |
| Network Load Balancer | * TCP level |
| ELB (Elastic Load Balancer) – classic load balancer | * Older loader balancer * Supports both HTTP/TCP levels * Can mix with internal load balancers * Supports single region |

|  |  |
| --- | --- |
| Auto Scaling | * Adds more EC2 instances as needed * Specify conditions/policy for when add/remove instances * Create launch config (what create if need new instance), group (constraints on what create) and policy (when to scale) * Limit to 20 EC2 instances per region |
| Listener | * Checks for connection requests to load balancer |
| Target | * Destination for traffic based on rules |
| Target groups | * 1+ targets * Target can be in multiple groups * Can do health check by target group |

# Basic Storage

|  |  |
| --- | --- |
| S3 (Simple Storage Service) | * Object data up to 5TB * Can access by URL * API to get data; not associated with specific server * Can access via HTTP/HTTPS * Objects grouped into S3 buckets. Can have up to   100. Can set policies on buckets.   * Can replicate across regions * Durability is always 11 nines. Means probability of losing an object. * Availability is 4 nines for standard and 3 nines for SIA (standard infrequent access) |
| EBS (Elastic Block Store) | * Block storage * Storage for EC2 * Persistent data * General Purpose (SSD), Provisioned IOPS (SSD), magnetic * Automatically replicated within AZ. Can copy to other region for recovery * Snapshots are backups |
| EFS (Elastic File System) | * File storage for EC2 |

# Advanced Storage/Data

|  |  |
| --- | --- |
| Aurora | * Managed database service * 5x faster than MySQL/Postgres * Faster version of MySQL * Defaults to replicating twice in each of 3 AZs |
| RDS (Relational Database Service) | * Supports Aurora, MySQL, PostgresSQL, Oracle, MS SQL Server and MariaDB * Set up own IP, subnet, access control, etc |

|  |  |
| --- | --- |
|  | * Automatically generates standby database in another AZ * Can create read replicas in different region for all but Oracle and MS SQL Server |
| DynamoDB | * Managed NoSQL service * Access by query (key) or scan (non-key attribute) |
| RedShift | * Managed data warehouse service * Uses SQL * Supports petabytes of data * OLAP |
| Snowball Edge | * Physically transport 100TB of data |
| Snowball | * Physically transport petabytes of data |
| Snowmobile | * Physically transport up to 100 petabytes of data |
| Glacier | * Data archiving * Each archive up to 40TB * Infrequent access * Data encrypted by default * Archive – document stored * Vault – container for storing archives. Has access policy and lock policy (can’t alter when locked) * Data comes from S3 (via lifecycle policies), SDK, CLI or snowball/snowmobile import * Takes minutes or hours to retrieve data depending on cost Bulk/Standard/Expedited |
| Transfer Acceleration | * Transfer files over the internet across long distances with S3 bucket |
| DMS (Data Migration Service) | * Migrate non-AWS database to cloud |
| EMR (Elastic map reduce) | * Hadoop |
| Glue | * ETL (extract load transform) |
| Storage Gateway | * Links to on premises data environment |
| Athena | * Serverless queries |
| Kinesis | * Streaming data |
| Kinesis Firehose | * Data load |
| Neptune | * Graph database |

# “Simple” services

|  |  |
| --- | --- |
| SES (Simple email  service) | * Email |
| SNS (Simple Notification Service) | * Publish messages * Supports HTTP/S, Email, Email JSON, SMS, SQS |
| SQS (Simple Queue Service) | * Hosted queue * Visible for 12 hours by default |

|  |  |
| --- | --- |
| SWF (Simple Workflow) | * Workflow * Activity worker implements a task |

Security

|  |  |
| --- | --- |
| NACL (network access control list) | * Stateless * Like passport control * Checks access each time on entry/exit * Optional * At subnet level |
| Security Groups | * Built in firewall for virtual servers * Set up rules * Can control by protocol/port/IP * By default, controls inbound (blocks all) and outbound traffic (allows all) |
| Shield | * Protects against DDoS (distributed denial of service) * Free level built into EC 2 * Two levels * Advanced level requires Business plan or higher |
| WAF (Web Application Firewall) | * Blocks common attacks (ex: XSS) * Global service |
| Shared responsibility model | * Amazon – “of the cloud” * Customer – “in the cloud” |
| Guard Duty | * Threat detection |

IAM

|  |  |
| --- | --- |
| IAM (Identity and Access Management) | * Control access * Can’t recover lost credentials * Allows each user up to two active keys * Global service |
| Identities | * People/processes/services * Unit of authentication |
| Groups | * Collections of users |
| Root user | * Initial user created * Unrestricted access * Only use to create initial other users * Required to use CLI * Recommended to delete access keys |
| Role | * Identity with permission policies * Does not have own credentials * Used for apps * Used for SSO where authenticated at company |
| Temporary credentials | * Credentials with restricted permission for a specific task |

|  |  |
| --- | --- |
| Policy | * Applied to user/role/group to grant permissions |
| Access types | * Programmatic access * Management console access |

Monitoring

|  |  |
| --- | --- |
| TCO (Total Cost of Ownership) Calculator | * Determine costs before using * Don’t need to be AWS customer yet * Compares on-prem and collocation to pure AWS |
| Trusted Advisor | * Check security, fault tolerance, performance, cost savings. * For existing customers * Red (immediate action), yellow (investigate), green (good) * Can get notification when checks fail * Focuses on services |
| Cost Explorer | * Billing visibility for current customers * Can see last 13 months of data * Forecasts costs for next three months |
| Budgets | * Alerts when costs exceed plan |
| Cost and Usage Report | * Shows costs by category |
| CloudTrail | * Records user activity/API calls |
| CloudWatch | * Monitoring logs * Aggregates logs * Can set billing alarm * Basic and Detailed plans * Defaults to 5 minute granularity for basic and 1 minute for detailed |
| Inspector | * Find possible security issues * Focuses on S3 level * Automated compliance |
| Artifact | * View compliance reports |
| Migration Hub | * Track progress of migrations across AWS and partners |

For Programmers and Dev/Ops

|  |  |
| --- | --- |
| AWS SDKs | * APIs |
| OpsWorks | * DevOps platform * Uses Chef |
| CodeStar | * UI for Development |
| CodeCommit | * Version control |
| CodeDeploy | * Automated deployment |
| CodePipeline | * Continuous Delivery |

Pillars of Architecture

|  |  |
| --- | --- |
| Operational Excellence | * Operations as code * Annotate documentation * Make frequent, small, reversible changes * Refine operations procedures frequently * Anticipate failure * Learn from operational failures |
| Security | * Implement a strong security foundation * Enable traceability * Apply security at all layers * Automate security best practices * Protect data in transit and at rest * Prepare for security events |
| Reliability | * Test recovery procedures * Automatically recover from failure * Scale horizontally to increase aggregate system availability   Stop guessing capacity   * Manage change in automation |
| Performance Efficiency | * Democratize advanced technologies * Go global in minutes * Use serverless architectures * Experiment more often * Mechanical sympathy |
| Cost Optimization | * Adopt a consumption model * Measure overall efficiency * Stop spending money on data center operations * Analyze and attribute expenditure * Use managed services to reduce cost of ownership |

Recovery

|  |  |
| --- | --- |
| Pilot Light | * Quick recovery option> Minimal version always running |
| Slowest to fastest | * Backup & Restore * Pilot Light * Warm Standby * Multi Site |
| Fault tolerance | * Stays up even if parts fail * More strict than High Availability |

Random other services

|  |  |
| --- | --- |
| CloudFront | * CDN (content delivery network) * Can act as a cache to serve objects from S3 |