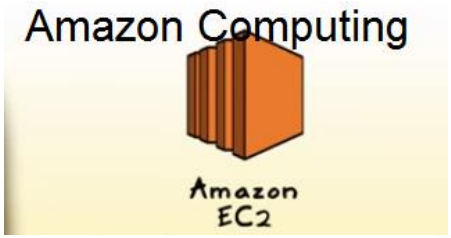
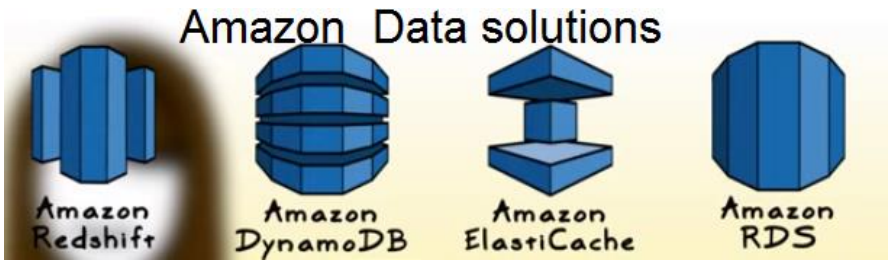
AWS BASIC INFRASTRUCTURE

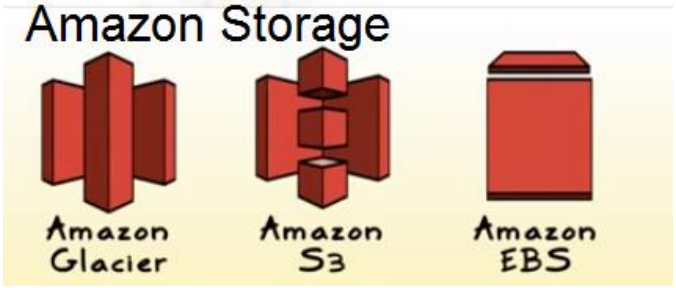
Dr. Khurram S. Khattak

Amazon Web Services

- AWS is a collection of remote computing services
  - Elastic Compute Cloud (EC2) provides scalable virtual private servers
  - Simple Storage Service (S3) provides Web Service based storage.
  - SimpleDB allows developers to run queries on structured data. It provide "the core functionality of a database."
  - Elastic MapReduce allows developers to easily and cheaply process vast amounts of data. It uses a hosted Hadoop
  - Virtual Private Cloud (VPC) creates a logically isolated set of Amazon EC2 instances which can be connected to an existing network using a VPN
  - And More…







AWS EC2

Amazon's EC2

- Amazon Elastic Compute Cloud (EC2)
  - Web service that provides resizable compute capacity in the cloud
- An EC2 instance appears physical HW, provides users complete control over nearly entire sw stack, from the kernel upwards
  - Load Variety of operating system
  - Install Custom applications
  - Manage network access permission
  - Run image using as many/few systems as you desire

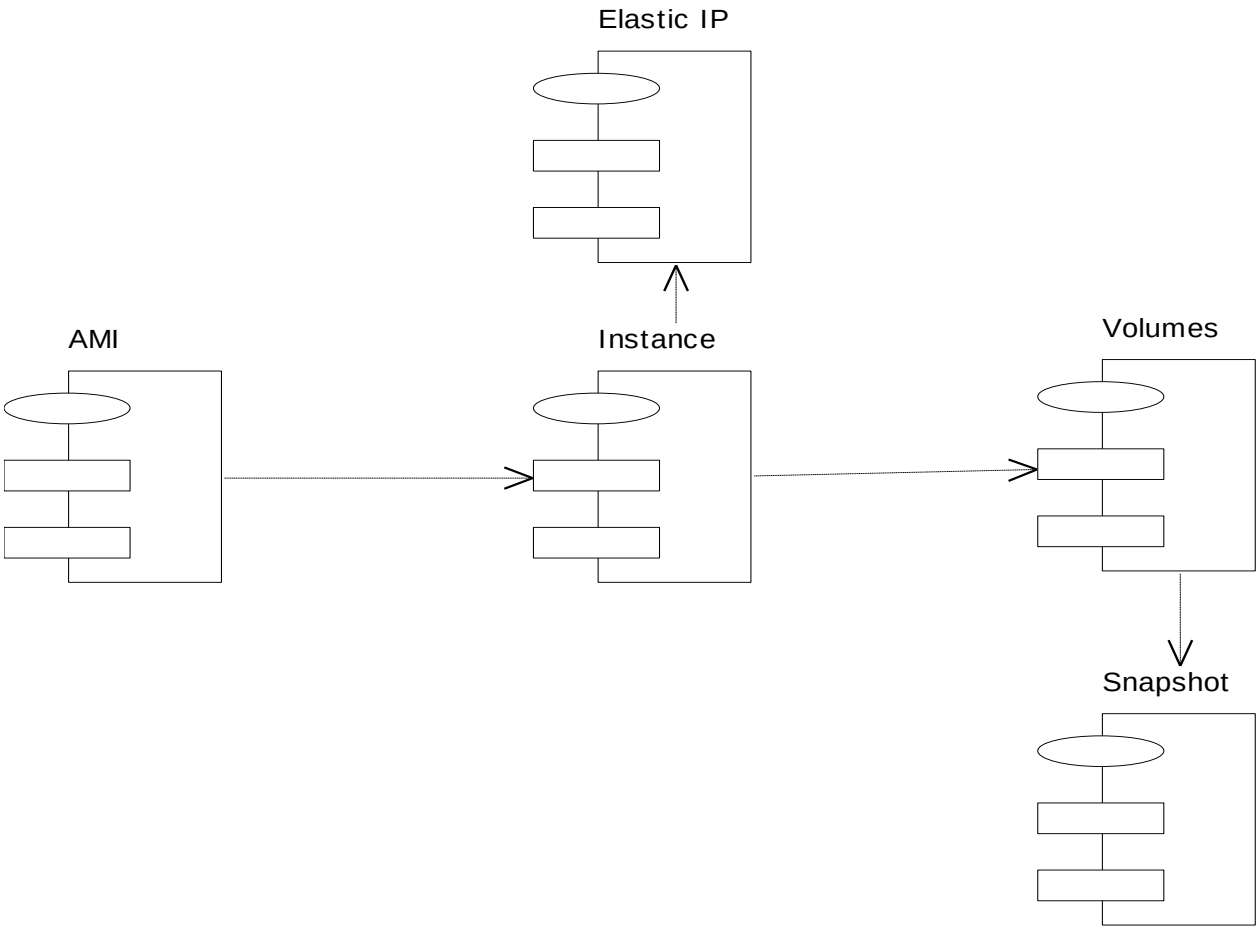
Amazon's EC2 features

- Elastic capacity
  - Elastic resource config/reconfig; Elastic num of instances
- Completely Control
  - Root access/access to console output/data store/ reboot
- Reliable
  - Multiple locations
  - Elastic IP addresses
- Secure
  - Firewall config
  - Virtual Private Cloud
- Performance
  - Auto Scaling
  - Auto local balancing

Amazon's EC2 Instances

- On-Demand Instances
  - Pay for capacity without long-term commitment
- Reserved Instances
  - Standard Instances
  - Micro Instances
  - High-Memory Instances
  - High-CPU Instances
  - High-I/O instance
  - High Storage Instances
- Spot Instances
  - Bid on unused Amazon EC2 capacity, run those instances for as long as their bid exceeds the current Spot Prices

EC2 Components



Amazon Machine Images (AMI)

- **Public AMIs:** Use pre-configured, template AMIs to get up and running immediately. Choose from Fedora, Movable Type, Ubuntu configurations, and more
- **Private AMIs:** Create an Amazon Machine Image (AMI) containing your applications, libraries, data and associated configuration settings
- **Paid AMIs:** Set a price for your AMI and let others purchase and use it (Single payment and/or per hour)
  - AMIs with commercial DBMS

AMI's with Special Softwares

- IBM DB2, Informix Dynamic Server, Lotus Web Content Management, WebSphere Portal Server
- MS SQL Server, IIS/Asp.Net
- Hadoop
- Open MPI
- Apache web server
- MySQL
- Oracale 11g
- …

Elastic Block Store (EBS)

- Provides persistent block level storage volumes for use with *EC2* instances; suitable for database applications, file systems, and applications using raw data devices.
- A volume appears to an application as a raw, unformatted and reliable physical disk; the range 1 GB -1 TB.
- Each user can use up to 5000 EBS volumes
- An *EC2* instance may mount multiple volumes, but a volume cannot be shared among multiple instances.
- EBS supports the creation of snapshots of the volumes attached to an instance and then uses them to restart the instance.
- EBS can be used as boot partition for instances: fast startup time
- The volumes are grouped together in Availability Zones and are automatically replicated in each zone.

Growing the EBS

- This AMI has a drive size of 8 GB
- It can be “grown”
- Take a snapshot, launch a new EBS instance using the snapshot, and

Amazon's EC2 Operating

- STEP 1 Create an Amazon Machine Image (AMI) containing your applications, libraries, data and associated configuration settings. Or use pre-configured, templated images to get up and running immediately.
- STEP 2 Choose the types of instances and OS, then start, terminate, and monitor as many instances of your AMI as needed, using the web service APIs or the variety of management tools provided.
- STEP 3 Determine whether you want to run in multiple locations, utilize static IP endpoints, or attach persistent block storage to your instances.
- NOTE: Pay only for the resources that you actually consume, like instance-hours or data transfer.

Pricing (2017)

Region: US East (N. Virginia)

	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage
General Purpose - Current Generation					
t2.nano	1	Variable	0.5	EBS Only	$0.0065 per Hour
t2.micro	1	Variable	1	EBS Only	$0.013 per Hour
t2.small	1	Variable	2	EBS Only	$0.026 per Hour
t2.medium	2	Variable	4	EBS Only	$0.052 per Hour
t2.large	2	Variable	8	EBS Only	$0.104 per Hour
m4.large	2	6.5	8	EBS Only	$0.12 per Hour
m4.xlarge	4	13	16	EBS Only	$0.239 per Hour
m4.2xlarge	8	26	32	EBS Only	$0.479 per Hour
m4.4xlarge	16	53.5	64	EBS Only	$0.958 per Hour
m4.10xlarge	40	124.5	160	EBS Only	$2.394 per Hour
m4.16xlarge	64	188	256	EBS Only	$3.83 per Hour
m3.medium	1	3	3.75	1 x 4 SSD	$0.067 per Hour
m3.large	2	6.5	7.5	1 x 32 SSD	$0.133 per Hour
m3.xlarge	4	13	15	2 x 40 SSD	$0.266 per Hour
m3.2xlarge	8	26	30	2 x 80 SSD	$0.532 per Hour

AWS EC2 Features

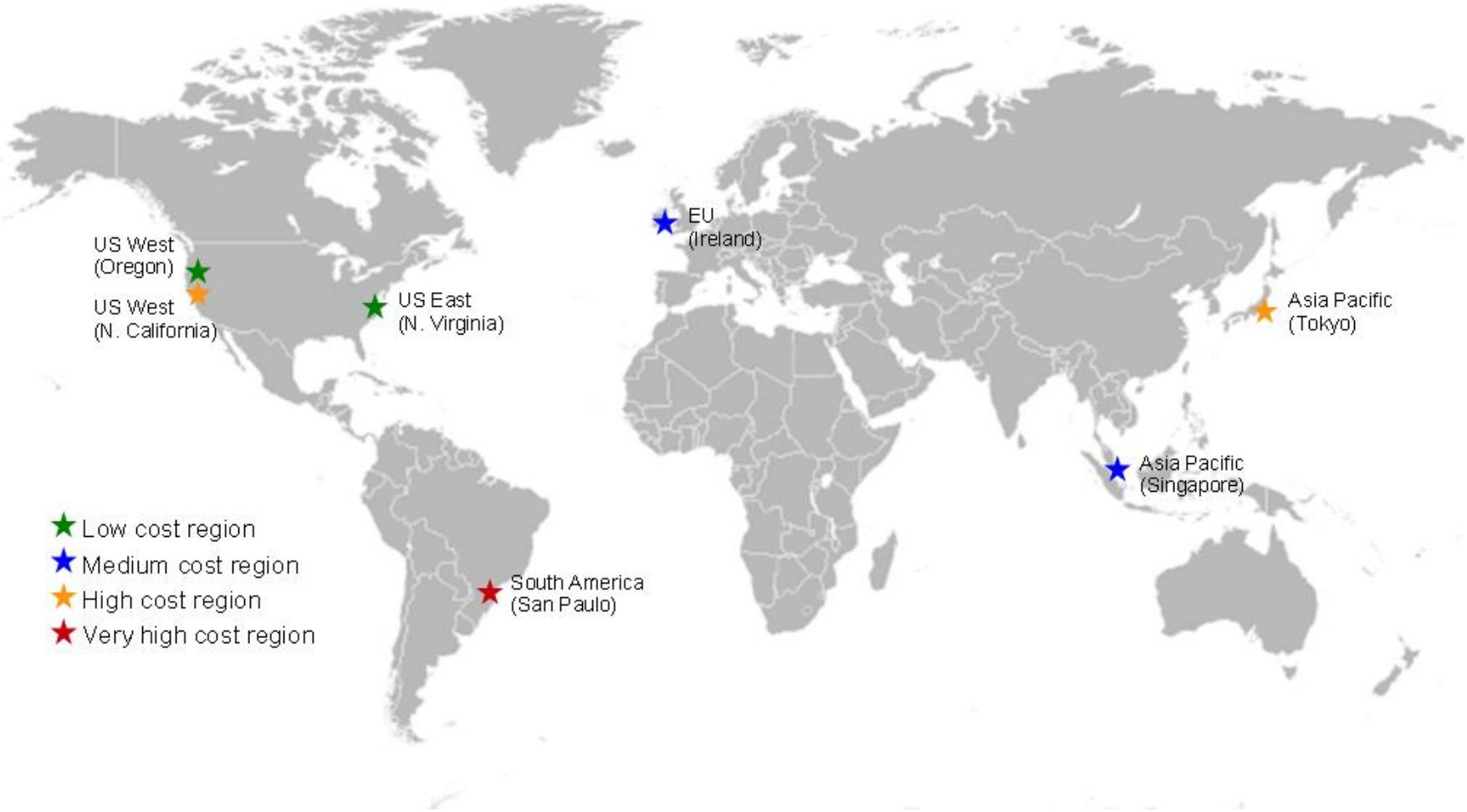
AWS regions and availability zones

- Amazon offers cloud services through a network of data centers on several continents.
- In each *region* there are several availability zones interconnected by high-speed networks.
- An *availability zone* is a data center consisting of a large number of servers.

Region	Location	Availability zones	Cost
US West	Oregon	us-west-2a/2b/2c	Low
US West	North California	us-west-1a/1b/1c	High
US East	North Virginia	us-east-1a/2a/3a/4a	Low
Europe	Ireland	eu-west-1a/1b/1c	Medium
South America	Sao Paulo, Brazil	sa-east-1a/1b	Very high
Asia Pacific	Tokyo, Japan	ap-northeast-1a/1b	High
Asia Pacific	Singapore	ap-southeast-1a/1b	Medium

- Regions do not share resources and communicate through the Internet.

AWS regions and availability zones



Evolving AWS Worldwide Infrastructure

amazon web services™

Static IP

- By default when you launch a new instance Amazon dynamically assign a private and a public IP.
- While this is fine for development purposed, for a real launch of a web accessible service, we need static IP.
- Amazon makes available what are classes elastic IPs for this purpose. Up to 5 elastic IPs can be assigned to an instance.
  Elastic IPs cost money even if you don't use them; assigning and reassigning strains the system; so it cost money
- Allocate elastic ip and associate it with an instance.

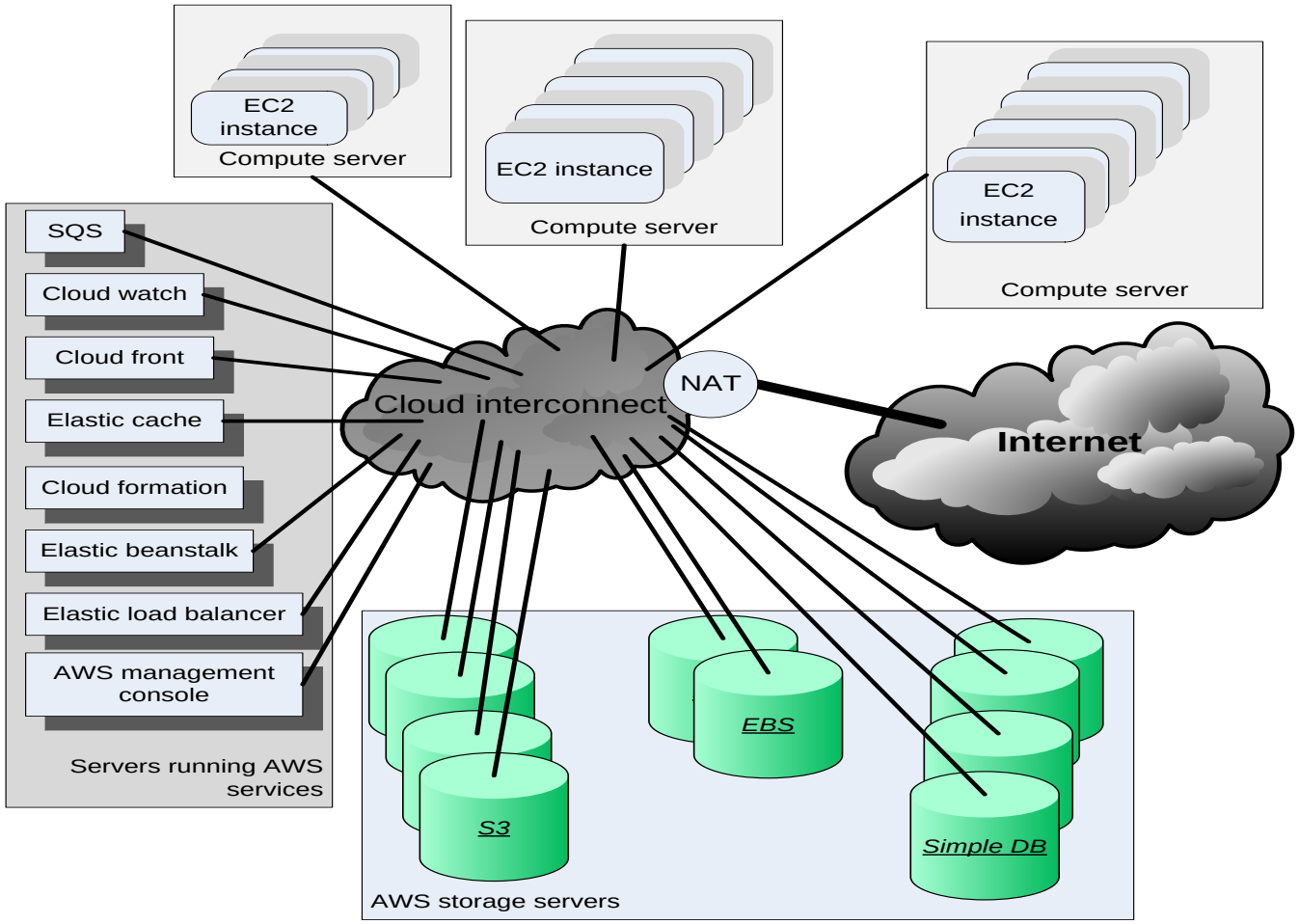
Security Group

- You want to create an “instance” of a server from an already established “AMI: Amazon Machine Image”
- It has a elastic IP for the whole world to interact with it.
- How to control access to this?
- While creating the instance, create a new security group that will specify the policy or “rules” about the access methods
- Security group is somewhat similar to network segment protected by a firewall
- Once the server is started you cannot change the security group: so plan ahead

Security Group

- Network Address Translation (NAT) maps external IP addresses to internal ones.
- The public IP address is assigned for the lifetime of an instance.
- An instance can request an *elastic IP address*, rather than a public IP address. The elastic IP address is a static public IP address allocated to an instance from the available pool of the availability zone.
- An elastic IP address is not released when the instance is stopped or terminated and must be released when no longer needed.

Security Group



Snapshot

- Snapshot is for saving a volume (of storage) is a feature of Amazon's elastic block storage.
- You can take a snapshot as often as needed.
- EC2 automatically saves the snapshots to S3, thus enabling a quick and powerful backup scheme.
- You can replay it by creating a volume from snapshot. See demo.

AWS Storage

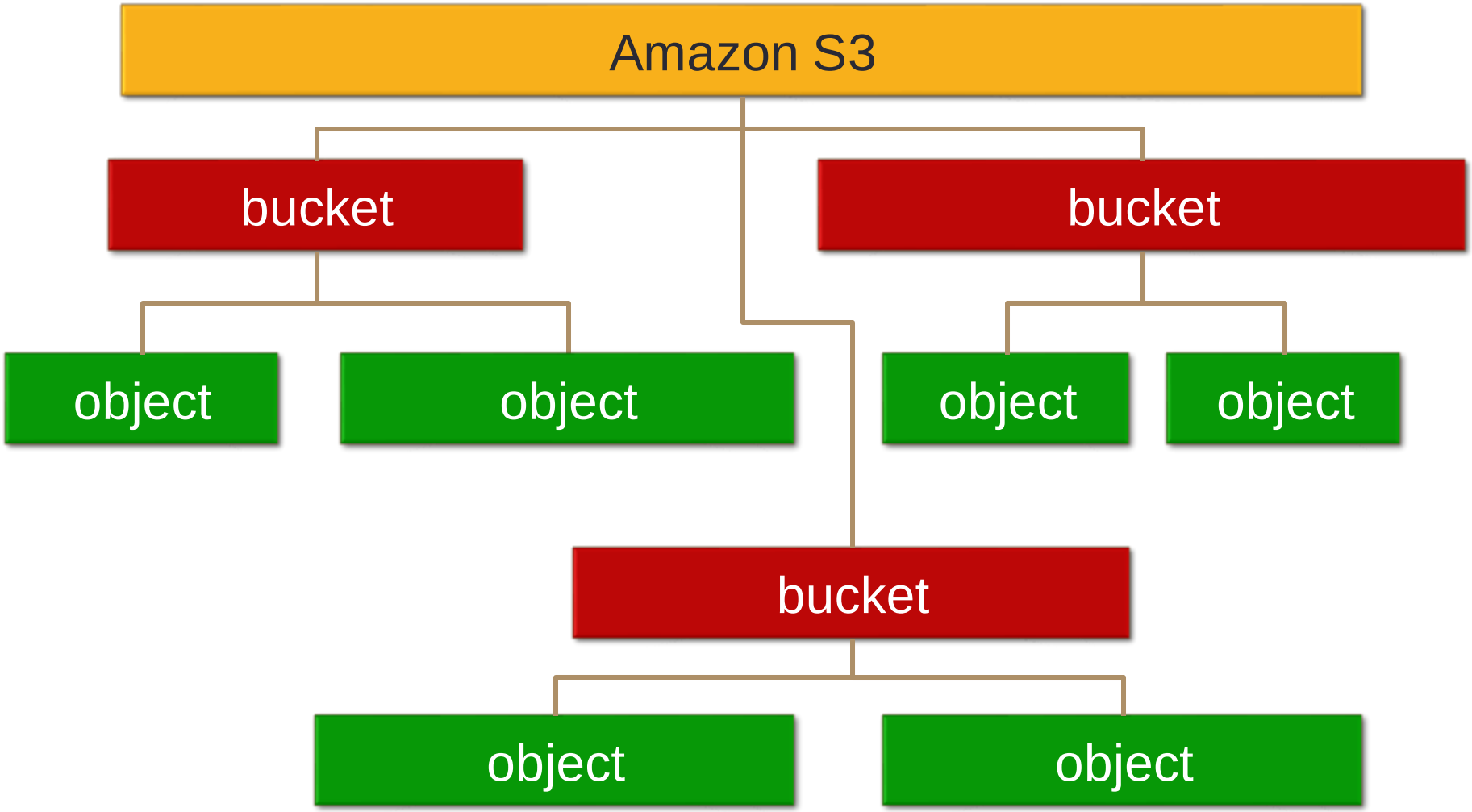
AWS: Storage

- *Simple Storage Service (S3):* provide persistent storage
  - Independent of EC2 instances
  - EC2 instances need to “download” data from S3 in order to access it (cannot issue read/write to S3)
- Amazon Glacier: low-cost storage service that provides secure and durable storage for data archiving and backup
  - Advantage over S3: offload the administrative burdens of operating and scaling storage + cost
  - Disadvantage: slower than S3
- Storage Gateway: securely store data to the AWS cloud for scalable and cost-effective storage
  - All data is securely transferred to AWS over SSL and stored encrypted in Amazon S3 using AES 256-bit encryption
- *Elastic Block Store (EBS):* provide block level storage volumes (virtual disk, i.e., disk-like) to EC2 instances
  - Persistent even after instances are terminated
  - Instances have to mount EBSs (EFS)

Amazon's S3

- Amazon Simple Storage Service (S3)
  - Storage for the Internet.
- Features
  - Unlimited Storage
  - Highly scalable
    - in terms of storage, request rate and concurrent users
  - Reliable
    - Store redundant data in multiple facilities and on multiple devices
  - Secure
    - Flexibility to control who/how/when/where to access the data
  - Performance
    - Choose region to optimize for latency/minimize costs
- Work with other AWS products
  - EC2/Elastic MR/Amazon Import/Export…

S3 namespace



AWS S3 (Accessing Objects)

- Bucket: keke-images, key: jpg1, object: a jpg image
  accessible with
  https://**keke-images**.s3.amazonaws.com/**jpg1**

- mapping your subdomain to S3
  with DNS CNAME configuration
  e.g. media.yourdomain.com → media.yourdomain.com.s3.amazonaws.com/

AWS S3 (Access Control)

- Access log
- Objects are private to the user account
  - Authentication
- Authorization
  - ACL: AWS users, users identified by email, any user …
- Digital signature to ensure integrity
- Encrypted access: https

Amazon's S3

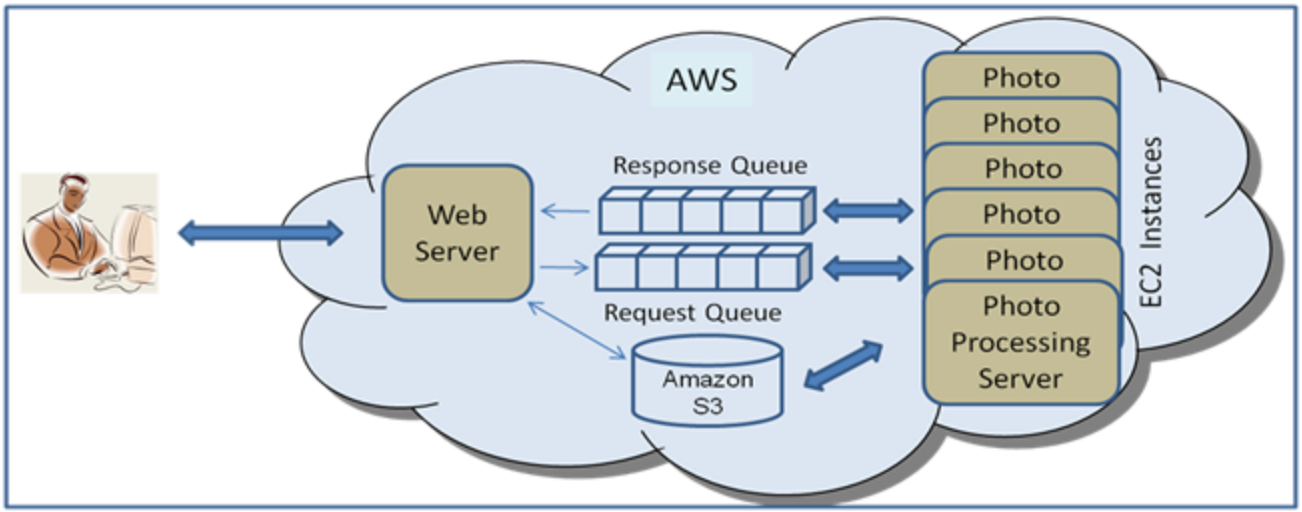
- Service designed to store large objects; an application can handle an unlimited number of objects ranging in size from 1 byte to 5 TB.
- An object is stored in a bucket and retrieved via a unique, developer-assigned key; a bucket can be stored in a Region selected by the user.
- Supports a minimal set of functions: write, read, and delete; it does not support primitives to copy, to rename, or to move an object from one bucket to another.
- The object names are global.
- S3 maintains for each object: the name, modification time, an access control list, and up to 4 KB of user-defined metadata.

Amazon's S3

- Authentication mechanisms ensure that data is kept secure.
- Objects can be made public, and rights can be granted to other users.
- *S3* computes the MD5 of every object written and returns it in a field called ETag.
- A user is expected to compute the MD5 of an object stored or written and compare this with the ETag; if the two values do not match, then the object was corrupted during transmission or storage.

Example : online photo processing service

- Photo operation
  - red eye reduction/cropping/customization/re-coloring/teeth whitening, etc
- Procedure
  - Web server receive request
  - Put request message in the queue
  - Pictures stored in S3
  - Multiple EC2 instances run photo processing
  - Put back in the queue
  - Return



MORE ABOUT AWS

Auto Scaling

- Auto Scaling helps maintain application availability and allows users to scale an Amazon EC2 capacity up or down automatically according to defined conditions.

- To ensure that you are running your desired number of Amazon EC2 instances.

- Auto Scaling can also automatically increase the number of Amazon EC2 instances during demand spikes to maintain performance and decrease capacity during low demands to reduce costs.

Amazon Data solutions

- **RedShift (PostgreSQL) - setup to be able to work with larger data than RDS, optimized over RDS.**
- **Relational Database Service(RDS) -- MySQL, Oracl, SQL Server, PostgreSQL**
- **DynamoDB – non traditional data base operations (and less scalable "older" Amazon SimpleDB)**
- **ElastiCache – cache solution**
- **CloudFront – content delivery network**
- **Elastic Block Store - backing up instances in an cloud infrastructure as a service for recovery from failure**

SimpleDB

- Non-relational data store. Supports store and query functions traditionally provided only by relational databases.
- Supports high performance Web applications; users can store and query data items via Web services requests.
- Creates multiple geographically distributed copies of each data item.
- It manages automatically:
  - The infrastructure provisioning.
  - Hardware and software maintenance.
  - Replication and indexing of data items.
  - Performance tuning.

SQS - Simple Queue Service

- Hosted message queues are accessed through standard SOAP and Query interfaces.
- Supports automated workflows - *EC2* instances can coordinate by sending and receiving SQS messages.
- Applications using SQS can run independently and asynchronously, and do not need to be developed with the same technologies.
- A received message is “locked" during processing; if processing fails, the lock expires and the message is available again.
- Queue sharing can be restricted by IP address and time-of-day.

CloudWatch

- Monitoring infrastructure used by application developers, users, and system administrators to collect and track metrics important for optimizing the performance of applications and for increasing the efficiency of resource utilization.
- Without installing any software a user can monitor either seven or eight pre-selected metrics and then view graphs and statistics for these metrics.
- When launching an Amazon Machine Image (AMI) the user can start the CloudWatch and specify the type of monitoring:
  - Basic Monitoring - free of charge; collects data at five-minute intervals for up to seven metrics.
  - Detailed Monitoring - subject to charge; collects data at one minute interval.

Elastic Beanstalk

- Handles automatically the deployment, capacity provisioning, load balancing, auto-scaling, and monitoring functions.
- Interacts with other services including *EC2*, *S3*, SNS, Elastic Load Balance and AutoScaling.
- The management functions provided by the service are:
  - Deploy a new application version (or rollback to a previous version).
  - Access to the results reported by CloudWatch monitoring service.
  - Email notifications when application status changes or application servers are added or removed.
  - Access to server log files without needing to login to the application servers.
- The service is available using: a Java platform, the PHP server-side description language, or the .NET framework.

CloudFront

- For content delivery: distribute content to end users with a global network of edge locations.
  - “Edges”: servers close to user’s geographical location
- Objects are organized into distributions
  - Each distribution has a domain name
- Distributions are stored in a S3 bucket

Edge servers

- US
- EU
  - US and EU are partitioned to different regions
- Hongkong
- Japan

AWS services

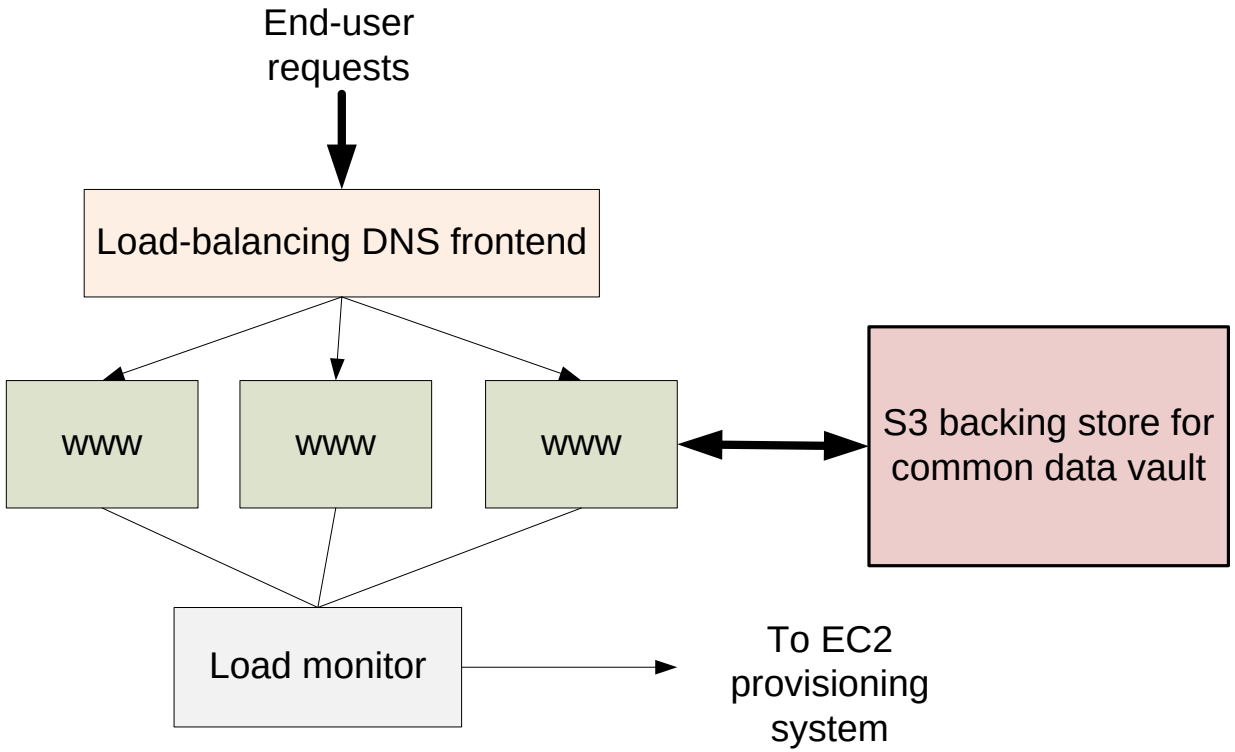
- *Route 53* - low-latency DNS service used to manage user's DNS public records.
- *Elastic MapReduce (EMR)* - supports processing of large amounts of data using a hosted Hadoop running on *EC2*.
- *Simple Workflow Service (SWF)* - supports workflow management; allows scheduling, management of dependencies, and coordination of multiple *EC2* instances.
- *ElastiCache* - enables web applications to retrieve data from a managed in-memory caching system rather than a much slower disk-based database.
- *DynamoDB* - scalable and low-latency fully managed NoSQL database service.

AWS services

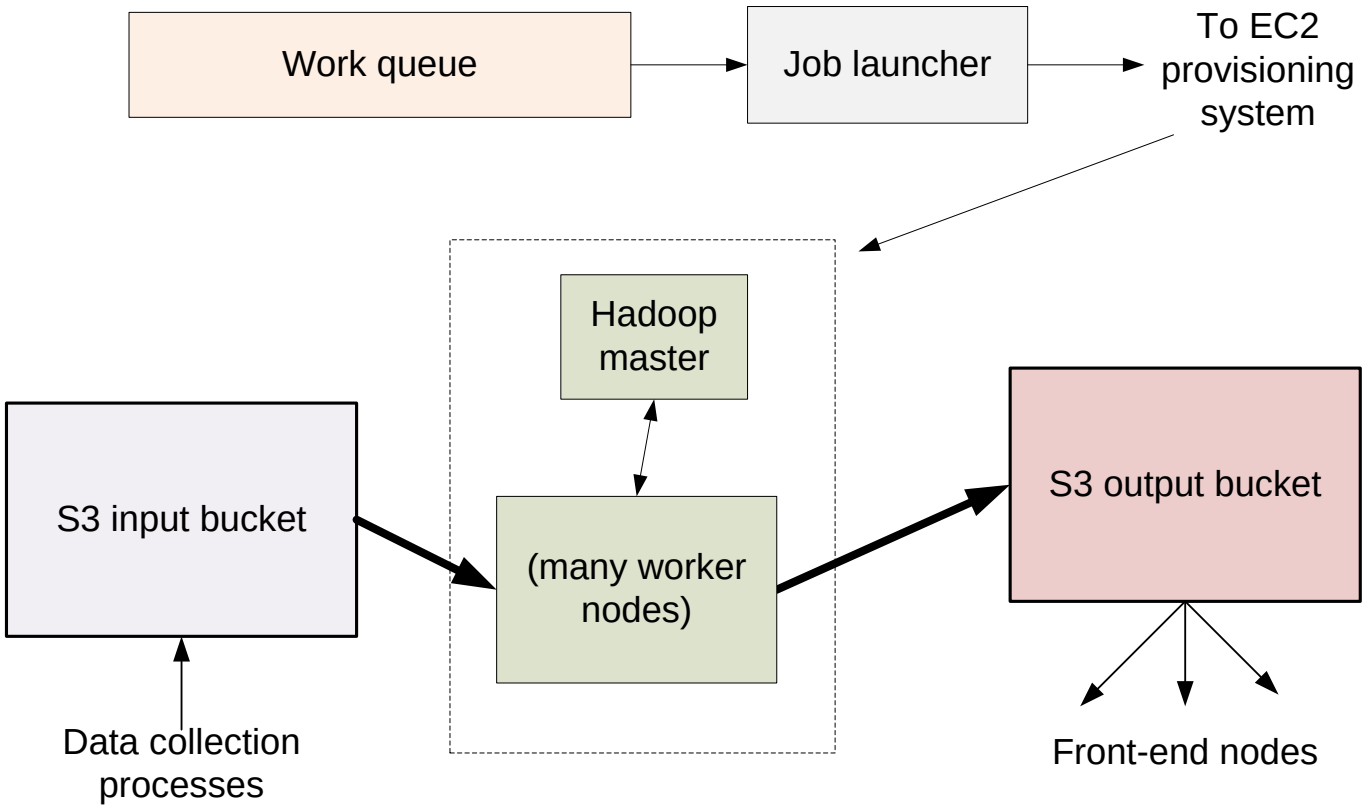
- *CloudFront* - web service for content delivery.
- *Elastic Load Balancer* - automatically distributes the incoming requests across multiple instances of the application.
- *Elastic Beanstalk* - handles automatically deployment, capacity provisioning, load balancing, auto-scaling, and application monitoring functions.
- *CloudFormation* - allows the creation of a stack describing the infrastructure for an application.

MORE ON USAGE

Self-Scaling Applications



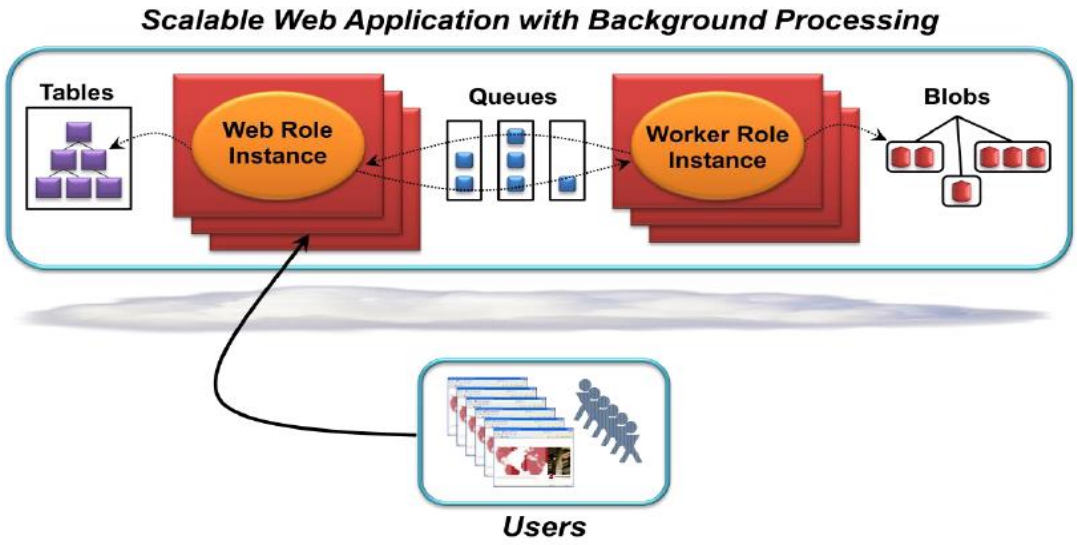
Self-Scaling Backends



CASE STUDIES

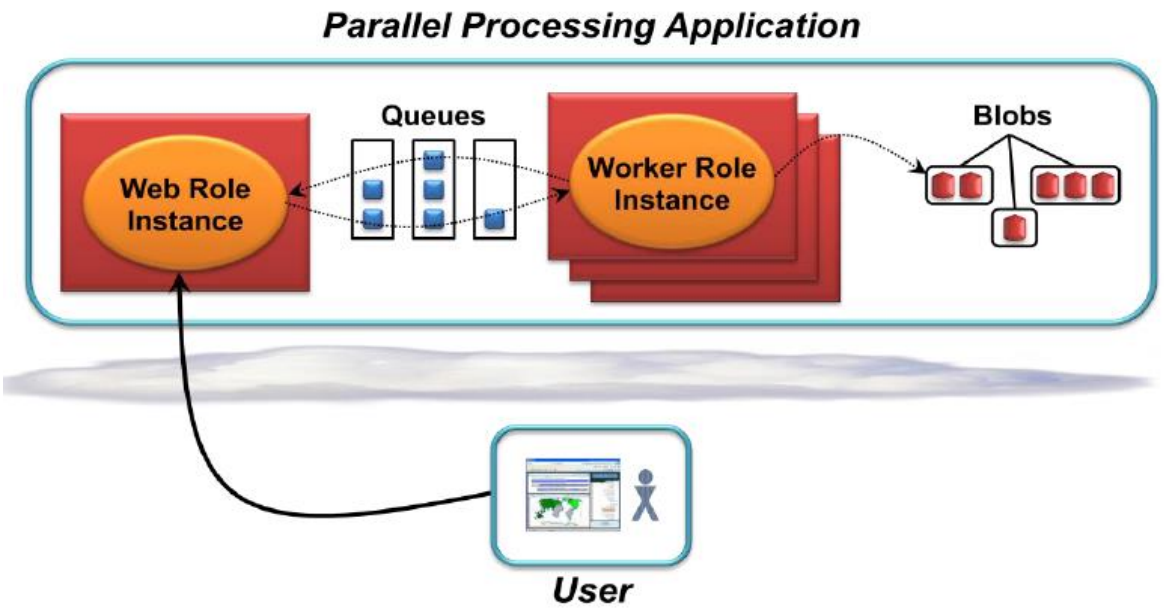
Case Study 1

- Large scale web applications with occasional huge spikes and background processing.
  - Video sharing site
- Deployment
  - A number of web instances based on demand
  - Table storage for information
  - Many works for processing
  - Blobs storage for large data set



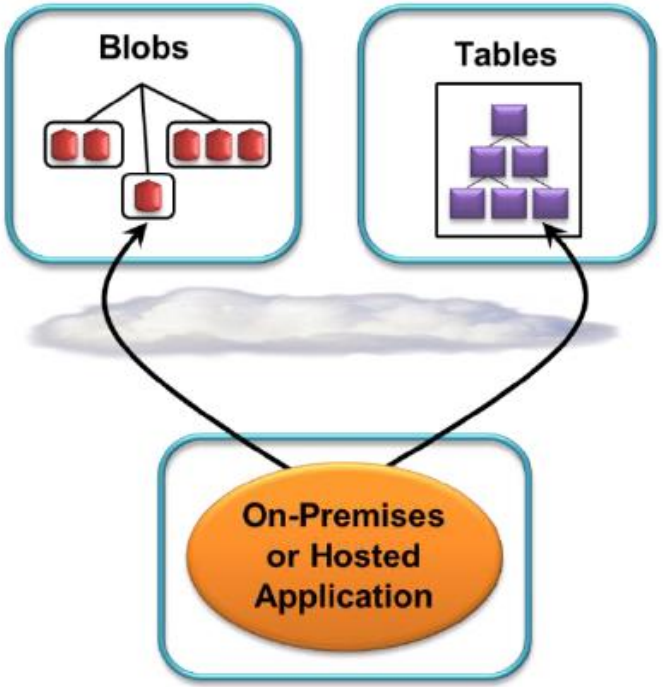
Case Study 2

- Parallel processing applications
  - Financial modeling at a bank
  - New drug testing simulations in a pharmaceutical company
- Deployment
  - Web role for access interface
  - Many workers for processing
  - Large data set stored in blobs



Case Study 3

- Using storage from an on-premises or hosted application
  - Archive old email
  - User log file
- Deployment
  - Connect on-premises application with Azure

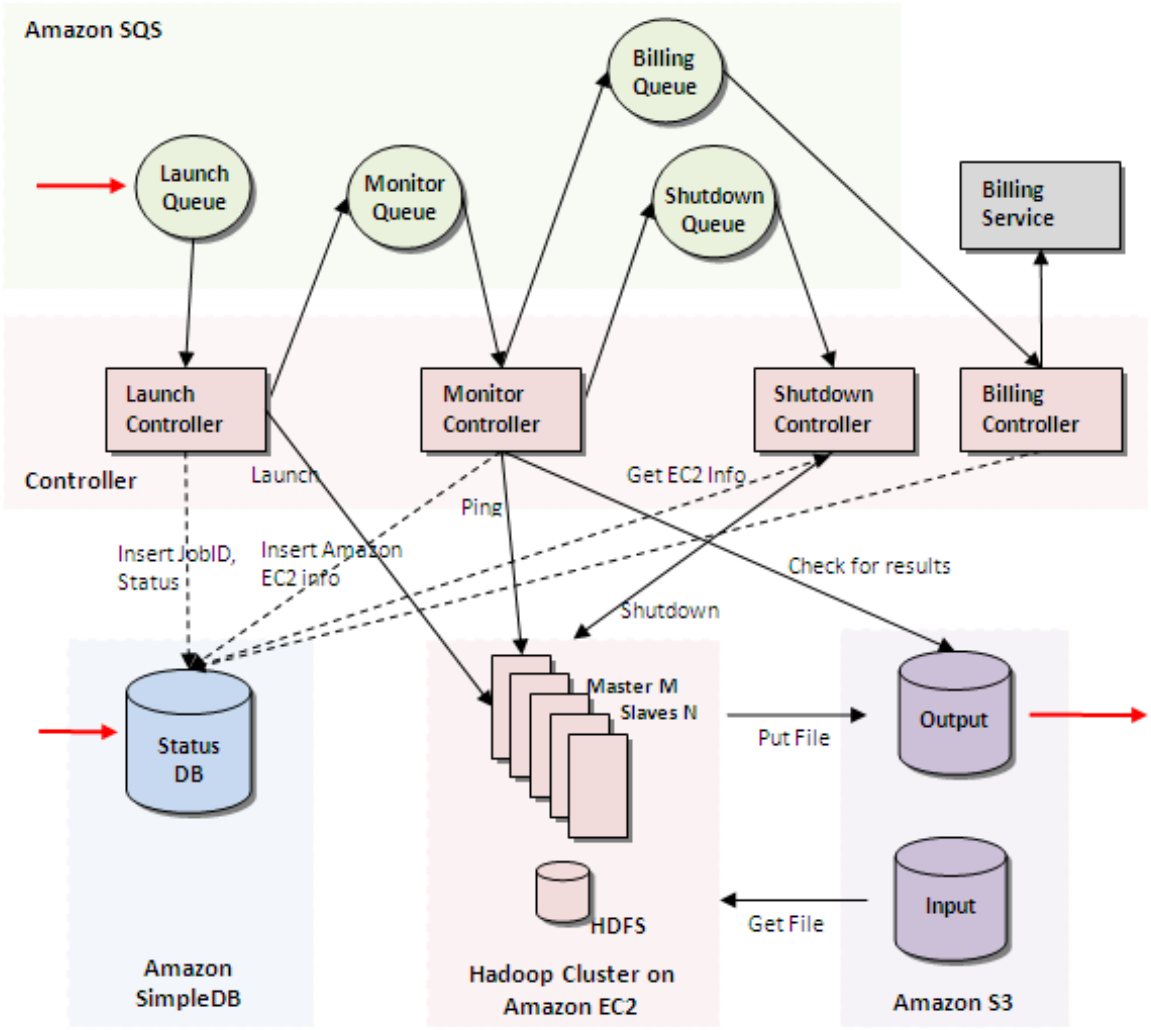


Case Study 4

- Crawling the web
- Large web crawl data is stored in S3
- Users can submit regular expression to the “search” program – “GTW: grep the web”
  - uses Hadoop to search for data
  - Puts your results in an output bucket and notifies you when it’s ready

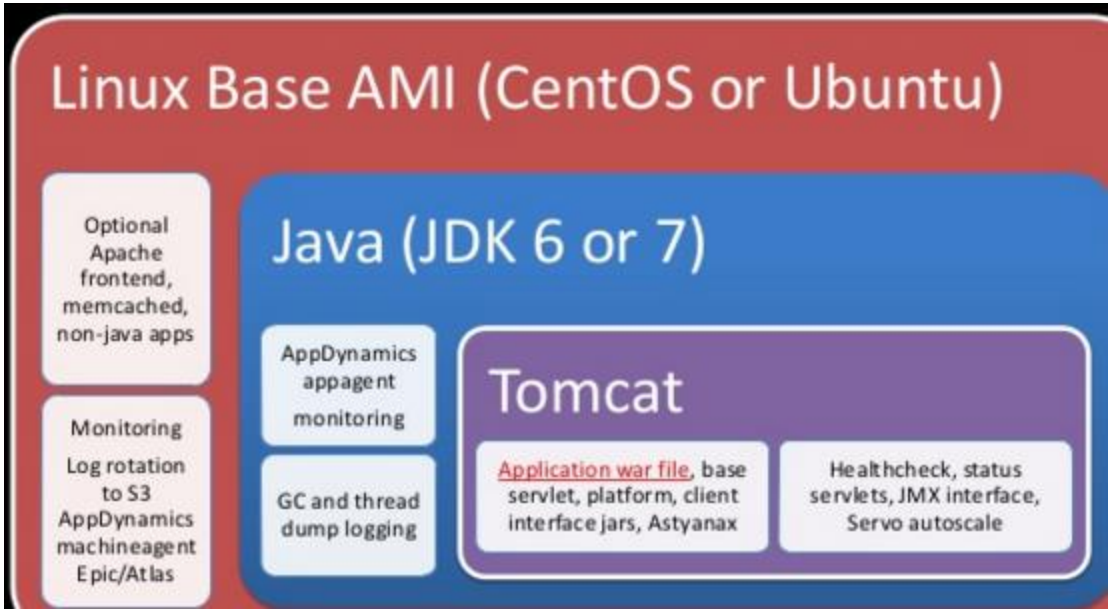
Figure 3: Phases of GrepTheWeb Architecture

Figure 4: GrepTheWeb Architecture - Zoom Level 3



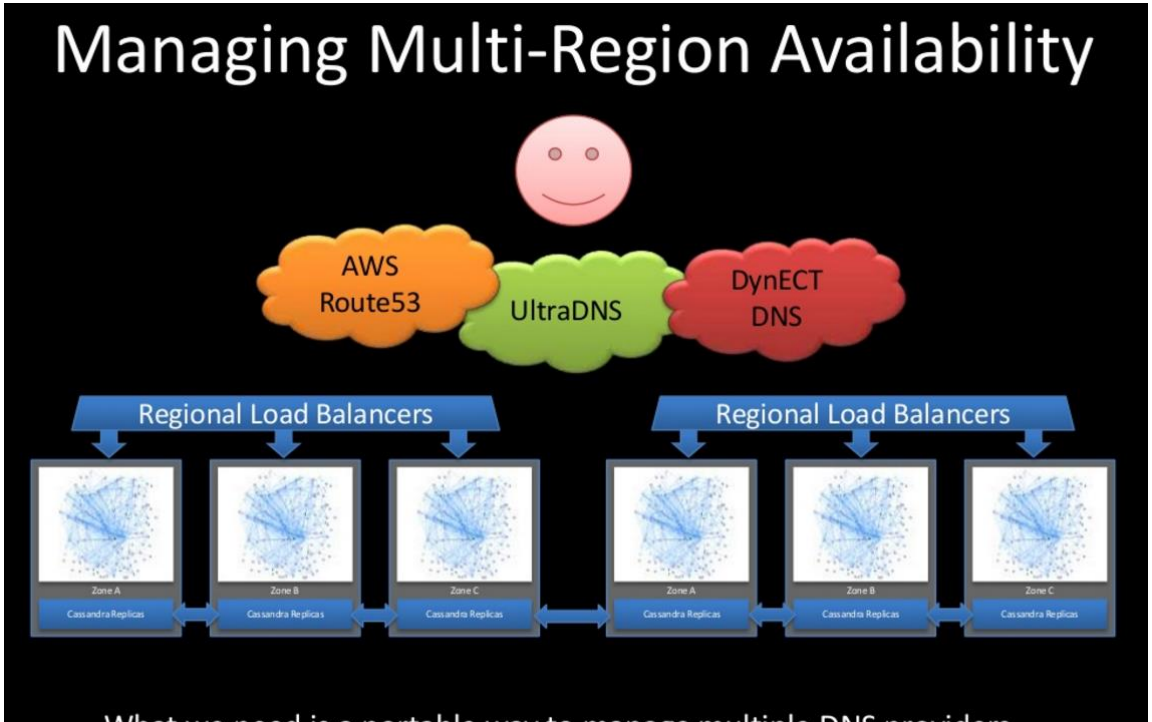
Case Study 5 - netflix

- “Netflix runs tens of thousands of AWS EC2 instances, _.”



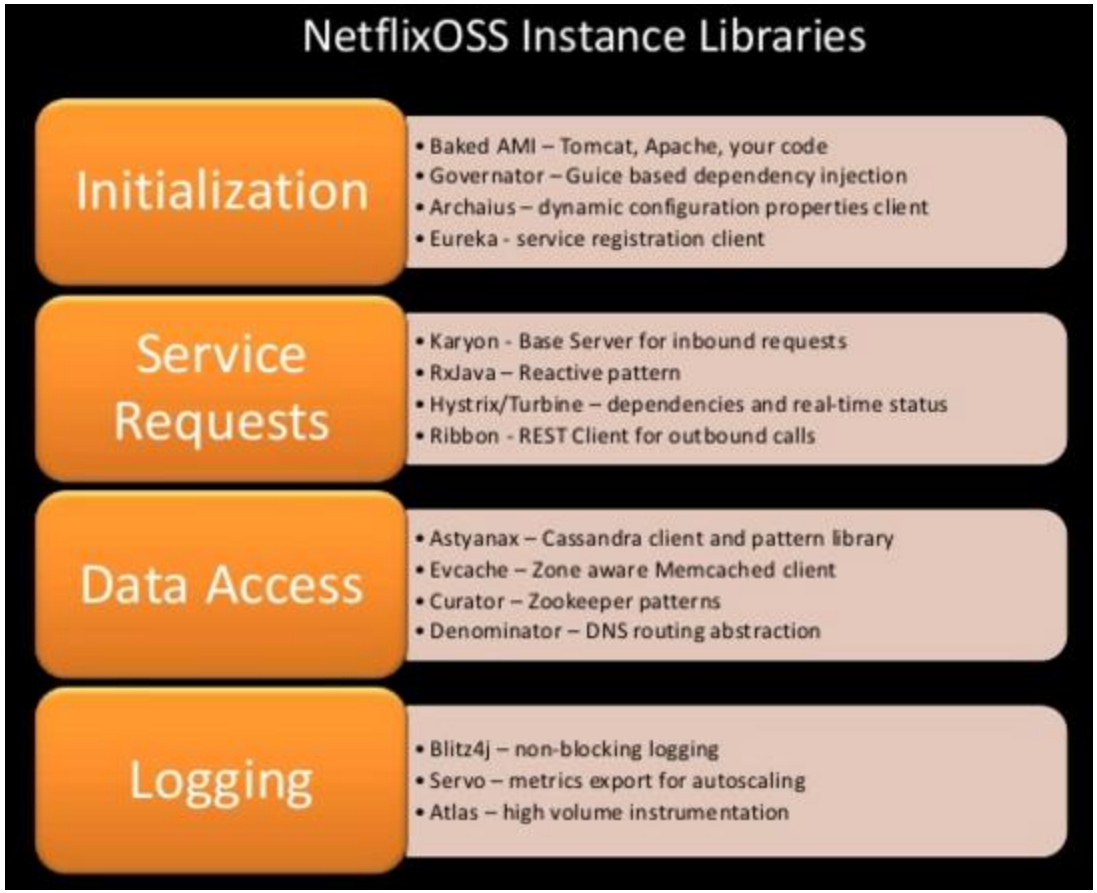
Netflix ---develop own management tools

- “trying to track and manage that number of resources via the AWS Management Console would be unworkable
- Netflix created its own AWS management tools to manage any of its applications running in AWS.
- at www.slideshare.net/adrianco/netflix-and-open-source .”



Netflix –OSS –management tools

- Some of it….



Case study

- The New York Times used AWS to create PDF files of its whole archive
  - 100 Amazon EC2 instances running Hadoop application
  - Processed 4TB of raw TIFF image data (stored in S3) into 11 million finished PDFs
  - Running time: 24 hours
  - Cost: $240 (not including bandwidth)

New York Times report

WAYS TO ACCESS AWS

More than the console

Accessing AWS

- AWS Management Console
- Command-line interface
- API SDKs (java, python, php, ruby, .net, android, ios,more…)
  - We can use APIs to bridge the gap between hardware and applications.
  - IDE integraton
    - Example Eclipse pluggin develop, debug, integrate, migrate, and deploy Java-based applications that use the AWS resources platform