

Infrastructure as a Service (IaaS)

Week 3

School of Software

Faculty of Engineering and Information Technology

University of Technology Sydney



SCHOOL OF SOFTWARE

Learning Objectives

- Understand Computing Infrastructure in Enterprises
- Understand and define Infrastructure as a Service (IaaS)
- Understand Generic Architecture of IaaS Service Provisioning
- Introduction to Amazon Web Services
- Understand traditional web hosting and Cloud-based web hosting

Computing Infrastructure in Enterprises

- Web Servers
 - Load balancers between Web Servers ... etc.
- Application servers
 - Load balancers between application servers etc.
- Database Servers
 - Load Balancers between Database Servers
 - Query Optimizers ...etc.
- Memory
 - RAM (SD-, RD-, DDR-, DDR2-)
 - Non-Volatile Memory
- Variations and enhancements in processing speed, capacity, latency, and cost over the years etc

Computing Infrastructure in Enterprises

- Enterprise applications have specific hardware requirements
 - Different applications have different hardware requirements (one-size does not fit all)
 - Large scale data mining application vs. Microsoft office applications
- Implications of procuring, maintaining and de-registering of IT Infrastructure
 - Budgeting for equipment;
 - Personnel hiring and management;
 - Real-estate
- Would make (business) sense to rent these (depending on the specific application requirements), and use them.
- Cloud computing provides a solution to the above issues in the form of Infrastructure-as-a-Service (IaaS)

Infrastructure-as-a-Service

- The NIST Definition of IaaS

“The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources, where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications.” (Mell and Grance 2011)

Understanding the IaaS definition

- Key elements of the IaaS definition
 - Provision (processing, storage and other fundamental computing resources) on demand;
 - Cloud Consumer is able to deploy and run arbitrary software; (operating system or application)
 - The cloud consumer does not manage the underlying cloud infrastructure
- Implications of IaaS:
 - No up-front capital investment for the cloud consumer;
 - No administrative or maintenance expenses for the cloud consumer;
 - No real-estate expenses for the cloud consumer;
 - Cloud Consumer is charged for resources (Pay-as-you-go)
- Simplest (or basic) flavour of cloud offering

Generic Architecture of IaaS Service Provisioning Process

- IaaS provider has to setup the IaaS stack
 - IaaS Service Provisioning Process is carried out by the IaaS Stack
- “Layers” of the IaaS Stack are
 - Co-location Layer
 - Hardware Layer
 - Virtualization Layer
 - Service Layer

Pictorial Representation of IaaS Service Provisioning

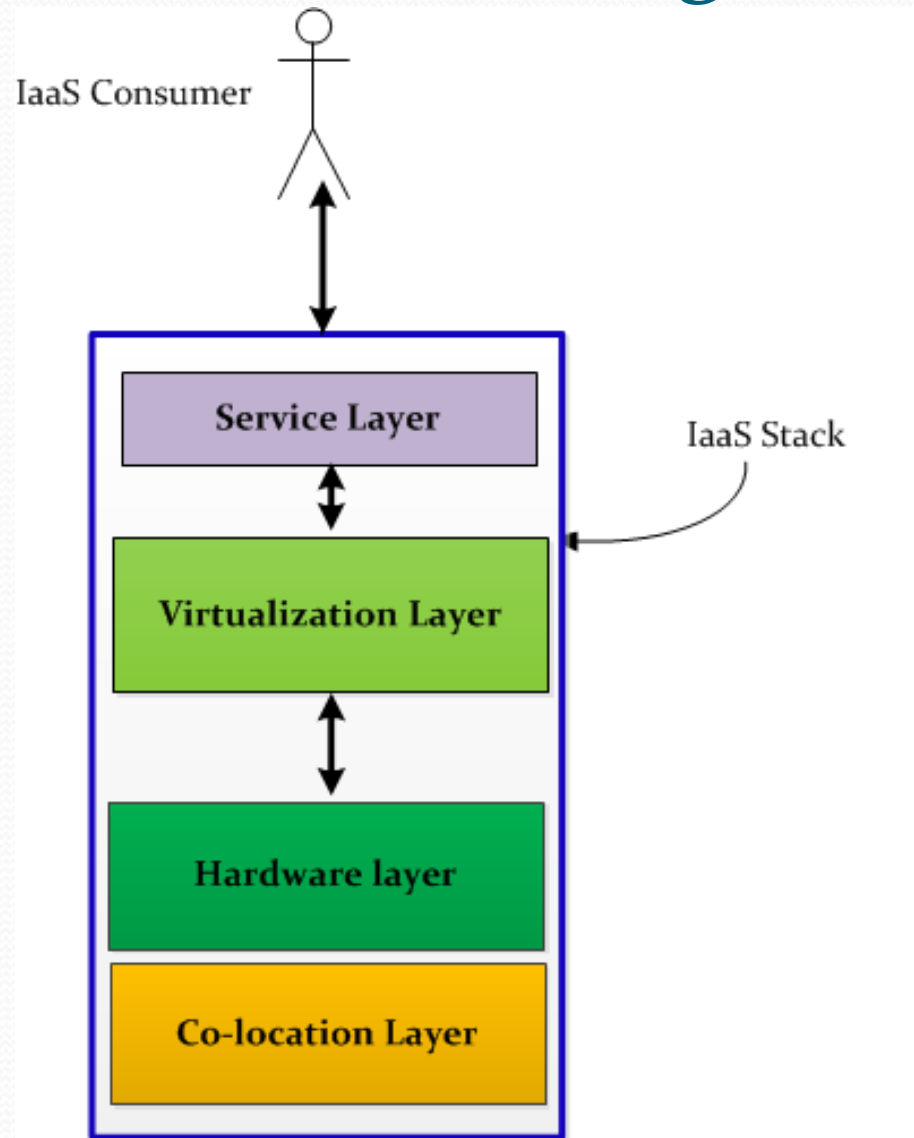


Figure: IaaS Stack

Co-location Layer

- Provide the basic requirements needed to deliver IaaS services
 - Electrical Power
 - Redundant electric power supply lines
 - Power supply lines from multiple electricity suppliers
 - Backup power supply
 - Bandwidth
 - Redundant network connectivity
 - Fault-tolerant network connectivity (multiple “physical” links from different network providers)
 - Cooling (efficient design, major cost-constraint)
 - Real Estate (choice of location, cheap yet reliable, “safe” from natural disasters, political stability ...etc)

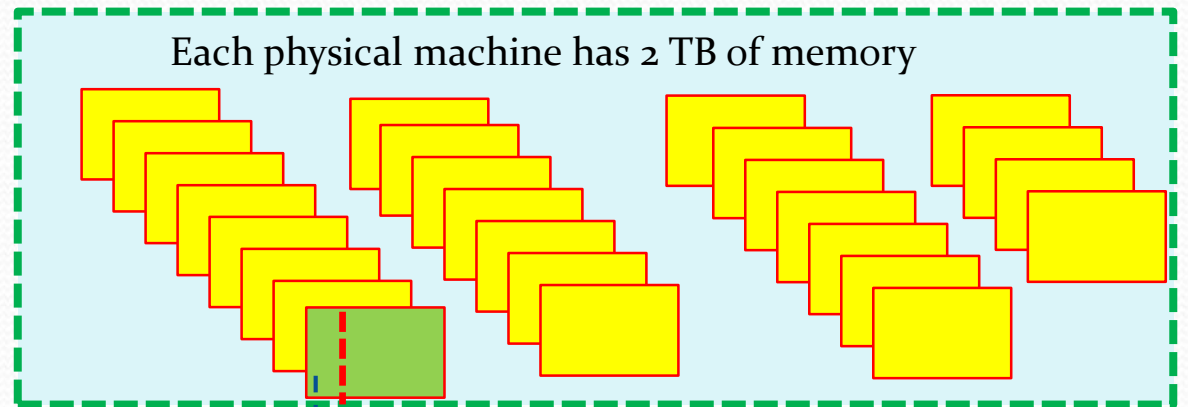
Hardware Layer

- Comprises of the physical hardware that will be provisioned “*virtually*” to the cloud consumers (data centres)
- Actual or physical resource present in the data centre
- Maximize the hardware capacity to give the cloud consumers the illusion of infinite capacity.

Virtualization Layer

- Key layer to enable IaaS Service Provisioning

Data centre

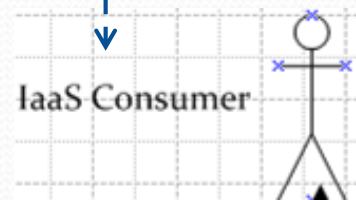


The virtualization layer is responsible for automatically:

- Finding an appropriate physical machine that can host the consumer's request;
- Slicing it to meet customer's requirements;
- Allocating the sliced part (virtual view) to the consumer automatically
- Maintaining a snapshot of resource usage (free/being used etc..)



50 GB of memory for
2 hours



Virtualization Layer

- The virtualization layer is responsible for automatically:
 - Finding an appropriate physical machine that can host the consumer's request;
 - Slicing it to meet customer's requirements;
 - Allocating the sliced part (virtual view) to the consumer automatically
 - Maintaining a snapshot of resource usage (free/being used etc..)

Service Layer

Define the interface to the IaaS Consumer to interact with the provisioned resources (which have been virtualized)

Navigation

Region: US East (Virginia)

EC2 Dashboard

Scheduled Events

INSTANCES

- Instances
- Spot Requests
- Reserved Instances

IMAGES

- AMIs
- Bundle Tasks

ELASTIC BLOCK STORE

- Volumes
- Snapshots

NETWORK & SECURITY

- Security Groups
- Elastic IPs
- Placement Groups
- Load Balancers
- Key Pairs
- Network Interfaces

Getting Started

To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.

[Launch Instance](#)

Note: Your instances will launch in the US East (Virginia) region.

Service Health

Service Status

Current Status	Details
Amazon EC2 (US East - N. Virginia)	Service is operating normally

[View complete service health details](#)

Availability Zone Status

Current Status	Details
us-east-1a	Availability zone is operating normally
us-east-1b	Availability zone is operating normally
us-east-1c	Availability zone is operating normally
us-east-1d	Availability zone is operating normally

Amazon Web Services (AWS)

- Amazon Web Services (AWS) is a collection of computing services offered over the Internet by Amazon.com
- Amazon.com is one of the heavily trafficked websites globally
- Houses and provides a wide range of computing infrastructure (compute, storage etc...); and provisions them “as-a-service” to the end-users on-demand
- AWS value proposition - *“You pay for what you use”*
- Amazon Infrastructure can be provisioned for any type of application usage
- Increasingly popular platform for website hosting and deployment



Architecture of Amazon Web Services

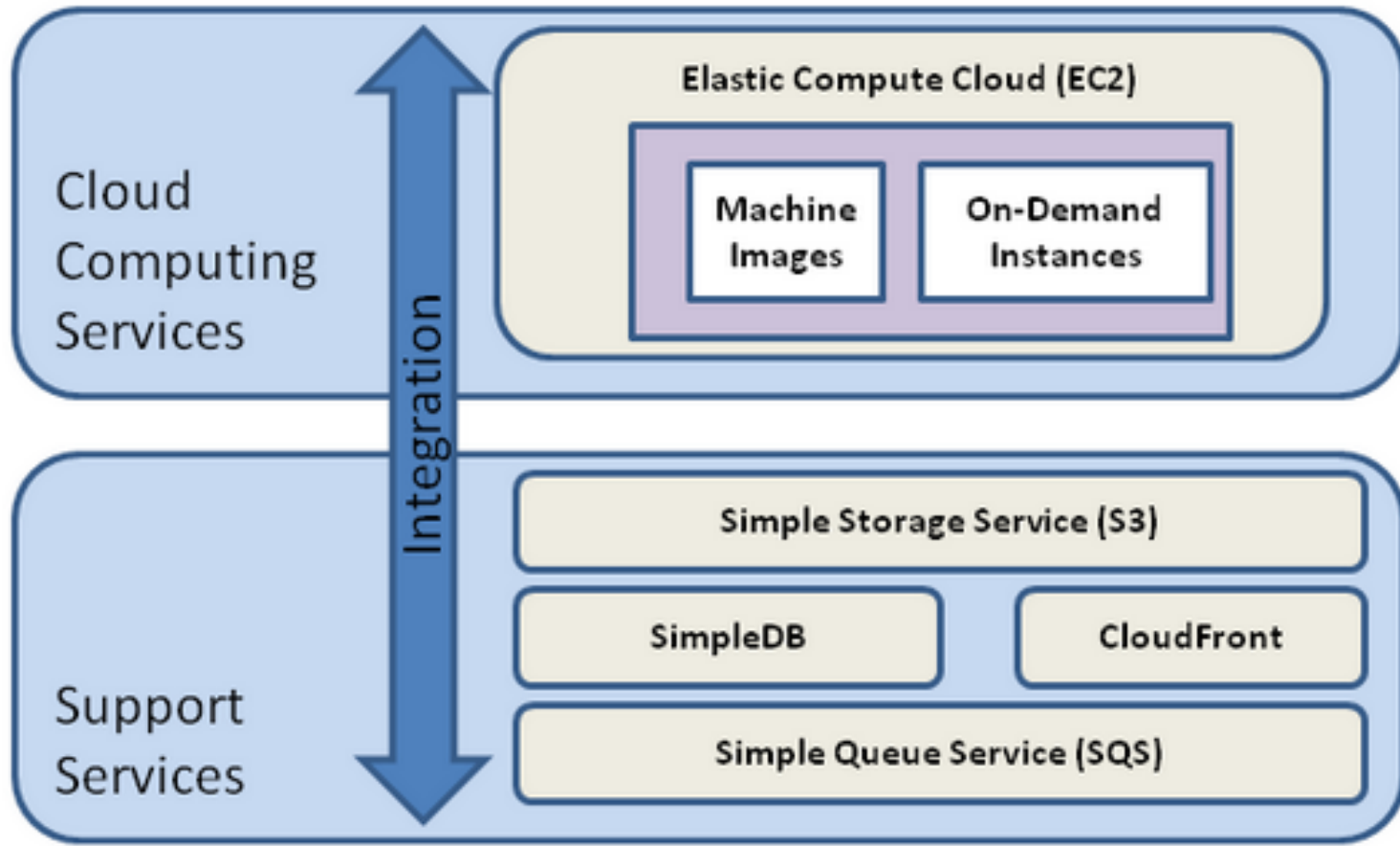
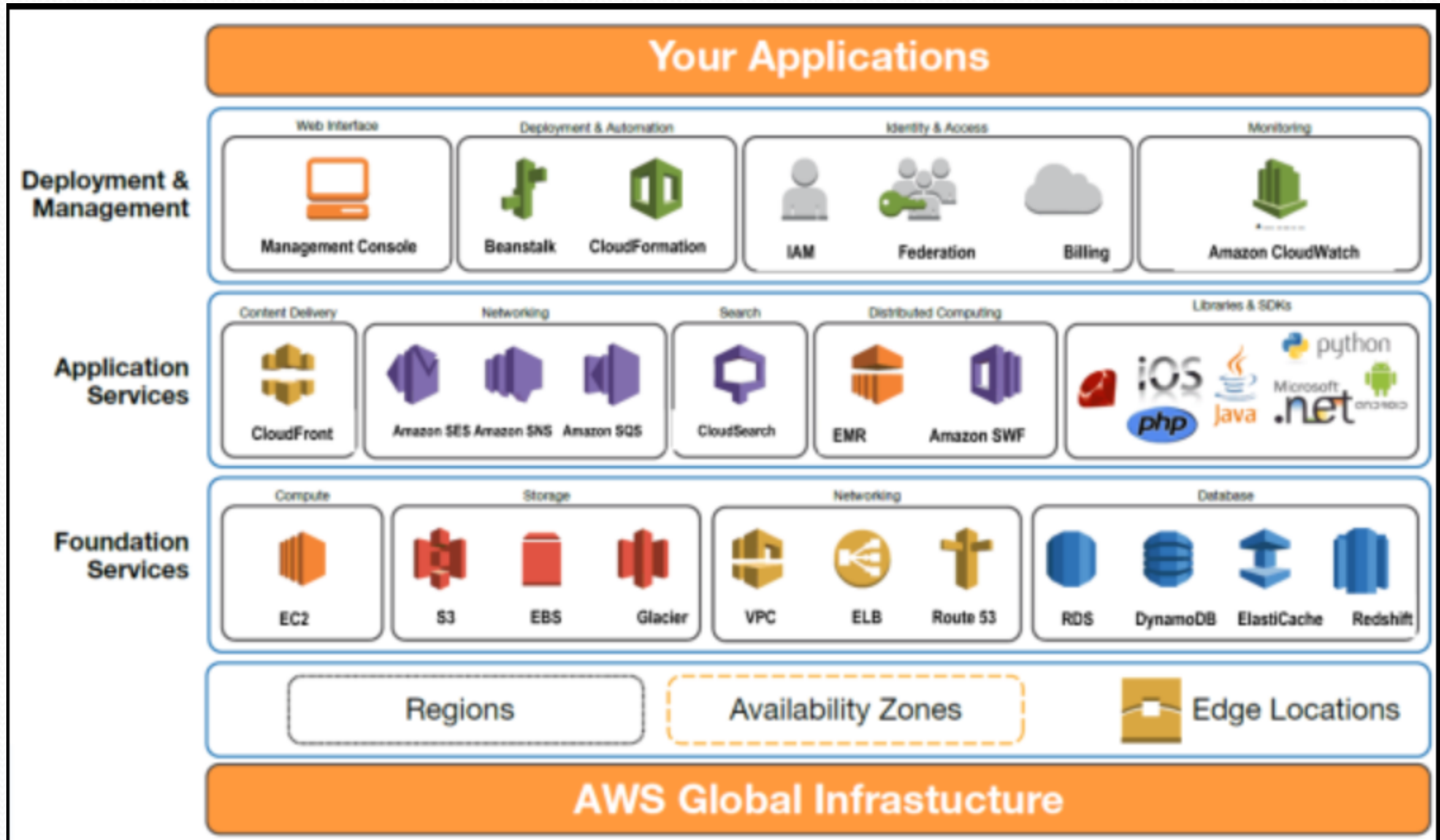


Figure: Architecture of Amazon Web Service (AWS)

AWS service offerings



Components of AWS

- Amazon Elastic Compute Cloud (EC2) Instances
- Amazon Simple Storage Service (S3)
- Amazon Relational Database Service
- Amazon Cloud Front Service
- Amazon Simple Queue Service (SQS)
- Amazon Glacier

Architecture of Amazon Web Service

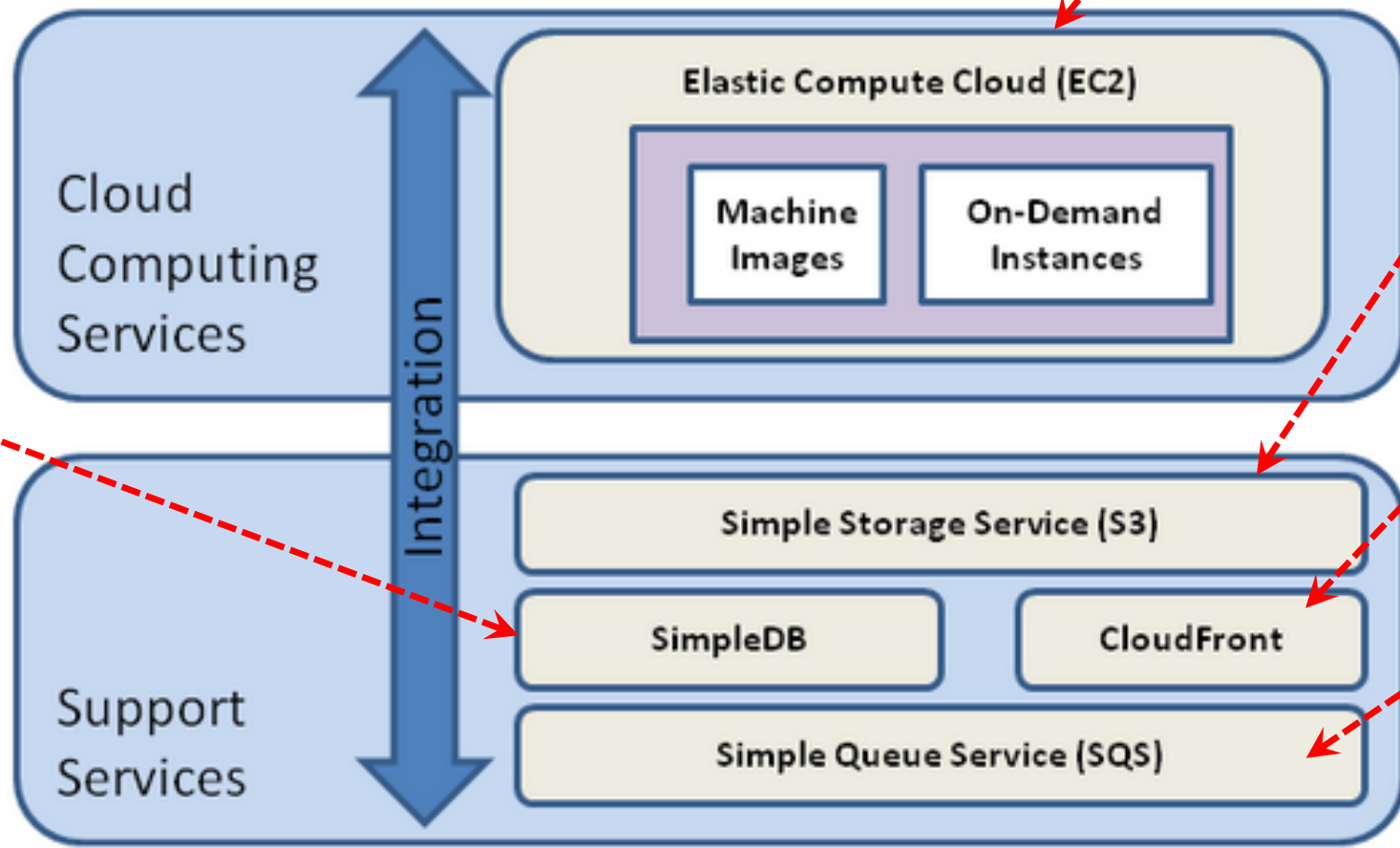


Figure: Architecture of Amazon Web Service (AWS)

AWS EC2 instances (on demand)



- Provide virtual machines to cloud consumers
- (As of 2019) EC2 instances are available in six different types
 - General;
 - Compute-optimized (lowest price per compute);
 - Memory-optimized (lowest price per GiB of RAM);
 - GPU-optimized (optimized for graphical and GPU-applications)
 - Storage-optimized (optimized for high I/O)
- Choose as per your application requirements
- Combine instances from different types

Different types of EC2 pricing models



- Comes with three pricing models
 - On-Demand Instance - Pay by the hour with no long term commitments;
 - Reserved Instance - Very cheap compute power (compared to the on-demand instances);
 - Up-front payment (All up-front, no up-front, partial up-front);
 - Fixed Date and time for resource usage
 - Spot Instances – Bid and use spare EC2 instances.

AWS EC2 instances



AWS Management Console > Amazon EC2

Navigation

Region: Asia Pacific (Singapore)

EC2 Dashboard

Scheduled Events

INSTANCES

- Instances
- Spot Requests
- Reserved Instances

IMAGES

- AMIs
- Bundle Tasks

ELASTIC BLOCK STORE

- Volumes
- Snapshots

NETWORK & SECURITY

- Security Groups
- Elastic IPs
- Placement Groups
- Load Balancers
- Key Pairs

Amazon EC2 Console Dashboard

Getting Started

To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.

[Launch Instance](#)

Note: Your instances will launch in the Asia Pacific (Singapore) region.

Service Health

Service Status

Current Status	Details
Amazon EC2 (APAC - Singapore)	Service is operating normally

[View complete service health details](#)

Availability Zone Status

Current Status	Details
ap-southeast-1a	Availability zone is operating normally

Figure: Launching an Amazon EC2 instance

Request Instances Wizard

CHOOSE AN AMI INSTANCE DETAILS CREATE KEY PAIR CONFIGURE FIREWALL REVIEW

Choose an Amazon Machine Image (AMI) from one of the tabbed lists below by clicking its **Select** button.

Quick Start My AMIs Community AMIs

Basic Amazon Linux AMI 2011.09
Amazon Linux AMI 2011.09, EBS boot with Amazon EC2 AMI Tools.
Root Device Size: 8 GB ☒ 64 bit ☐ 32 bit [Select](#)

Red Hat Enterprise Linux 6.2
Red Hat Enterprise Linux version 6.2, EBS-boot.
Root Device Size: 6 GB ☒ 64 bit ☐ 32 bit [Select](#)

SUSE Linux Enterprise Server 11
SUSE Linux Enterprise Server 11 Service Pack 1 basic install, EBS boot with Amazon EC2 AMI Tools preinstalled; Apache 2.2, MySQL 5.0, PHP 5.3, Ruby 1.8.7, and Rails 2.3.
Root Device Size: 10 GB ☒ 64 bit ☐ 32 bit [Select](#)

Ubuntu Server Cloud Guest 11.10 (Oneiric Ocelot)
Ubuntu Server version 11.10 (Oneiric Ocelot) optimized for use on AWS. Commercial support available at <http://www.canonical.com/enterprise-services/ubuntu-advantage/cloud>
Root Device Size: 8 GB ☒ 64 bit ☐ 32 bit [Select](#)

Ubuntu Server Cloud Guest 10.04 LTS (Lucid Lynx)
Ubuntu Server version 10.04 (Lucid Lynx) optimized for use on AWS. Commercial support available at <http://www.canonical.com/enterprise-services/ubuntu-advantage/cloud>
Root Device Size: 8 GB ☒ 64 bit ☐ 32 bit [Select](#)

Figure: Select an appropriate EC2 instance



EC2

Importing VM images to Amazon EC2

- VM Import
 - Provides the ability to import virtual machine images to Amazon EC2 instances
 - This feature comes pre-packaged with Amazon EC2 instances

Key Features of EC2 instances

- Multiple locations
 - Amazon has defined Availability Zones (AZ) and Regions
 - Multiple regions in total across the globe
 - Select EC2 instances from (multiple) zones within a region to ensure fault tolerance

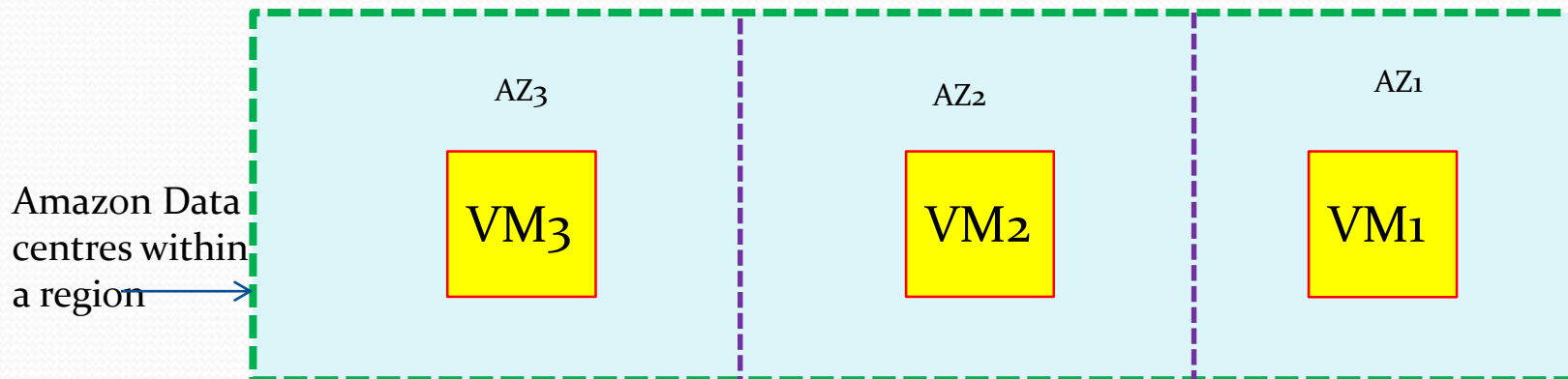
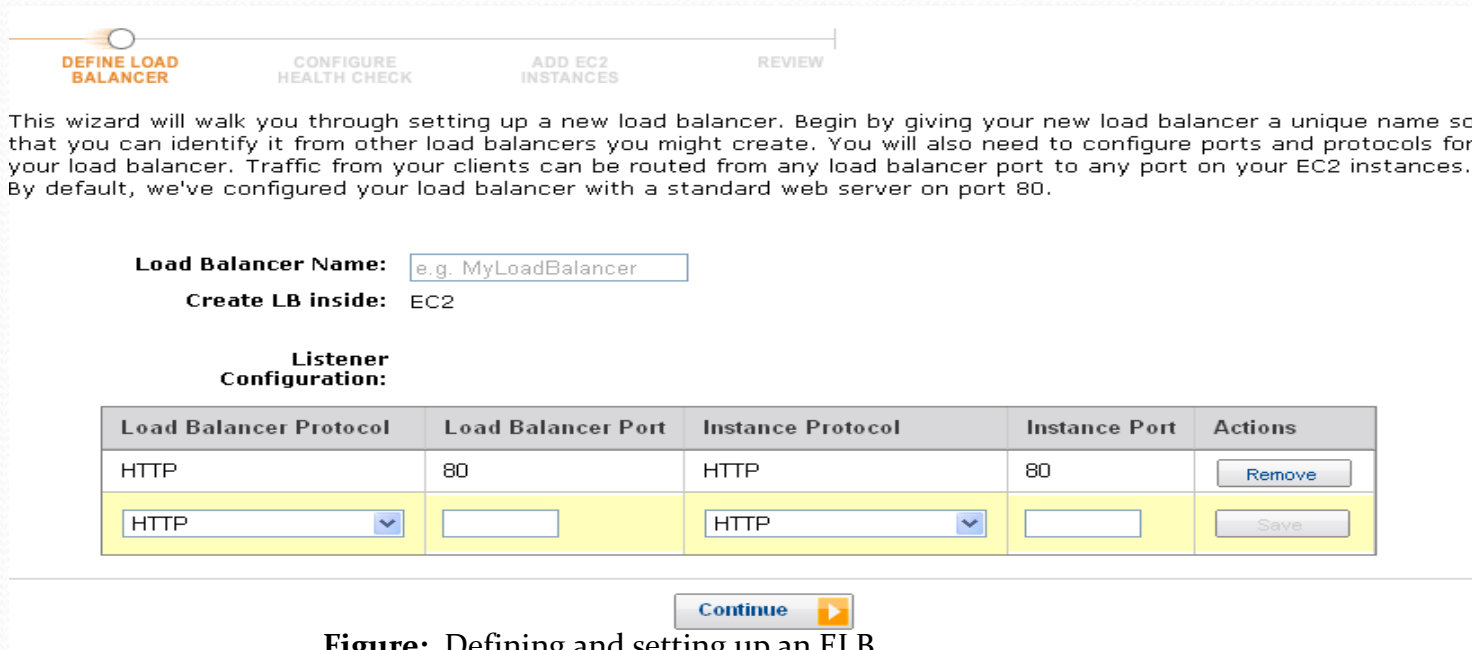


Figure: Illustration of setting up VM's in multiple AZ's

Key Features of EC2 instances

- Elastic Load Balancers (ELB)
 - This comes pre-packaged with EC2 instances
 - Perform load balancing across EC2 instances in an “elastic” manner
 - Set rules or policies to define load balancing (define your own rules)
 - Interacts with Amazon Cloud Watch for load reporting on EC2 instances



The screenshot shows the 'DEFINE LOAD BALANCER' step of a wizard. The progress bar at the top indicates the current step is 'DEFINE LOAD BALANCER', followed by 'CONFIGURE HEALTH CHECK', 'ADD EC2 INSTANCES', and 'REVIEW'.

This wizard will walk you through setting up a new load balancer. Begin by giving your new load balancer a unique name so that you can identify it from other load balancers you might create. You will also need to configure ports and protocols for your load balancer. Traffic from your clients can be routed from any load balancer port to any port on your EC2 instances. By default, we've configured your load balancer with a standard web server on port 80.

Load Balancer Name:

Create LB inside: EC2

Listener Configuration:

Load Balancer Protocol	Load Balancer Port	Instance Protocol	Instance Port	Actions
HTTP	80	HTTP	80	Remove
<input type="text" value="HTTP"/>	<input type="text"/>	<input type="text" value="HTTP"/>	<input type="text"/>	Save


[Continue](#) 

Figure: Defining and setting up an ELB

Elastic Load Balancers

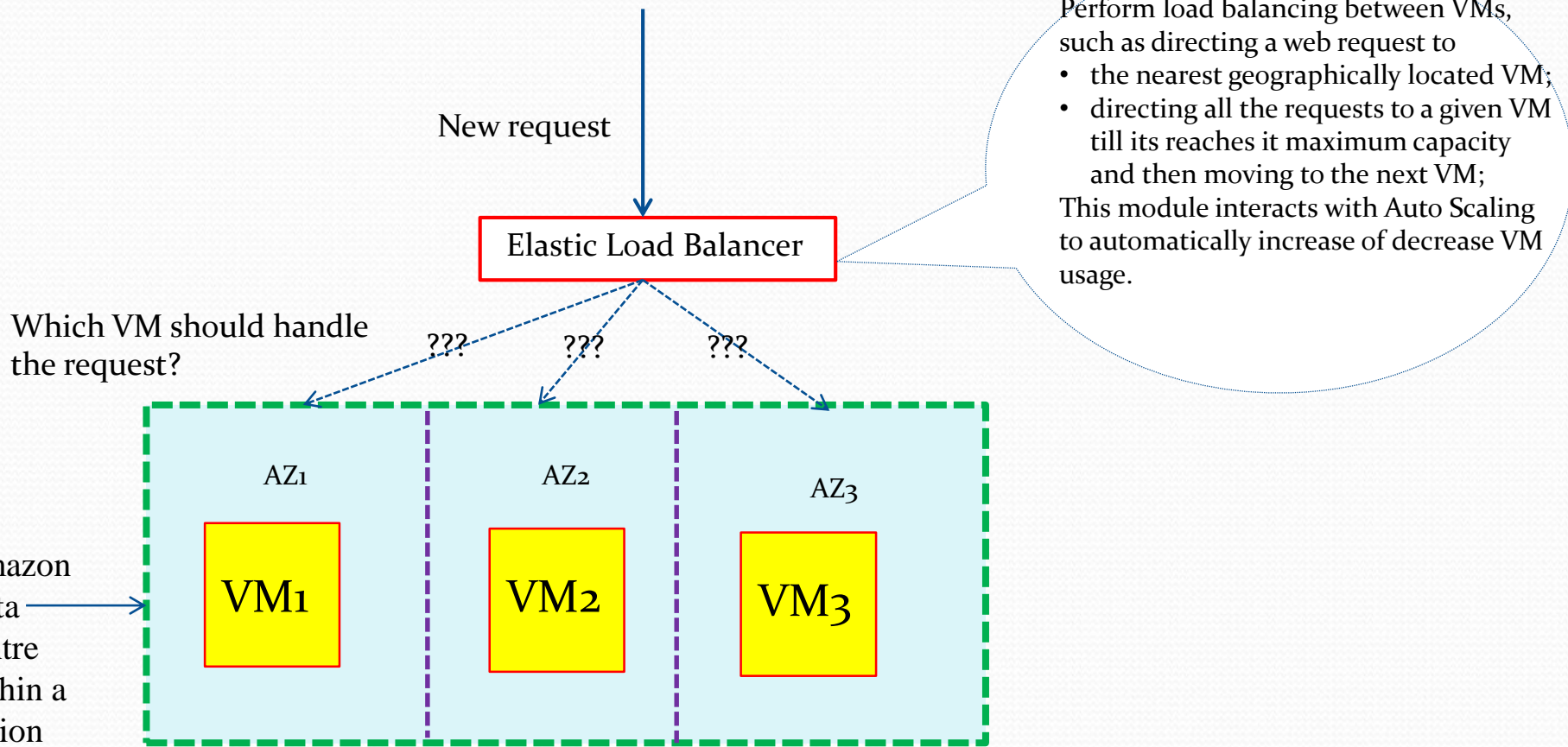


Figure: Illustration of working of the ELB



EC2

Key Features of EC2 instances

• Auto Scaling

- “Scale-up” or “scale-down” the number of used EC2 instances, depending on your requirements
- Could be across multiple or single availability zones
- You can set your own rules or policies to enable auto scaling (request count per VM, request latency per VM...)

When should an additional VM be automatically added? And in which AZ? When should an existing VM be removed?

Auto-scaling module

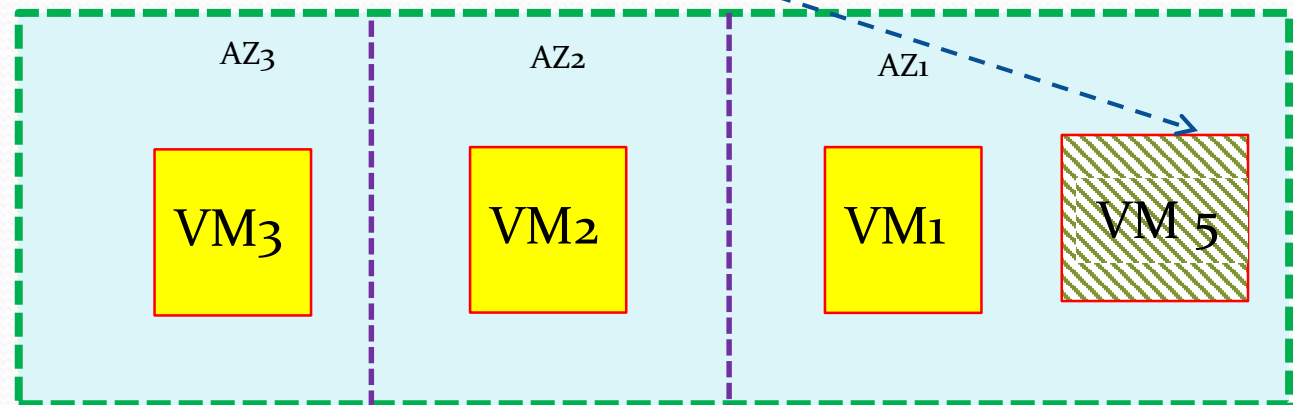


Figure: Illustration of Auto-scaling across multiple AZ's

Key Features of EC2 instances

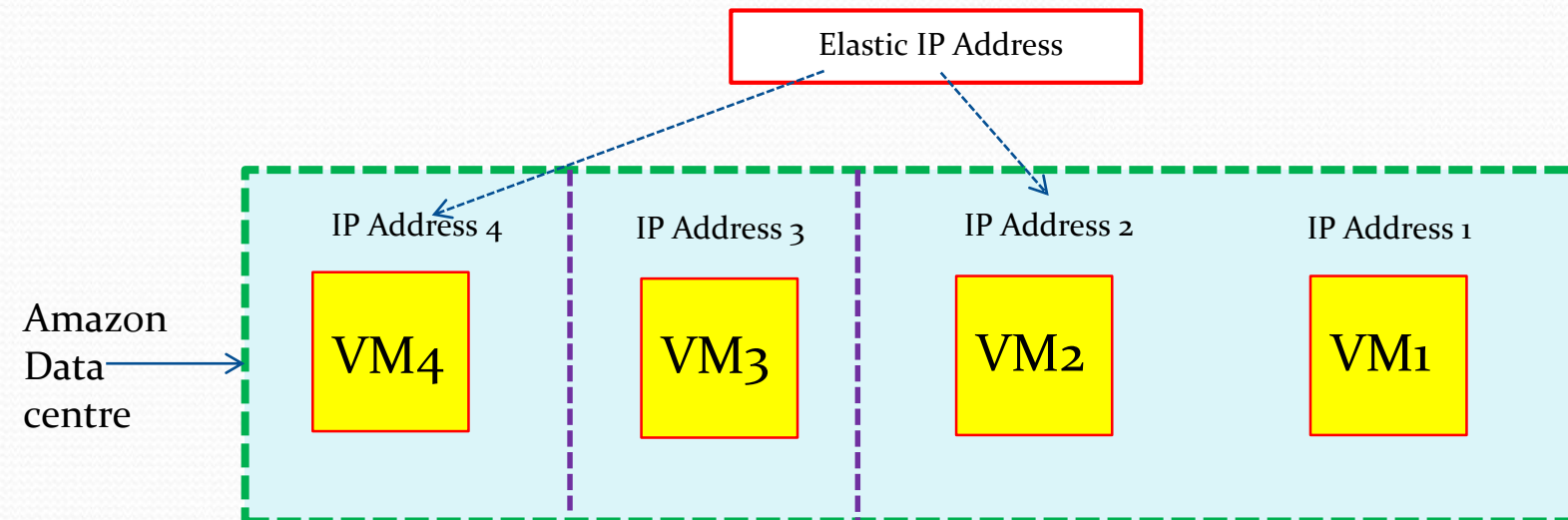
- Security Groups
 - Implements firewalls on EC2 instances using rules (or policies)
 - Rules applied to each EC2 instance
 - Needs to be customized by the consumer for each EC2 instance.
 - Done using EC2 management console



Figure: Defining and setting up a security group for an EC2 instance

Key Features of EC2 instances

- Elastic IP Address for EC2 instances
 - EC2 instances may be moved at run time (failure, performance issues, scaling up, scaling down ...etc)
 - Elastic IP addresses are associated with account rather than EC 2 instance
 - Is a proxy for the actual IP address
 - Mapping between Elastic IP address and actual IP address is performed by Amazon seamlessly



Persistence mechanisms in AWS

- AWS provides the following (web) services for data storage:
 - Amazon Relational Database Service (Amazon RDS)
 - Amazon Simple Storage Service (Amazon S3)
 - Amazon Glacier
- For different purposes and audience

Amazon Relational Database Service (Amazon RDS)



- Amazon RDS provides “relational database-as-a-service” to consumers
 - Set up and access your relational database in the cloud
- Users can select from MySQL, Oracle, or Microsoft SQLServer, Postgre SQL, Amazon Aurora, and Maria DB
 - Database administration and maintenance activities performed by Amazon
- Carries out regular backup and “point-in-time” (time stamped) recovery
- Provides the added advantage of scale-up, scale down of RDS database instances on demand

Amazon Relational Database Service (Amazon RDS)

- As of 2018, 2 flavours of Amazon RDS
- General purpose SSD storage
 - SSD-backed storage option that delivers a consistent baseline of 3 IOPS per provisioned instance
- Provisioned IOPS
 - SSD-backed storage option designed to deliver fast, predictable, and consistent I/O performance.
 - You specify an IOPS rate when creating a database instance, and Amazon RDS provisions that IOPS rate for the lifetime of the database instance.
 - This storage type is optimized for I/O-intensive transactional (OLTP) database workloads. You can provision up to 40,000 IOPS per database instance

Launching a RDS instance

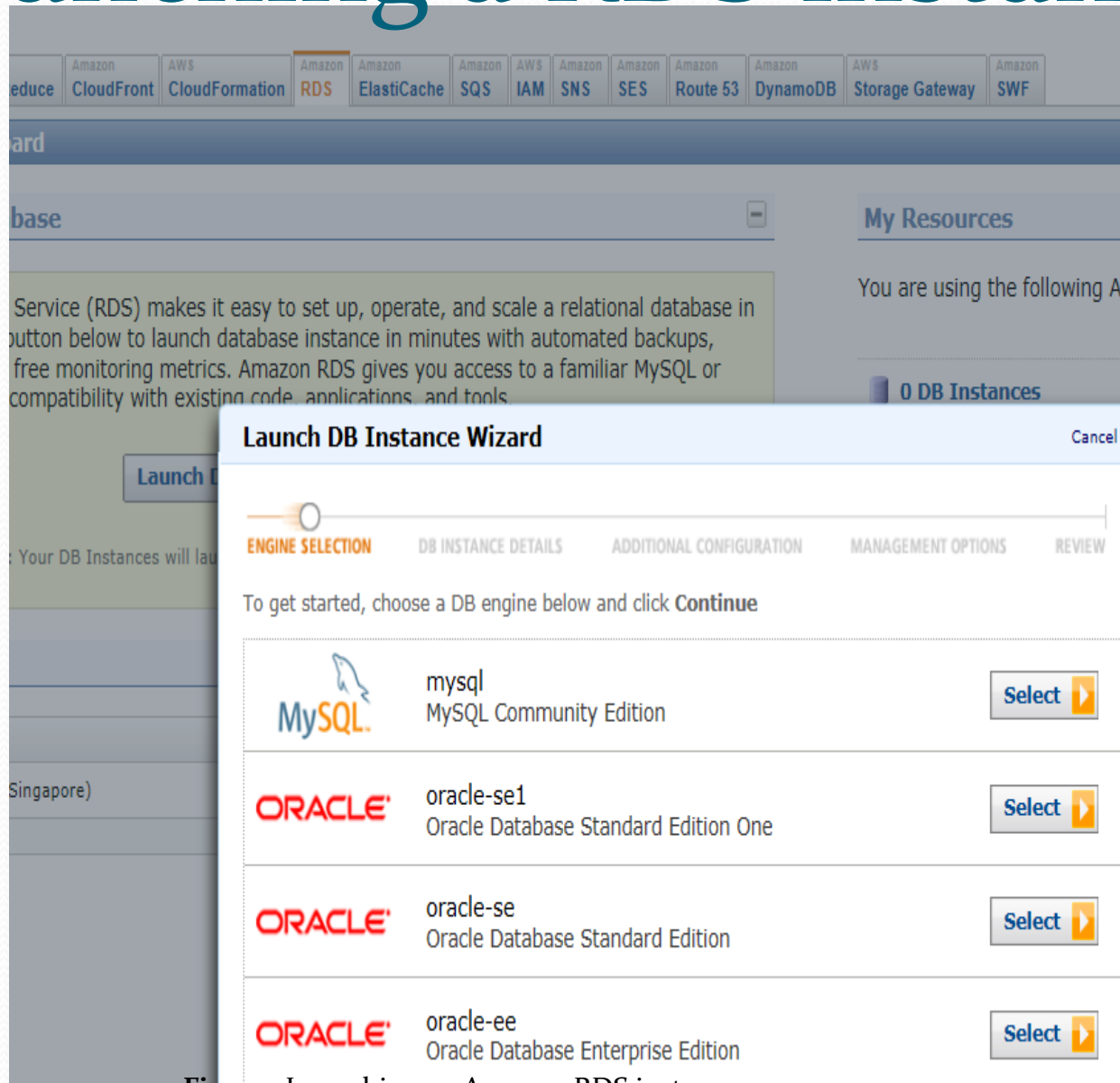


Figure: Launching an Amazon RDS instance

Amazon Simple Storage Service (Amazon S3)



- Amazon S3 is an Online storage service from Amazon
- Very useful for storage of non-structured information (information without an underlying data model) such as images, videos..etc.
- Objects are stored in “buckets”, associated with an Amazon User.
 - Data is stored as “objects”, each of which is assigned unique key for referencing it
 - Size of objects (1 byte to 5 terabytes)



Amazon Simple Storage Service (Amazon S3)



- Very popular web service from Amazon
- Has been used by many providers to provision innovative services (Ex: Dropbox)

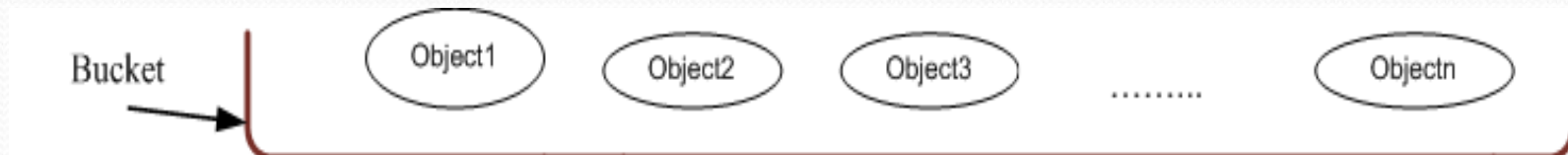


Figure: Bucket and objects in Amazon S3

Amazon Glacier

- Relatively new cloud-based storage service offered by Amazon
 - Useful for data archiving and backup
- Should be used for that is infrequently accessed and for which retrieval times of several hours are suitable.
- Very cheap to store data with Amazon Glacier (\$0.005 per gigabyte per month for Asia-Pacific (Sydney))
- Typical use-cases where it may be used:
 - Archiving media records or assets;
 - Very old transaction records

Amazon CloudWatch

- Monitor the performance of AWS resources
- Available in two flavours (basic and advanced) – vary in terms of frequency and “type” of monitoring provided to the cloud consumer
- Define your own monitoring metrics and define alarm for these metrics
- Present the “health” of AWS resource to the consumer (response time, latency...etc)

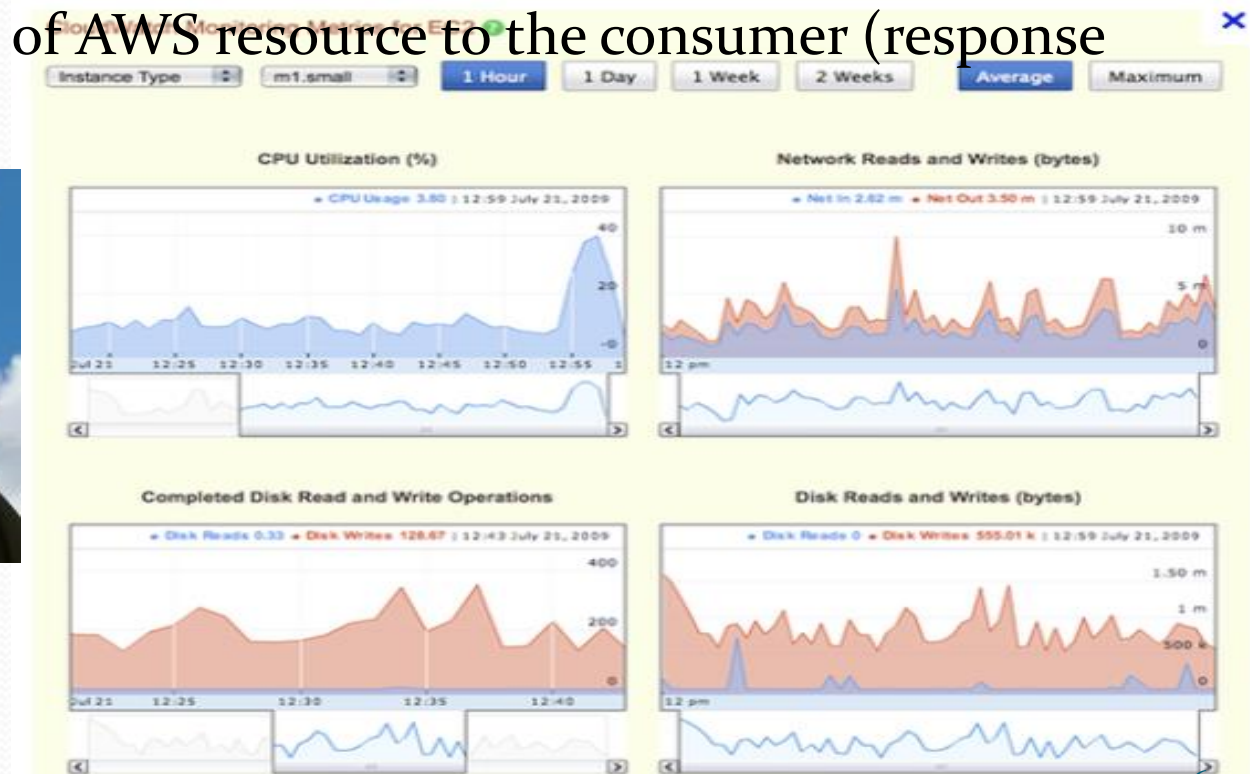


Figure: Amazon CloudWatch Monitoring

Example of AWS metering and billing

- Fine-grained resource usage
- Pay only for what you use!!

This Month's Activity as of December 25, 2010 ✕

The billing cycle for this report is December 1 - December 31, 2010. The AWS service usage charges on this page currently show activity through approximately 12/25/2010 16:59 GMT.




Expand All Services Collapse All Services Printer Friendly Version		Printer Friendly Version
		Totals
 Amazon CloudFront View/Edit Service	Download Usage Report	0.68
 Amazon Simple Storage Service View/Edit Service	Download Usage Report	0.17
 AWS Data Transfer (excluding Amazon CloudFront) View/Edit Service		0.01
Taxes Estimated Taxes (Due January 1, 2011)		0.00
Total Charges due on January 1, 2011†		\$0.86

Figure: Monthly activity statement from Amazon Web Services

(Traditional) Web Application Hosting

- Web Hosting is pervasive
- Critical for e-commerce
- Needs to be highly available, and scalable
- Highly researched area (traffic forecasting, load balancers, server optimization, ...)
- and yet ... we have outages every now and then!
- An example of a traditional web application hosting (three-tiered architecture)
 - Presentation Layer
 - Application Layer
 - Persistence Layer

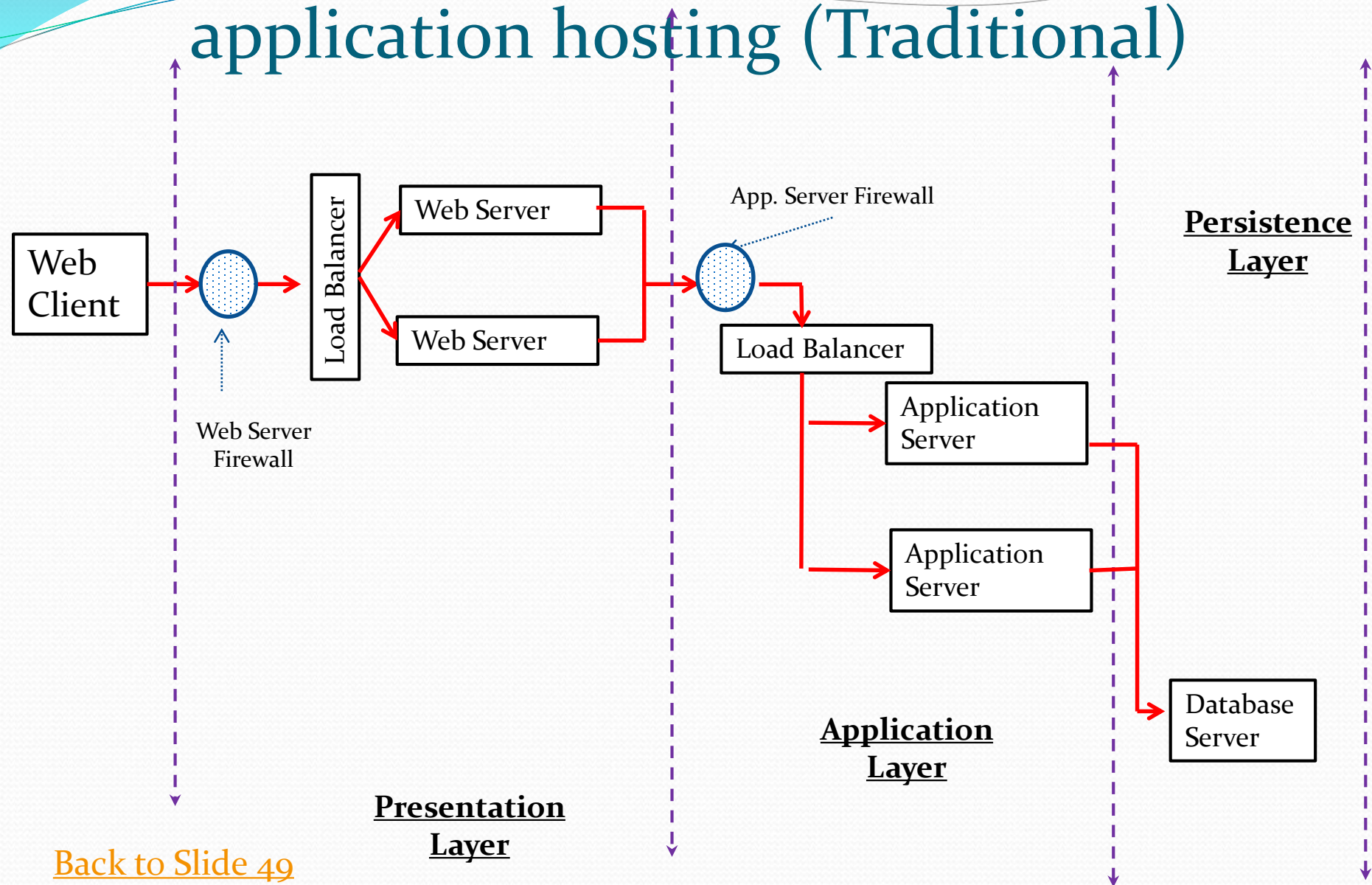
(Traditional) Web Application Hosting

- Web Client
 - Web browser to request a web page
- *Presentation Layer*
 - Web Server Firewall
 - External hardware or software firewall for checking and screening incoming messages to the Web Server
 - Web Server Load Balancer
 - Web traffic distribution to web servers to meet performance and availability requirements
 - Web Servers
 - Web servers for handling http:// or https:// requests from web clients
- *Application Layer*
 - Internal Firewall (or Application Server Firewall)
 - Internal firewall to protect any unauthorised access to application servers (or to the database servers) via web servers

(Traditional) Web Application Hosting

- *Application Layer (continued)*
 - Application Server Load balancer
 - Traffic distribution to application servers to meet performance and availability requirements
 - Application Servers
 - Application servers handling application specific request from web servers.
- *Persistence Layer*
 - Database Server
 - Database server running master and slave databases for providing persistence data storage
 - Backup Servers
 - Regular backups on CDs, tapes

Typical architecture of a three-tiered web application hosting (Traditional)



[Back to Slide 49](#)

Issues with traditional web hosting

- Upfront provisioning to handle peak load (corresponding increase in the need for Upfront capital investment)
 - Inability to handle (unexpected) traffic peaks
- Resource provisioning for testing, beta, pre-production environments

Wasted Capacity

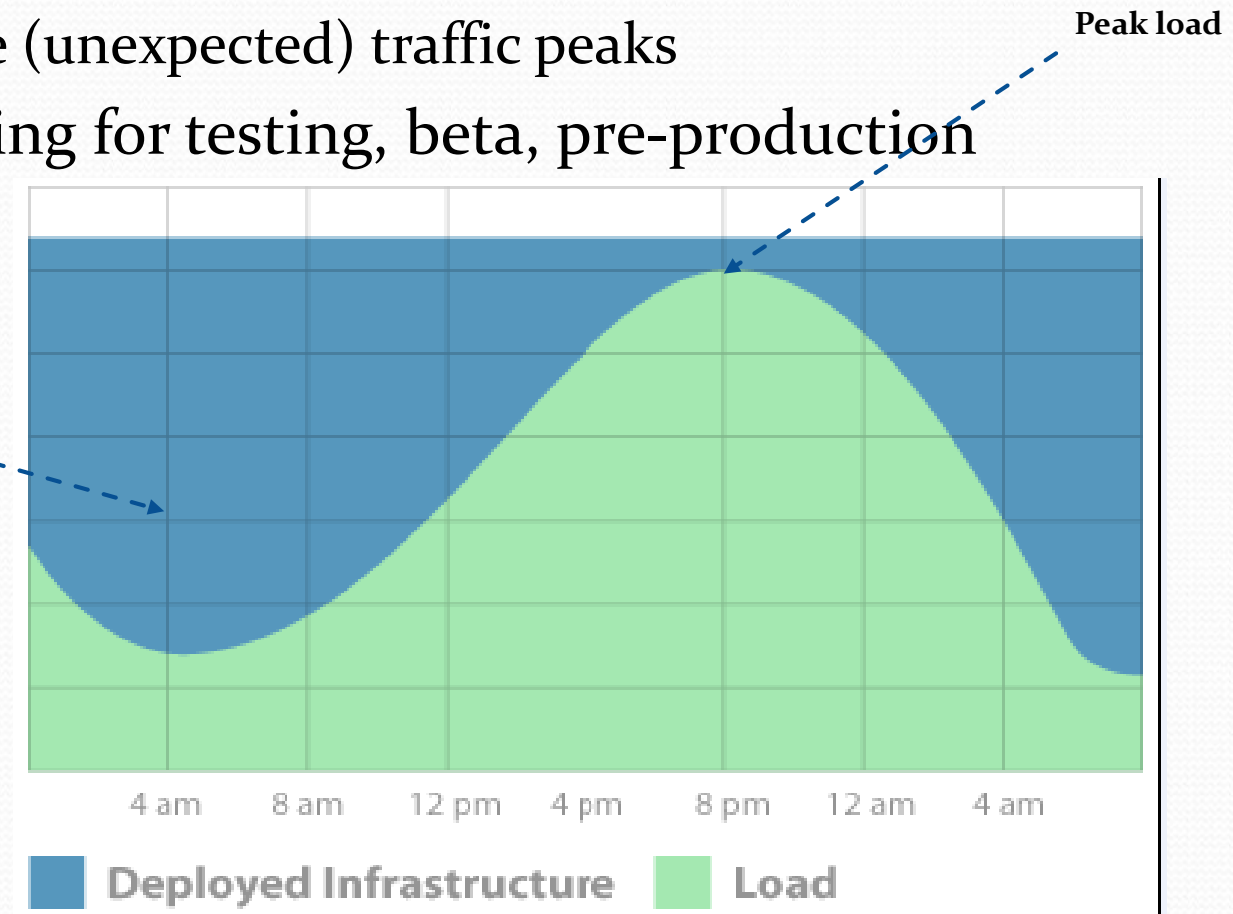


Figure: Typical provisioning pattern in traditional environments

Wasted Capacity in traditional web hosting

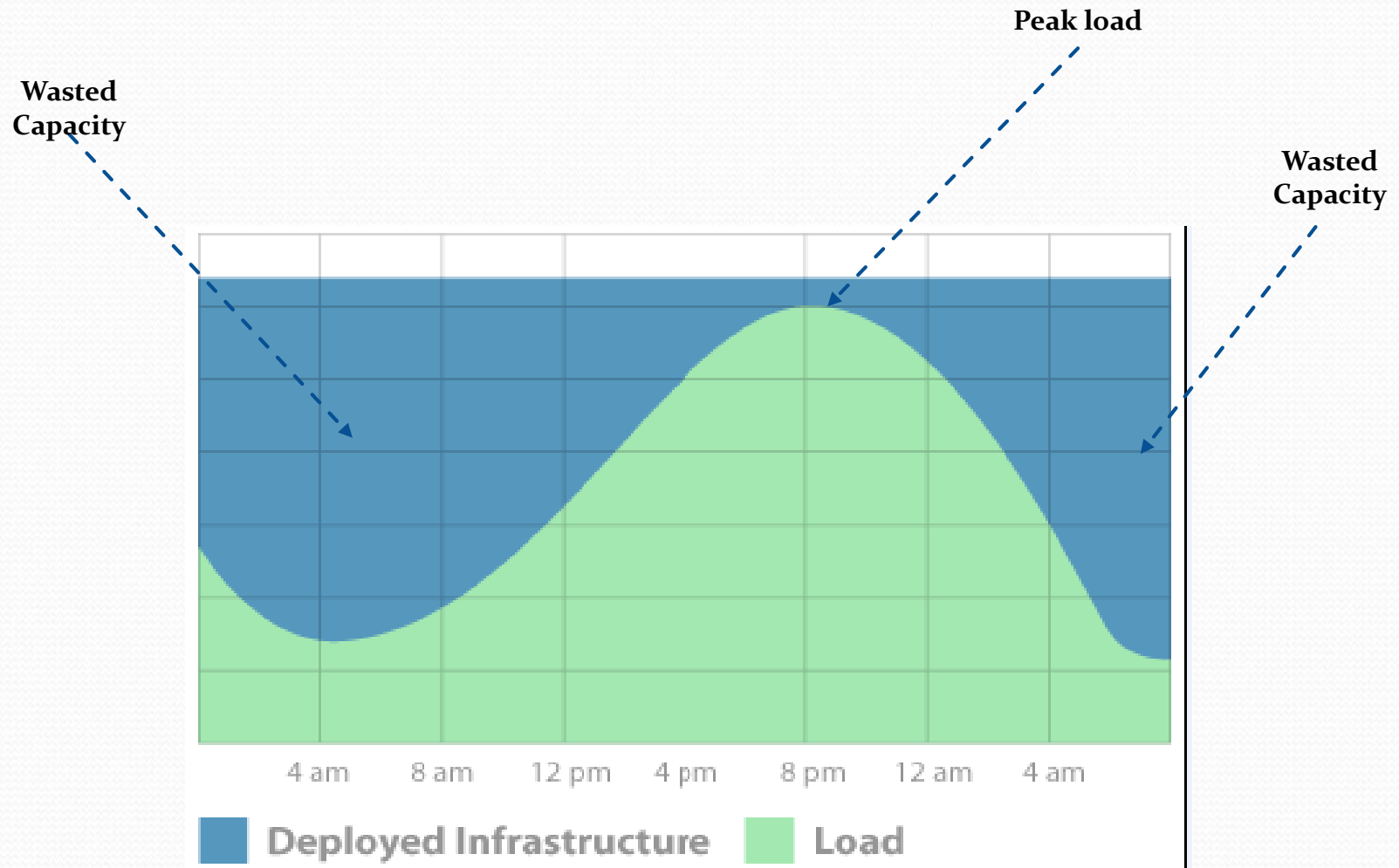


Figure: Wasted capacity during non-peak hours

Provisioning in Cloud-based Environments

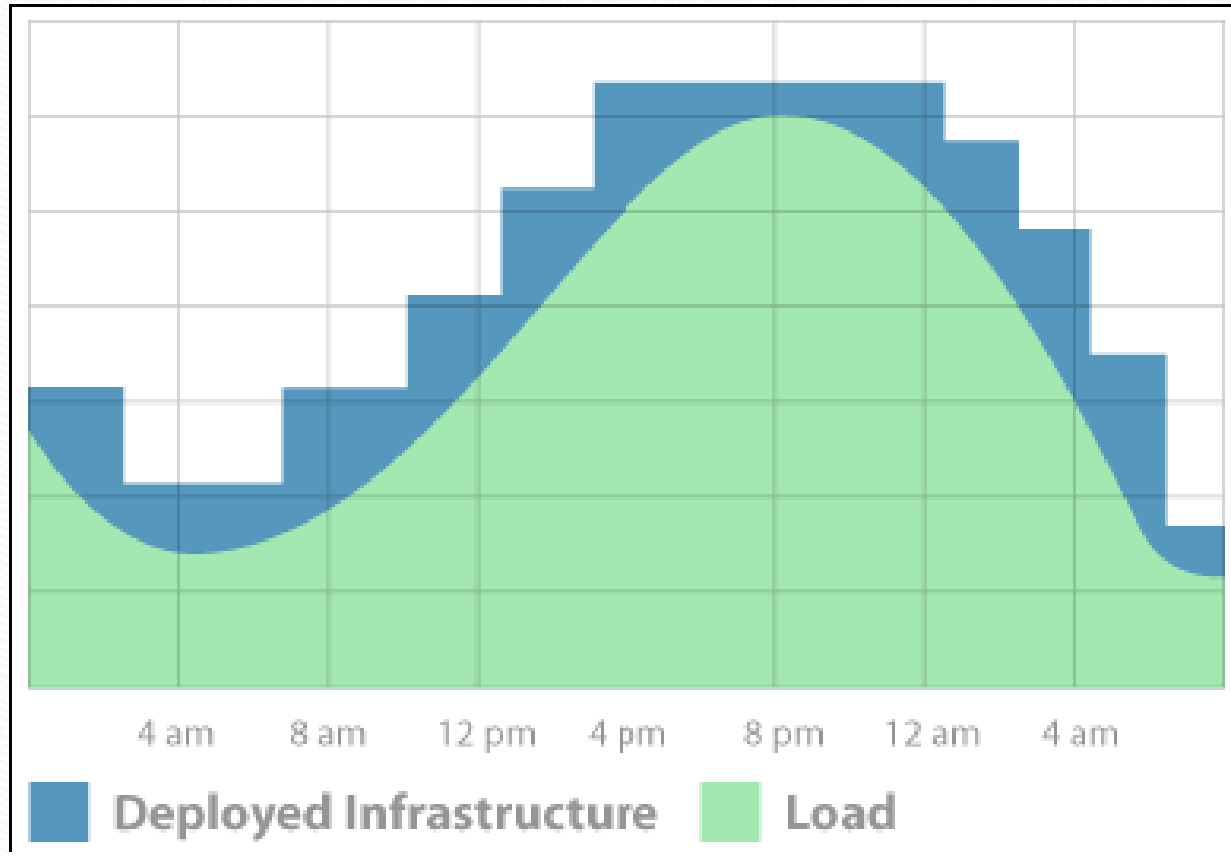


Figure: Provisioning in cloud-based environments

Understanding AWS-based web hosting

- Set up you an AWS account (<http://aws.amazon.com/>)
- Interact with AWS using an API
- Determine your computing requirements:
 - Number of web servers required
 - What is the accepted level of latency?
 - Who should access your web site and at what level?
 - Number of application servers required
 - Who do you wish to permit to access the application servers?
 - What are you storage requirements?
 - Do you have relational data? (Use Amazon RDS)
 - Do you have non-relational data? (Use Amazon S3)
 - Level of fault-tolerance required?
 - Choose OS requirements, processing capabilities of the chosen storage mechanism.

Understanding AWS-based web hosting

- Specify your requirements using the AWS API
- Start EC2 (elastic compute cloud) instances and Amazon S3/Amazon RDB instances using API
- Keep an eye on the performance of your system using Amazon Cloud Watch
- Dynamically request additional resources (if required)
- Pay for the invoice sent by Amazon

Components of AWS Web Hosting Solution

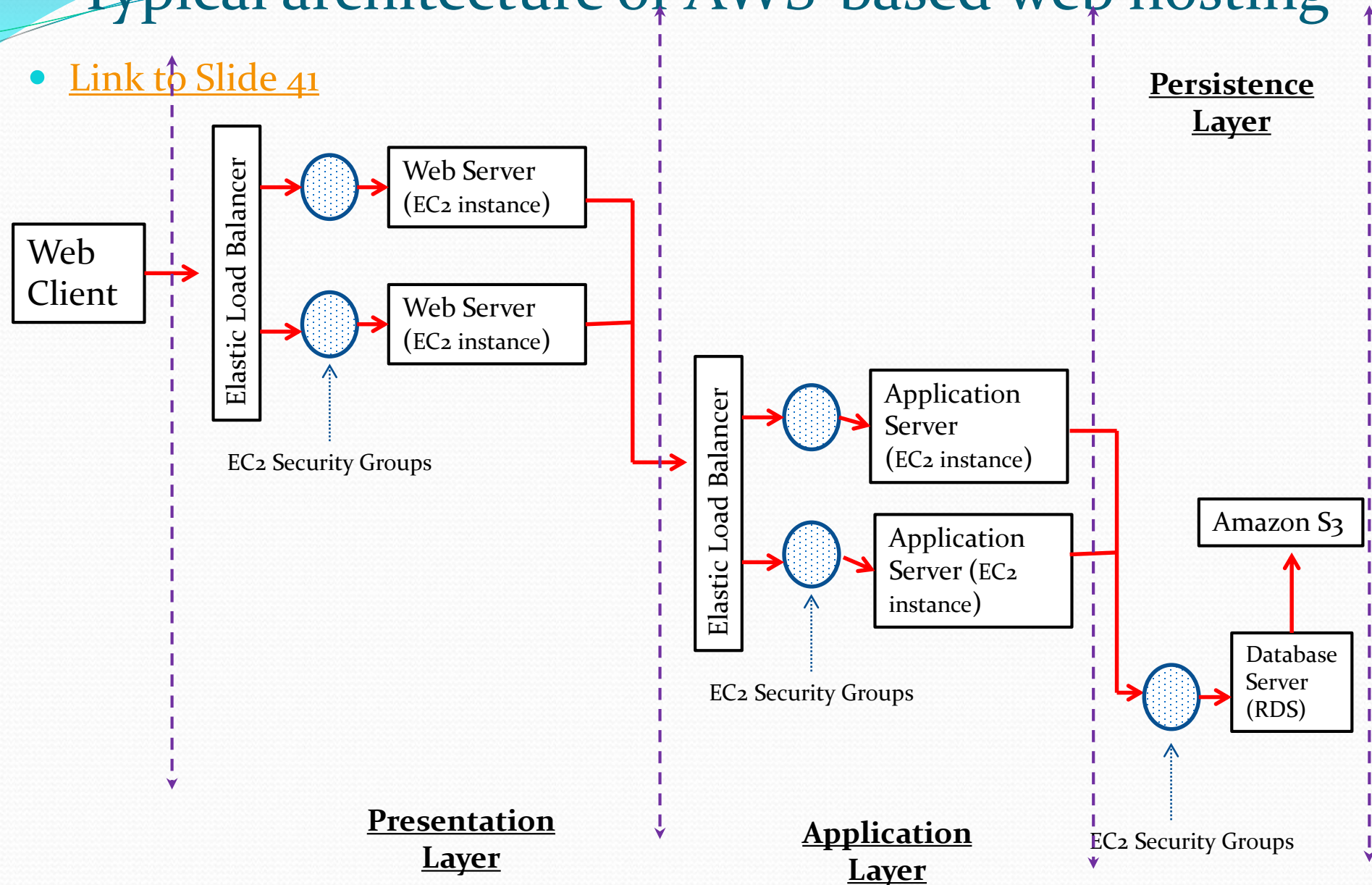
- Web Client
 - Web browser to request a web page
- *Presentation Layer*
 - Amazon EC2 instances for Web Servers
 - Each Web server is an EC2 instance for handling http requests.
 - Firewall
 - Firewall attached to each EC2 web server instance. Firewalls are implemented using Amazon EC2 security groups
 - Elastic Load Balancer
 - Web traffic distribution to the different Amazon EC2 instances to meet performance and availability requirements
- *Application Layer*
 - Amazon EC2 instances for Application Servers
 - Application server is a EC2 instance for handling application specific request from the web servers.

Components of AWS Web Hosting Solution

- *Application Layer (continued)*
 - Internal firewall
 - Internal firewall attached to each EC2 application server instance. This is to protect any unauthorised access to application servers from the web servers
 - Application Server Load balancer
 - This load balancer is used for the traffic distribution to the group of application servers to meet performance and availability requirements
- *Persistence Layer*
 - Amazon Relational Database Service (RDS)
 - Amazon RDS web service for relational data storage mechanism. Multiple Amazon RDS may be chosen for fault-tolerance
 - Amazon Simple Storage Service (Amazon S3)
 - Automatic and regular backups on S3 database – managed by AWS

Typical architecture of AWS-based web hosting

- [Link to Slide 41](#)



Additional References

1. M. Behrendt, B. Glasner, P. Kopp, R. Dieckmann, G., Breiter, S. Pappe, H. Kreger and A. Arsanjani. (2011). IBM Cloud Computing Reference Architecture 2.0.
2. P. Mell and T. Grance. (2011), The NIST Definition of Cloud Computing.
3. Tavis, M. (2010). Web Application Hosting in the AWS Cloud: Best Practices. Amazon Web Services
4. <http://aws.amazon.com/>