CS15-319: Cloud Computing

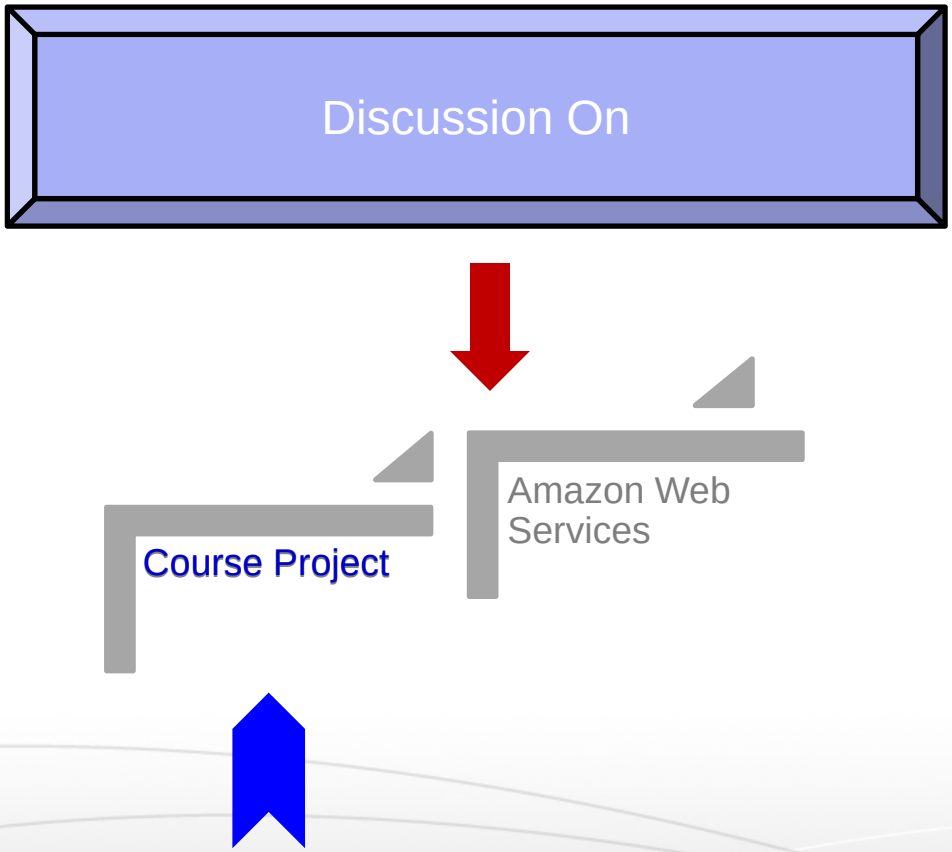
Lecture 3

Course Project and Amazon AWS

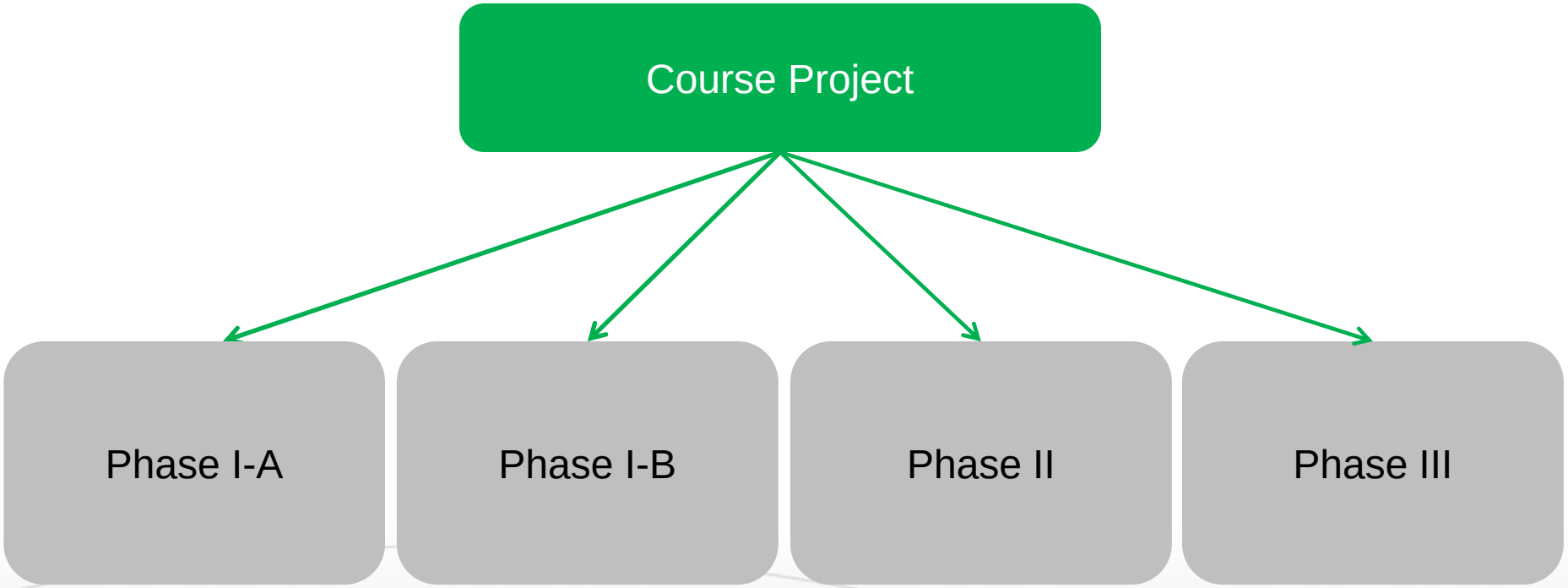
Majd Sakr and Mohammad Hammoud

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Lecture Outline

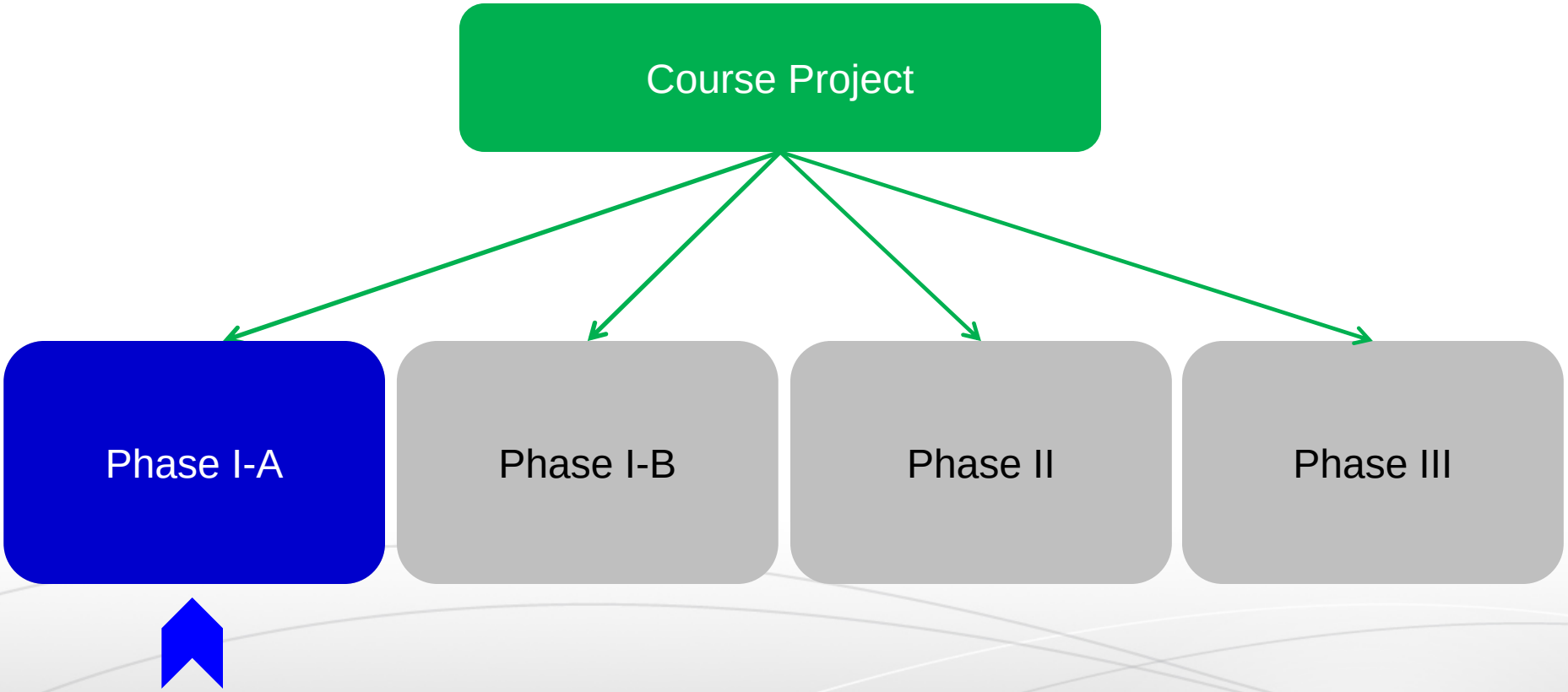


Course Project



جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Course Project

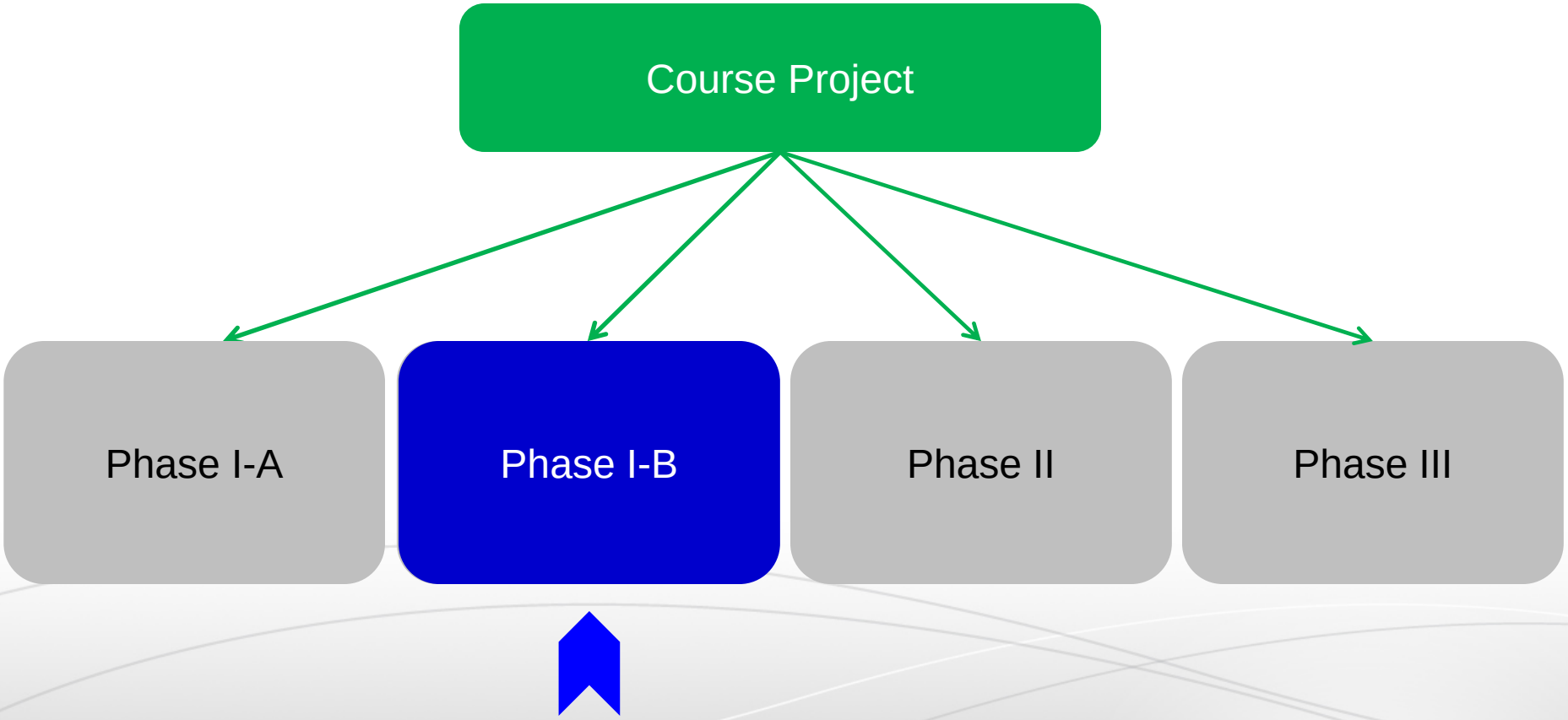


Introduction to Amazon AWS and Hadoop

Project Phase 1-A

- Get comfortable with the AWS web interface for VM management
- Provision a single instance manually and analyze instance performance
  - Run and analyze resource micro-benchmarks (CPU and Memory micro-benchmarks)
- Provision and utilize a Hadoop cluster on Amazon using Apache Whirr and Amazon EC2 and S3 web services
- Learn how to develop and deploy MapReduce jobs on your Amazon Hadoop cluster
- Perform scalability and sizing studies for MapReduce applications

Course Project



Project Management

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

What is Project Management?

- Project management is a series of flexible and iterative steps through which you identify:
  - Where you want to go
  - A reasonable way to get there
  - Specifics of what to do
  - Specifics of when to do what

- Project management consists of planning each part of your project via:
    - Stating your projected work
    - Tracking your work

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

The Statement of Work (1/2)

- The statement of work is a written document that clearly explains what the project is about

- Your statement of work should include the following sections:
  1. Purpose
  2. Objectives
  3. Constraints
  4. Assumptions

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

The Statement of Work (2/2)

1. The purpose section typically includes:

   - A Background Section: This is the first step of all good investigations. You typically provide a very brief review of literature

   - A Scope of Work Section: What will you do?– a brief statement describing the major work to be performed

   - A Strategy Section: How will you perform the work?

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

The Statement of Work (2/3)

2. Objectives are the end results achieved by the project
   2.1. Each objective should include a description of the desired outcome when the project is completed

3. Constraints are the restrictions on the project (e.g., wanting to complete all your experiments 2 weeks before the end of the semester)

4. Assumptions are the unknowns you assume in developing your plan– statements about uncertain information you will take as fact as you conceive, plan and perform the project

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Tracking the Work

- A project requires setting a schedule for a series of activities to be performed

- Your project schedule should:
  1. Include estimates of how long each activity will take
  2. Identify the order of experiments
  3. Show the relationship of experiments to each other (e.g., do they need to be done sequentially or can they be done in parallel)
  4. Identify bottlenecks

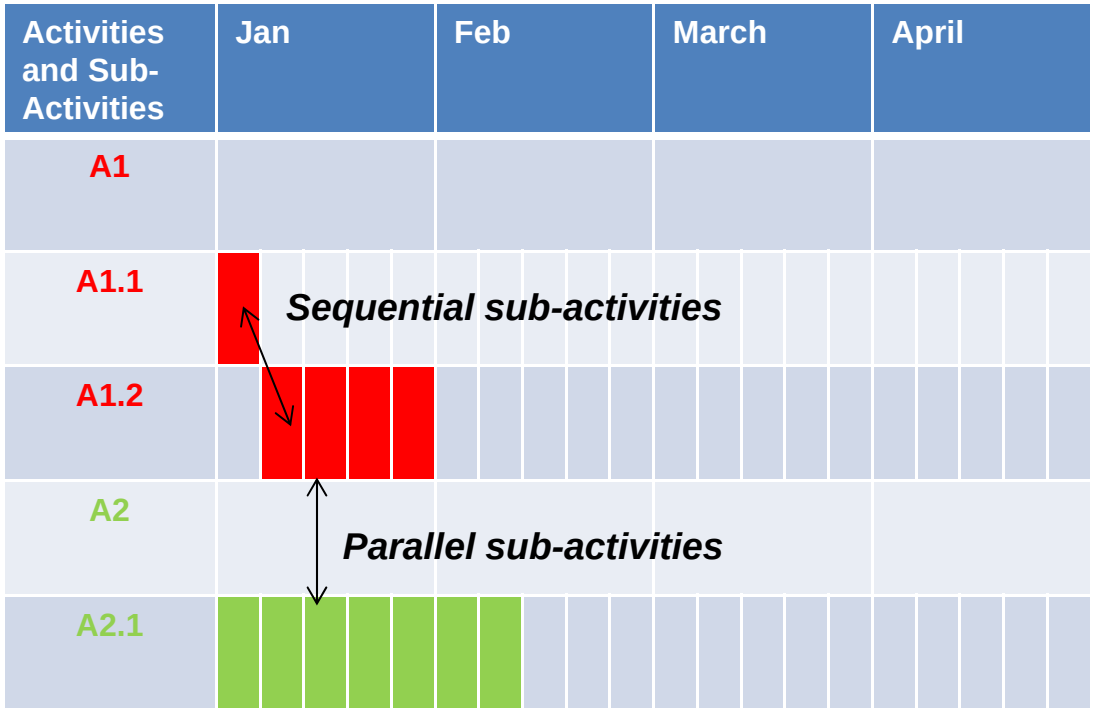
جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Tools for Tracking the Work

- Here are two tools that you can use to track your work:
  - Activities plan: A table showing activities (at a coarse-grain level) and their planned start and end dates
  - Gantt chart: A graph consisting of horizontal bars that depict the start date and duration for each sub-activity (at a fine-grain level)

Activity	Start Date	End Date
A1	--	--
A2	--	--

Activities Plan



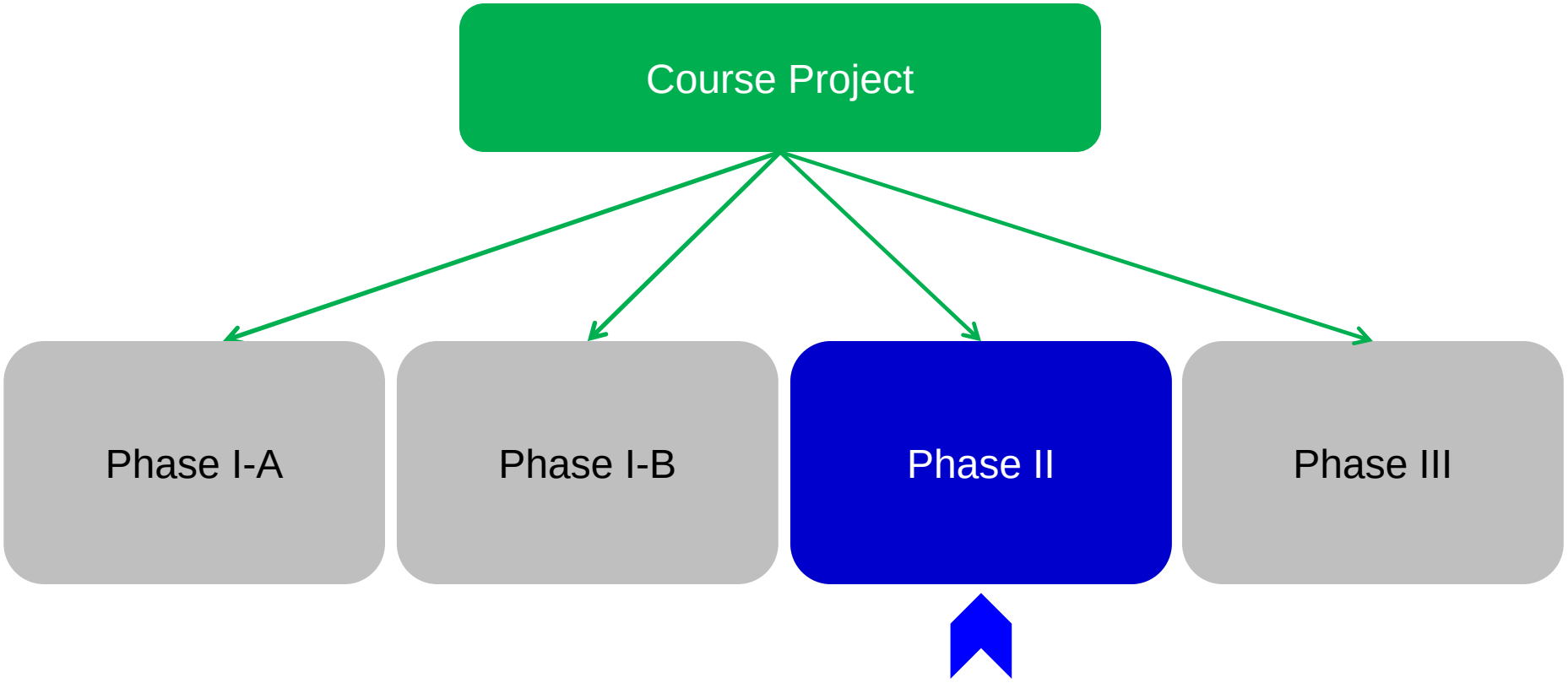
Gantt Chart

Keep in Mind…

- Project management is:
  - Not just a planning tool, it is also a training and communication tool
  - A process for identifying what to think about, not how to think about it
  - Merely a tool to help you plan and organize your work
    - It shouldn't become your work, bogging you down in complex manipulations or fancy tables/graphs that look impressive but don’t contribute to your progress in the project
- Effective project management demands that the components of a project be constantly monitored and revised with new information (especially as you make progress and get more experienced)

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Course Project



Project Phase II

- Develop a MapReduce application for a problem from following domains:

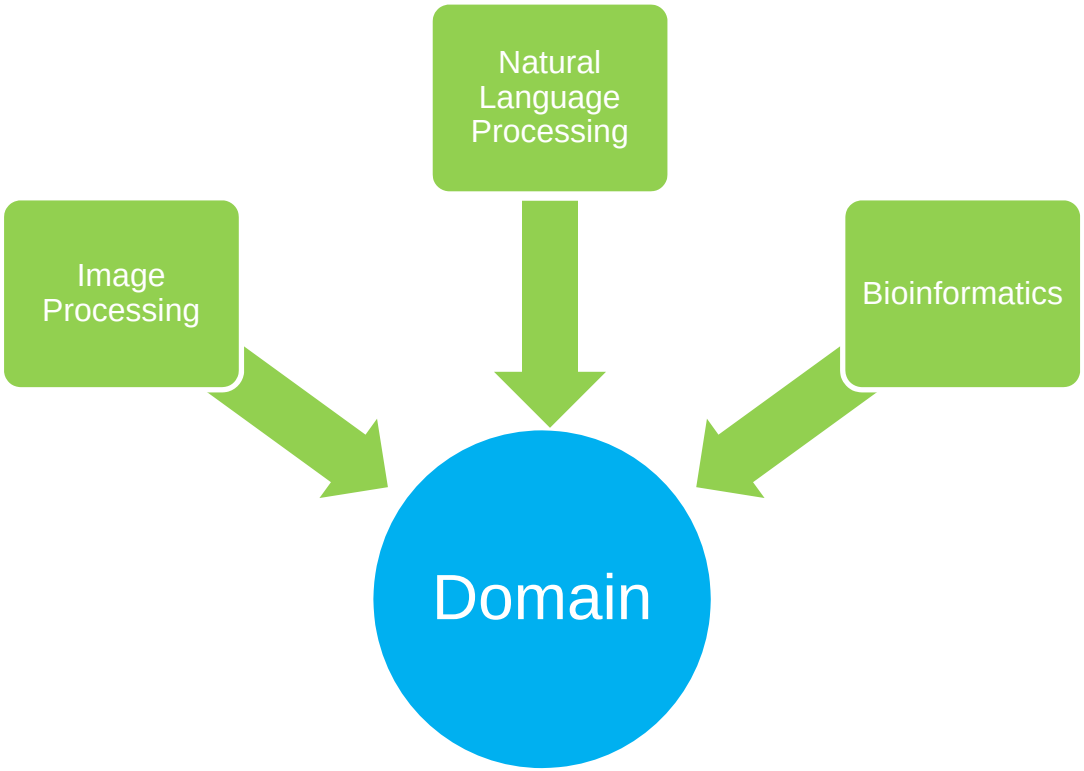
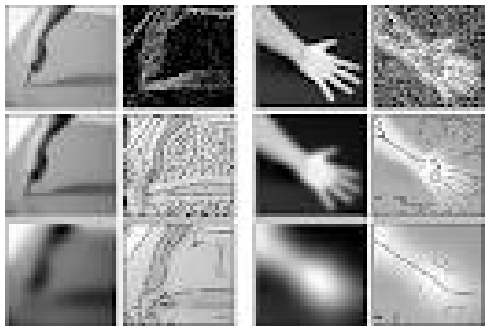


Image Processing

I. Edge Detection



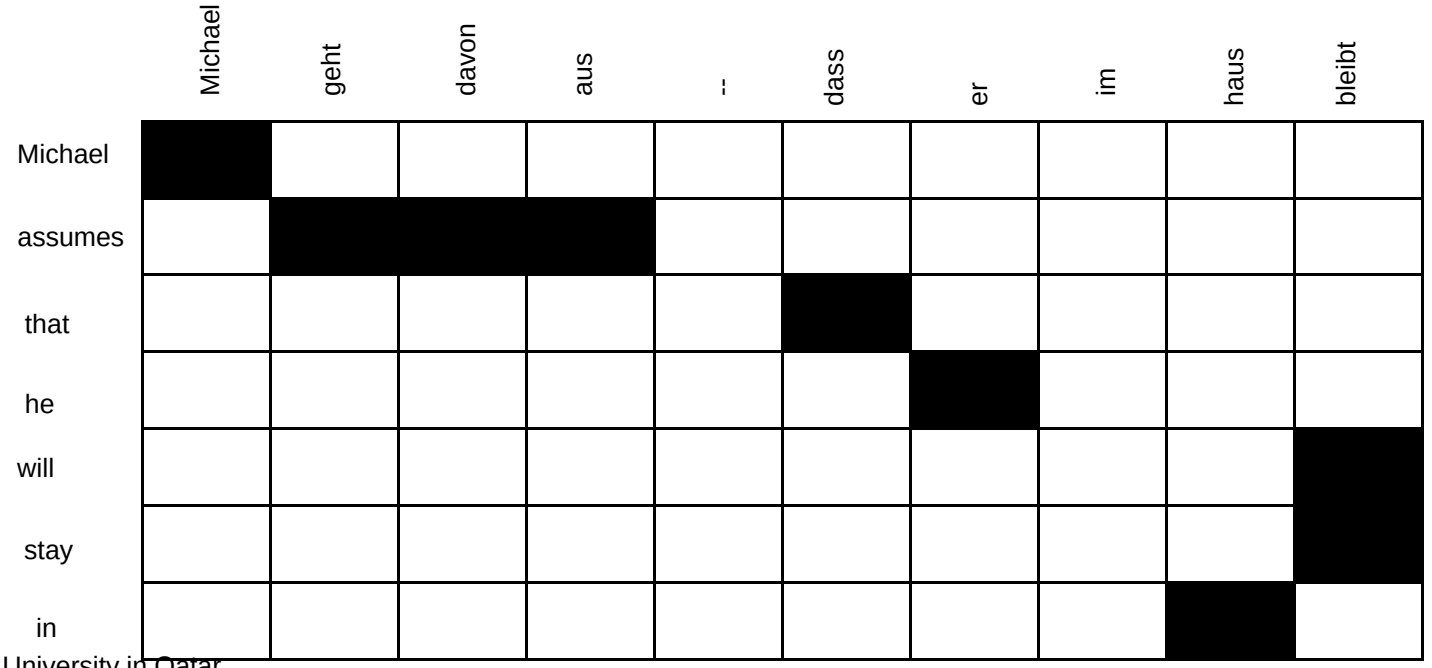
- Edge detection is a basic operation that is used in applications, such as image recognition
- Edge detection is representative of a class of applications that highlight or collect points on high contrast (e.g., edges of an object in an image) in the input
- One of the algorithms which you can implement is the Sobel algorithm

Natural Language Processing

II. Word Alignment

- You can implement word alignment, a problem of determining translational correspondence at the word level given a corpus of parallel sentences

- Automatic word alignment (Brown et al.,1993) has received extensive treatment in natural language processing as it is a vital component of all statistical machine translation approaches



Bioinformatics Benchmarks

III. Single Nucleotide Polymorphisms (SNPs)

- One of the applications that you can implement in Bioinformatics is SNP search, which uses the hill climbing search method

- SNPs are DNA sequence variations that occur when a single nucleotide is altered in a genome sequence

- Understanding the importance of many recently identified SNPs in human genes has become a primary goal of human genetics

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

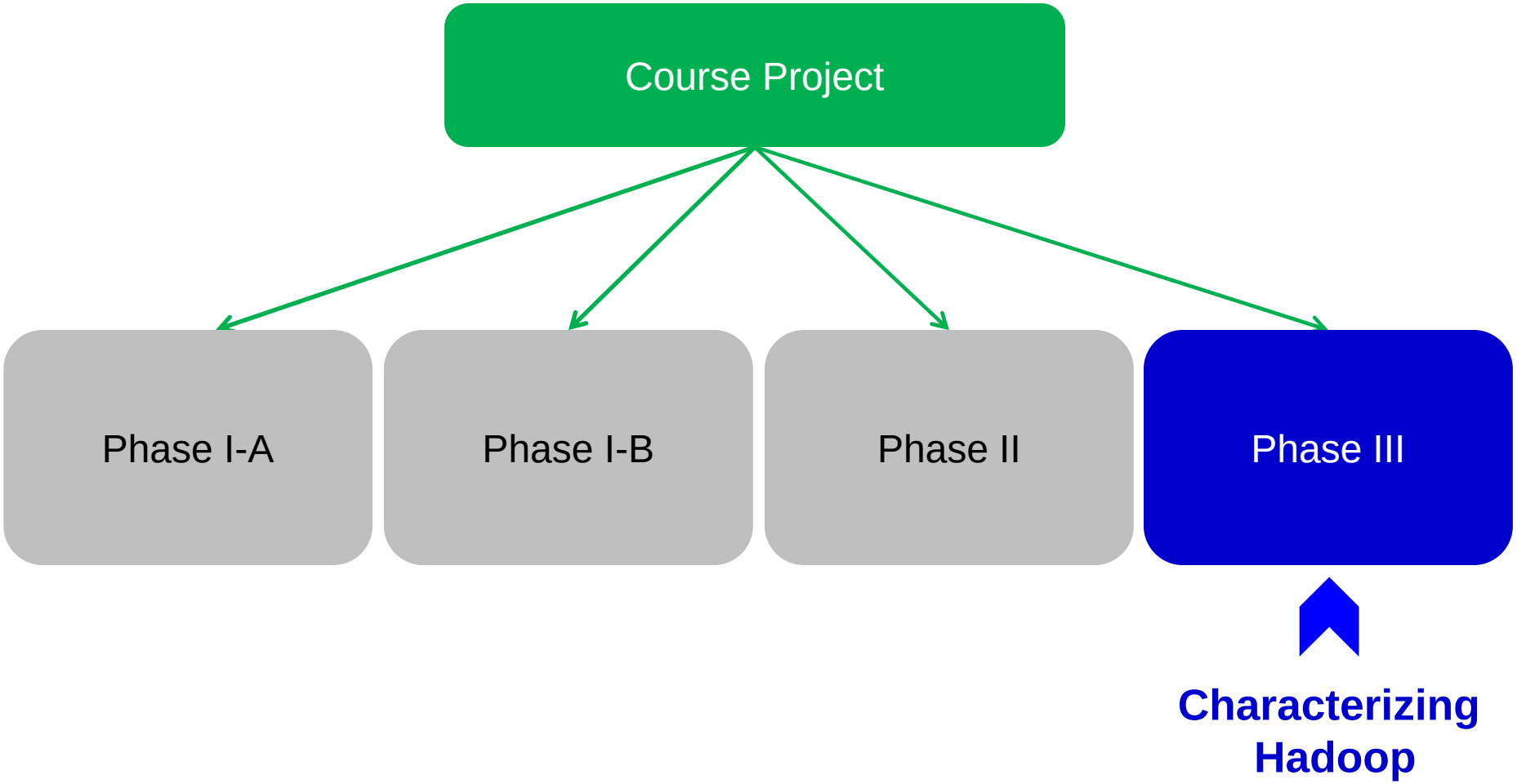
Amazon Deployments and Discussions

- You will deploy and test your application on the Amazon public cloud

- You will also provide a discussion on:
  - Your experience in applying MapReduce to a real problem
  - Your insights concerning the performance of a real MapReduce application
  - Your thoughts on the applicability of MapReduce to your selected problem
  - Your recommendations regarding the usage of MapReduce for algorithms similar to your selected one

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Course Project



Workload Characterization

- In order to capture the processing and the I/O characteristics of MapReduce applications, we have to perform what is referred to as workload characterization
- Workload characterization is a crucial component of any performance analysis process
- In this part of the project, you will characterize Hadoop using the MapReduce application you developed in Phase II

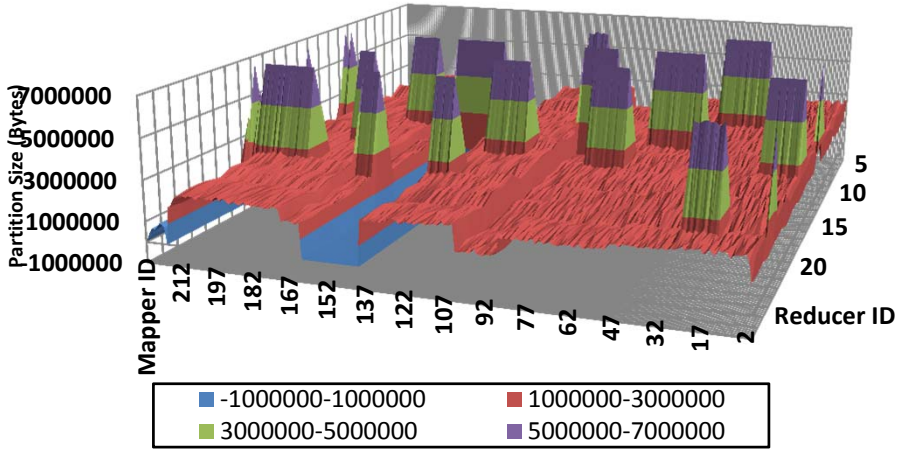
Why Workload Characterization?

- Through your workload characterization you can:
  - Provide insights into the Hadoop framework bottlenecks and intricacies and probably trigger framework improvements
  - Provide a quantitative foundation for MapReduce researchers and developers seeking validations for their hypotheses against real-world workloads
  - Help researchers and organizations adopt sound experimental methodology via:
    - Fine-tuning application parameters and achieve performance targets for applications similar to yours

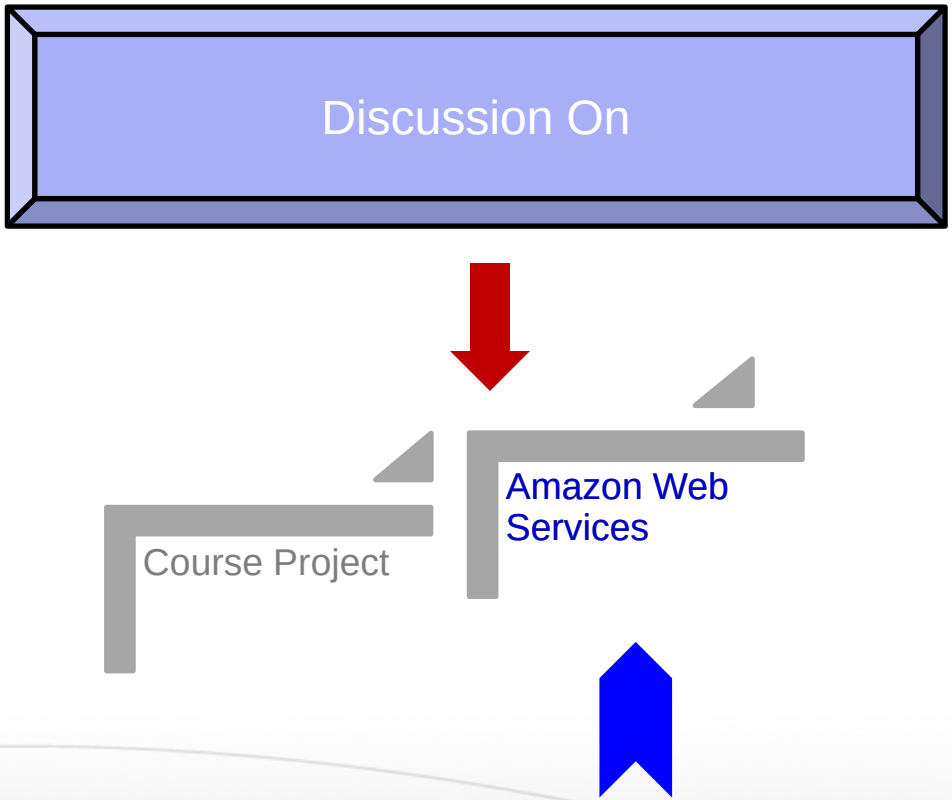
Characterization Dimensions

- Among the dimensions that you can use to pursue your workload characterization are:

1. Data Patterns
2. Concurrency
3. Phase Timelines
4. Dataset Types
5. Network Traffic Patterns
6. System Resource Utilization
7. Communication to Computation Ratio



Lecture Outline



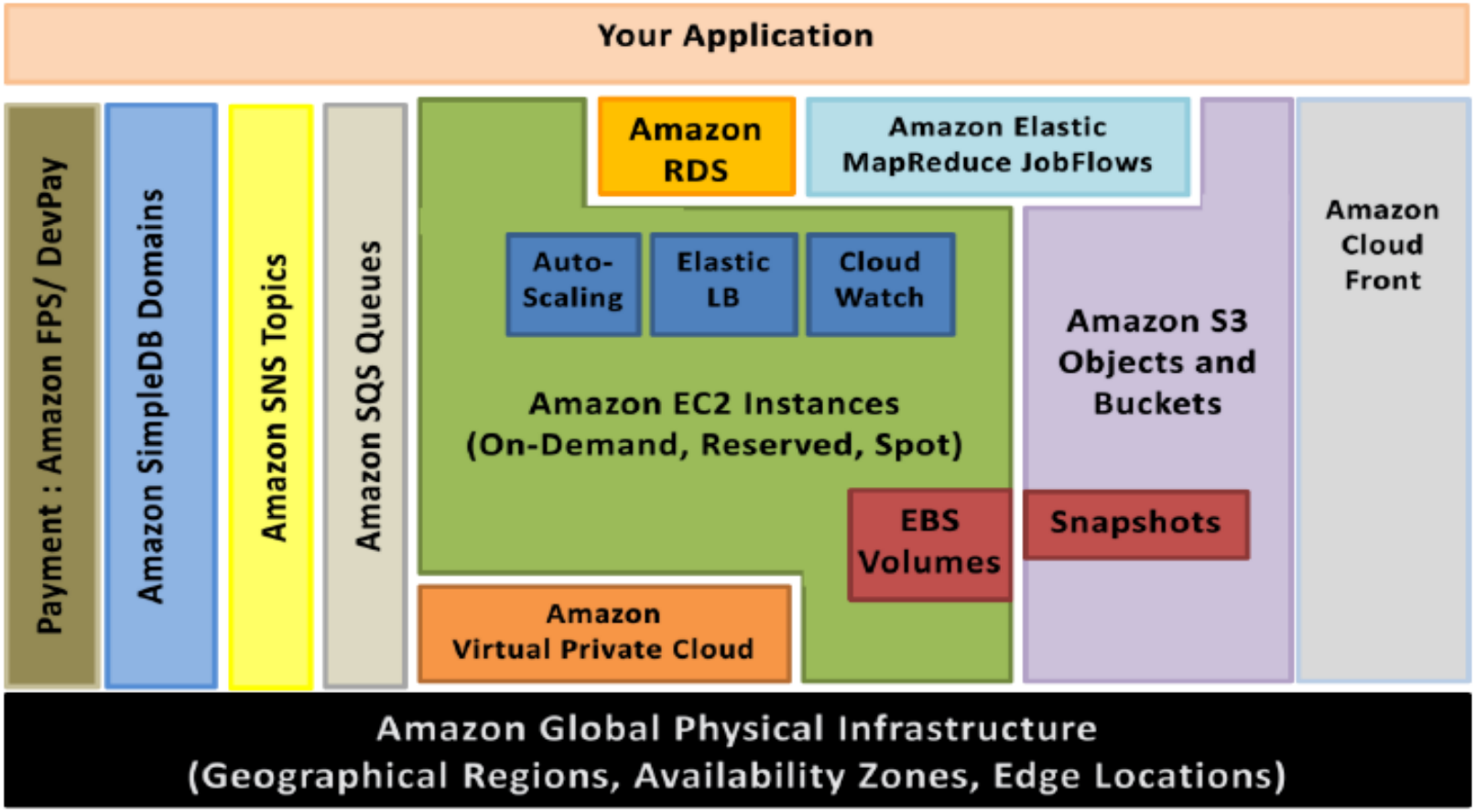
Amazon Web Services (AWS)

- AWS is a collection of remote computing services that together make up a cloud computing platform

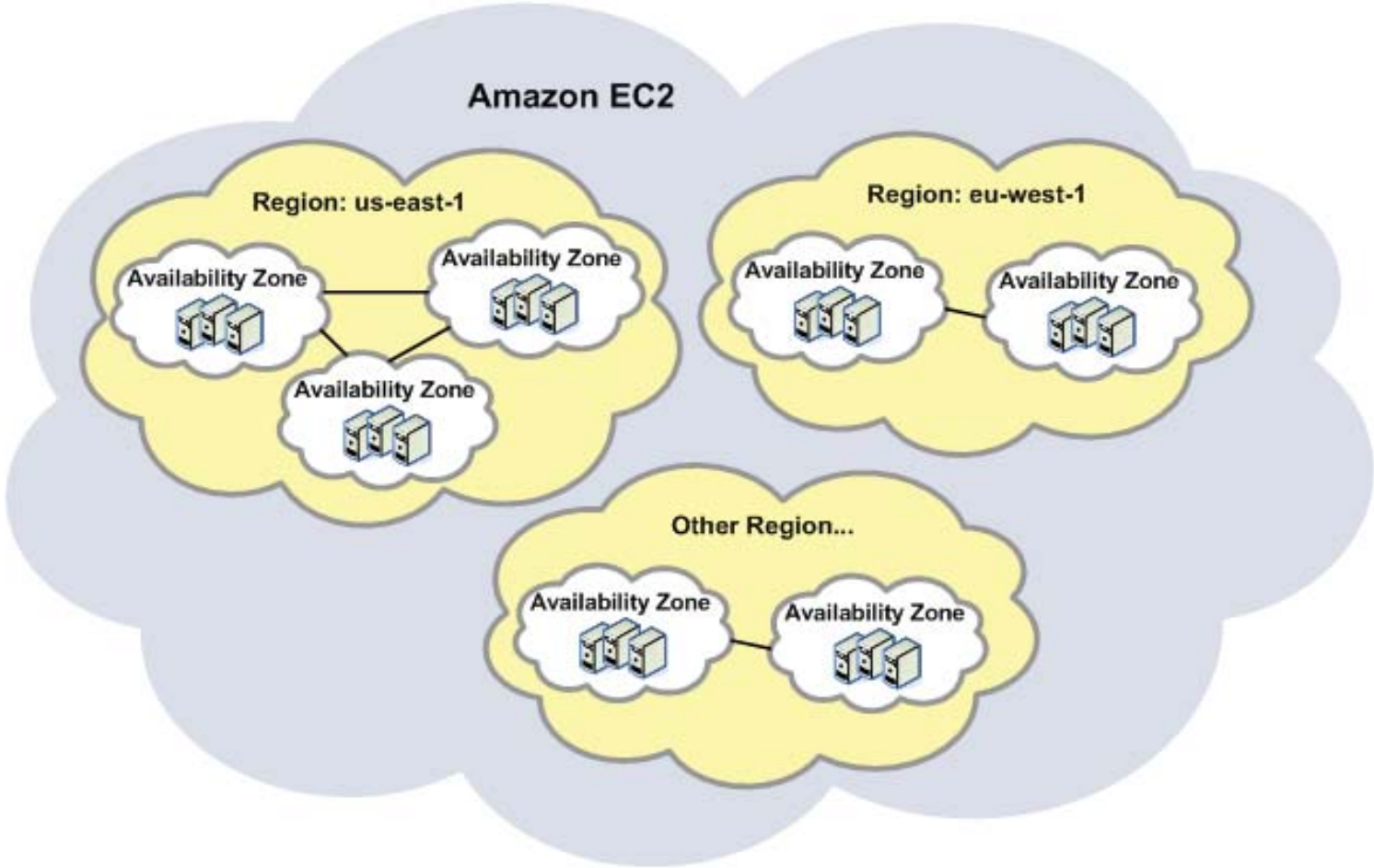
- AWS offers the following products:
  - Compute (e.g., EC2)
  - Storage (e.g., S3)
  - Database (e.g., RDS)
  - Networking (e.g., VPC)

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

AWS Ecosystem



Regions and Availability Zones



Regions and Availability Zones

- Amazon EC2 provides the ability to place instances in multiple locations
- Amazon EC2 locations are composed of Availability Zones and Regions
- Regions are dispersed and located in separate geographic areas (US, EU, etc.)
  - You can design your application to be closer to specific customers or to meet legal or other requirements
- Availability Zones are distinct locations within a Region
  - You can protect your applications from the failure of a single location

Amazon's Global Datacenters

Amazon EC2 is currently available in eight regions: US East (Northern Virginia), US West (Oregon), US West (Northern California), EU (Ireland), Asia Pacific (Singapore), Asia Pacific (Tokyo), South America (Sao Paulo), and AWS GovCloud

Amazon EC2

- Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud
- It has been offered in 2006
- It is deemed the first real cloud computing product
- You can rent EC2 instances by the hour
- Many *instance types* are available in Amazon

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Amazon EC2

- Amazon EC2 presents a true virtual computing environment, allowing you to:
  - Use web service interfaces to launch instances with a variety of operating systems (bundled into AMIs)
  - Load your instances with your custom application environment
  - Manage your network's access permissions
- Amazon EC2 reduces the time required to obtain and boot new server instances to minutes
  - This allows you to quickly scale capacity as your computing requirements change

EC2 Instance Models

1. On-Demand Instances
   - Pay-by-the hour
   - Start and stop as you wish

2. Reserved Instances
   - Pay-by-the-year
   - Reserved for you, can start-or-stop as needed

3. Spot Instances
   - Bid for unused EC2 capacity
   - Mention your *Spot Price* and if the market rate is less than your Bid, you get your instance
   - Instance automatically terminates if your Spot Price becomes less than the current market rate

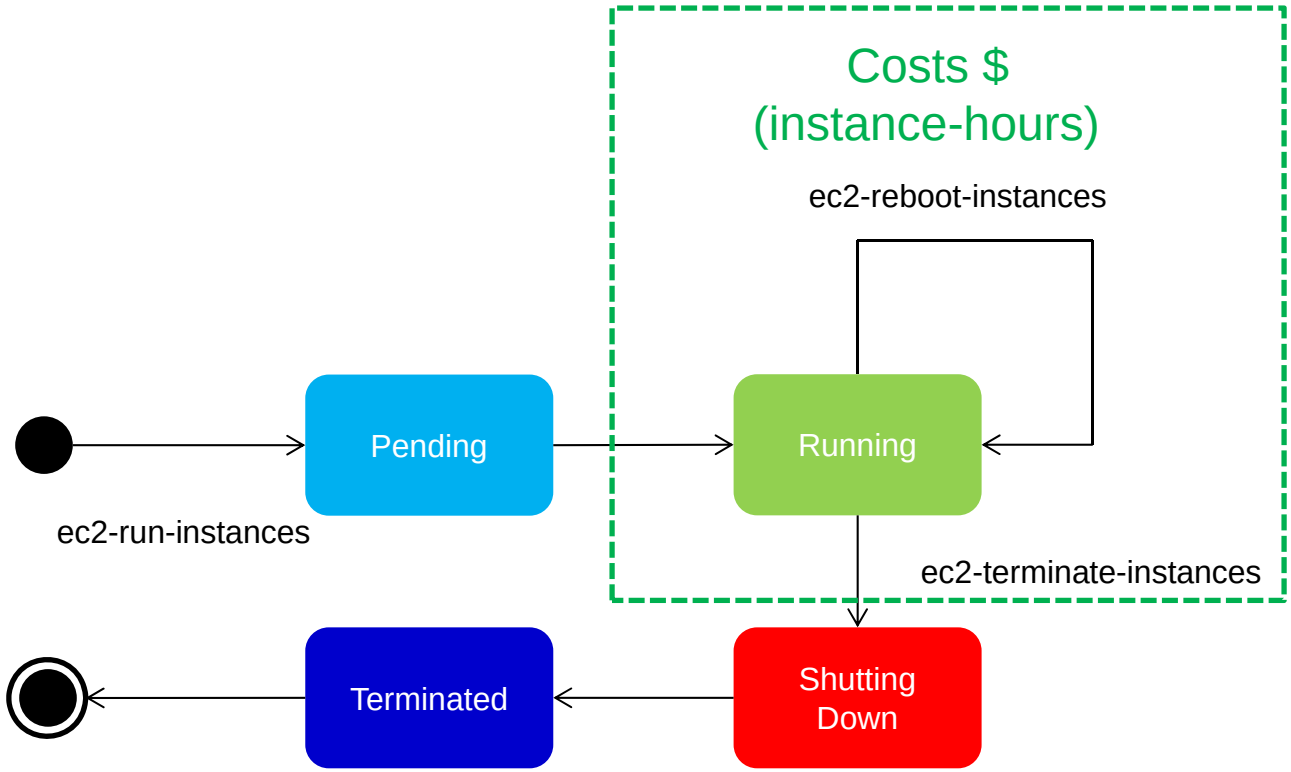
EC2 Instance Parameters

- CPU Power
  - Elastic Compute Unit (ECU) – Defined by amazon as the equivalent CPU capacity of a 1.0-1.2 GHz 2007 Opteron/Zeon processor
- Memory – GB
- I/O performance
  - Low/Moderate/High
  - High-end instances have 10 Gigabit Ethernet

EC2 Instance Types

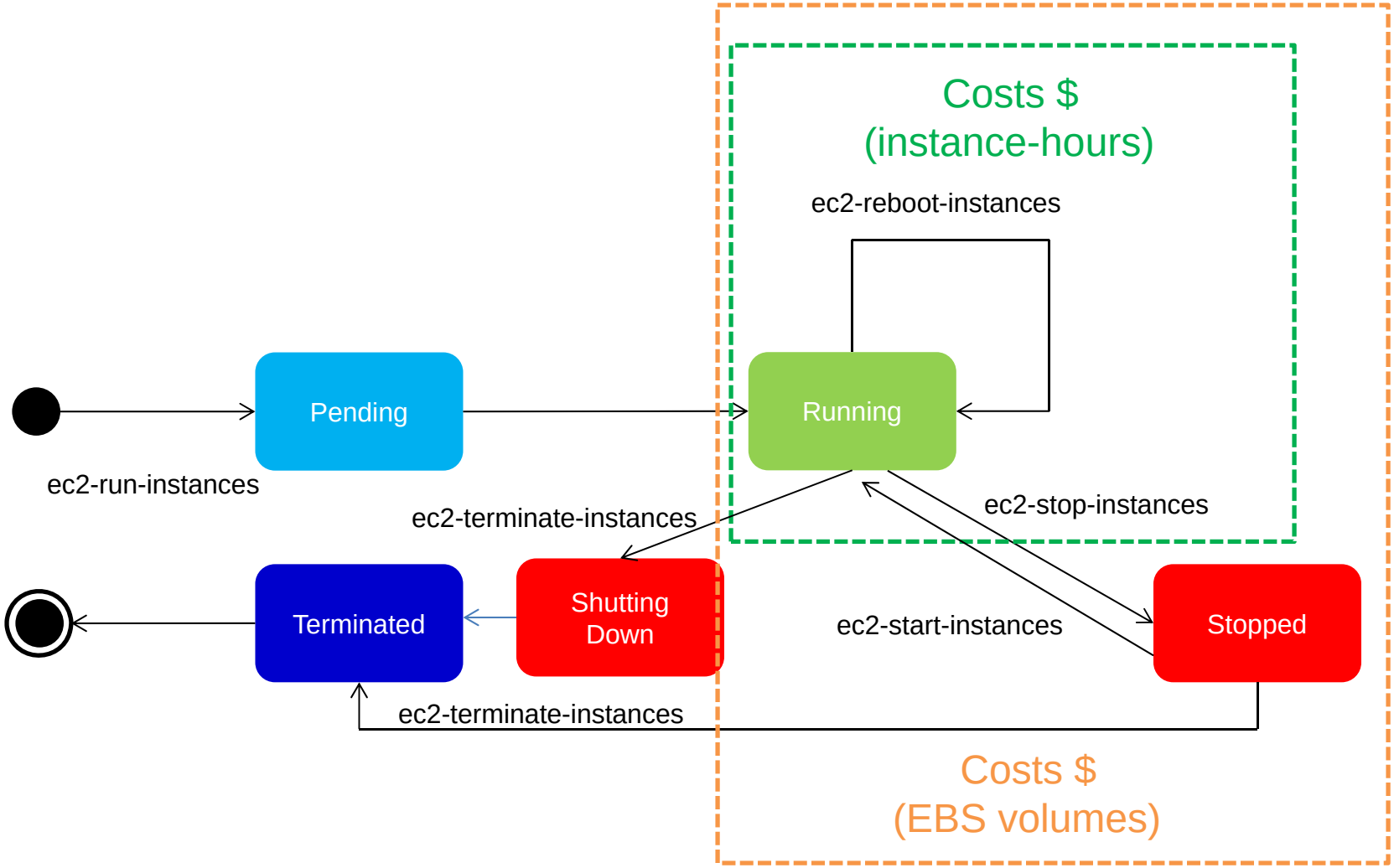
Instance Name	Mem (GB)	CPU Capacity	Disk (GB)	Platform	On-Demand Pricing / hour (Linux)
Micro	0.59	Upto 2 ECUs	--	32/64	$0.02
Small	1.7	1 core - 1 ECU	160	32	$0.085
Large	7.5	2 cores, 2 ECUs each	850	64	$0.34
Extra Large	15	4 cores, 2 ECUs each	1690	64	$0.68
High-Mem Extra Large	17.1	2 cores, 3.25 ECUs each	420	64	$0.50
High-Mem Double Extra Large	34.2	4 cores, 3.25 ECUs each	850	64	$1.00
High-Mem Quad Extra Large	68.4	8 cores, 3.25 ECUs each	1690	64	$2.00
High CPU Medium	1.7	2 cores, 2.5 ECUs each	350	32	$0.17
High CPU Extra Large	7	8 cores, 2.5 ECUs each	1690	64	$0.68
Cluster Compute Quad XL	23	33.5 ECUs	1690	64	$1.30
Cluster GPU Quad XL	22	33.5 ECUS + 2x GPUs	1690	64	$2.10

EC2 Instance Lifecycle (1/2)



Standard EC2 Instance
(Once terminated, data is lost)

EC2 Instance Lifecycle (2/2)



EBS-backed EC2 Instance (Data persists as long as EBS volume is Intact)

Amazon EC2 Characteristics

Characteristic	Description
Elastic	Amazon EC2 enables you to increase or decrease capacity within minutes, not hours or days
Completely Controlled	You have complete control of your instances
Flexible	You have the choice of multiple instance types, operating systems, and software packages
Extensible	Amazon EC2 works in conjunction with Amazon S3, Amazon RDS, Amazon SimpleDB and Amazon SQS to provide a complete solution for computing, query processing and storage across a wide range of applications
Reliable	Amazon EC2 offers a highly reliable environment. The Amazon EC2 Service Level Agreement commitment is 99.95% availability for each Amazon EC2 region
Secure	Amazon EC2 provides numerous mechanisms for securing your compute resources (e.g., Amazon VPC)
Inexpensive	You pay a very low rate for the compute capacity you actually consume

Amazon Simple Storage Service

- Amazon S3 can be used to store and retrieve any amount of data, at any time, from anywhere on the web
- You can write, read, and delete *objects* into S3 containing from 1B to 5TBs of data each (the number of objects you can store is unlimited)
- Each object is stored in a *bucket* and a bucket can be stored in one of several regions
- Objects stored in a region never leave the region unless you transfer them out

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Amazon Elastic Block Store

- Amazon Elastic Block Store (EBS) offers persistent storage for Amazon EC2 instances
  - Amazon EBS volumes provide off-instance storage that persists independently from the life of an instance

- Amazon EBS provides the ability to create point-in-time consistent snapshots of your volumes that are then stored in Amazon S3, and automatically replicated across multiple available zones

- These snapshots:
  - Can be used as the starting point for new Amazon EBS volumes
  - Can protect your data for long term durability
  - Can be easily shared with co-workers and other AWS developers

Amazon CloudWatch

- Amazon CloudWatch provides monitoring for AWS cloud resources and applications running on AWS

- It provides you with visibility into resource utilization, operational performance, and overall demand patterns—including metrics such as CPU utilization, disk reads and writes, and network traffic

- Amazon CloudWatch provides a reliable, scalable, and flexible monitoring solution that you can start using within minutes
  - You no longer need to set up, manage, or scale your own monitoring systems and infrastructure

Other AWS Products / Services / Features

- Compute
  - Auto Scaling
  - Elastic Load Balancing
  - Amazon Elastic MapReduce (EMR)
- Database
  - Amazon Relational Database Service (RDS)
  - Amazon SimpleDB
  - Amazon DynamoDB
- Networking
  - Amazon Virtual Private Cloud (VPC)
  - Amazon Route 52
  - Amazon Direct Connect
- Messaging
  - Amazon Simple Queue Service (SQS)
  - Amazon Simple Email Service (SES)
  - Amazon Simple Notification Service (SNS)
- Management and Deployment
  - Amazon Identity and Access Management (IAM)
  - Amazon Cloudwatch
  - Amazon Elastic Beanstalk
  - Amazon CloudFormation

…and much more