CS15-319 / 15-619
Cloud Computing

Recitation 5

Feburary 11th & 13th, 2014

Quiz2 Bugs

- Question 14
  - The question did not specify read/write ratio
  - Read speed: 12x (theoretical)
  - Write Speed: 6x (theoretical)
- Question 22, part3
  - Rounding issue
- Question 23
  - Some students considered leap years

We will manually grade these questions

Project2.1 Checkpoint Bugs

- Question 4 and Question 5 will be manually graded
  - This happens due to AWS variations in performance
  - We will manually grade Q4 & Q5
  - Please be patient, we will inform you once this process is completed

Announcements

- Do not cheat
  - We will find out
  - The penalties are severe
- Tag your instances
- Provide feedback on OLI
- Post on Piazza:
  - Private: a grading bug
  - Public: general questions
    - Search Piazza and the web before posting

Announcements

- Monitor AWS expenses regularly
  - EMR cost is “on top of” the EC2 cost of instance and EMR cost is fixed per instance type per hour
    - for example, m2.4xlarge EMR cost is $0.42 on-top-of the spot pricing ($0.14)
  - Suggestions
    - Terminate your instance when not in use
      - stop still costs money!
    - Use smaller instances to test your code
    - Use small sample dataset in EMR
    - Decrease the total number of requests when firing up the benchmark

CloudWatch Billing Alert

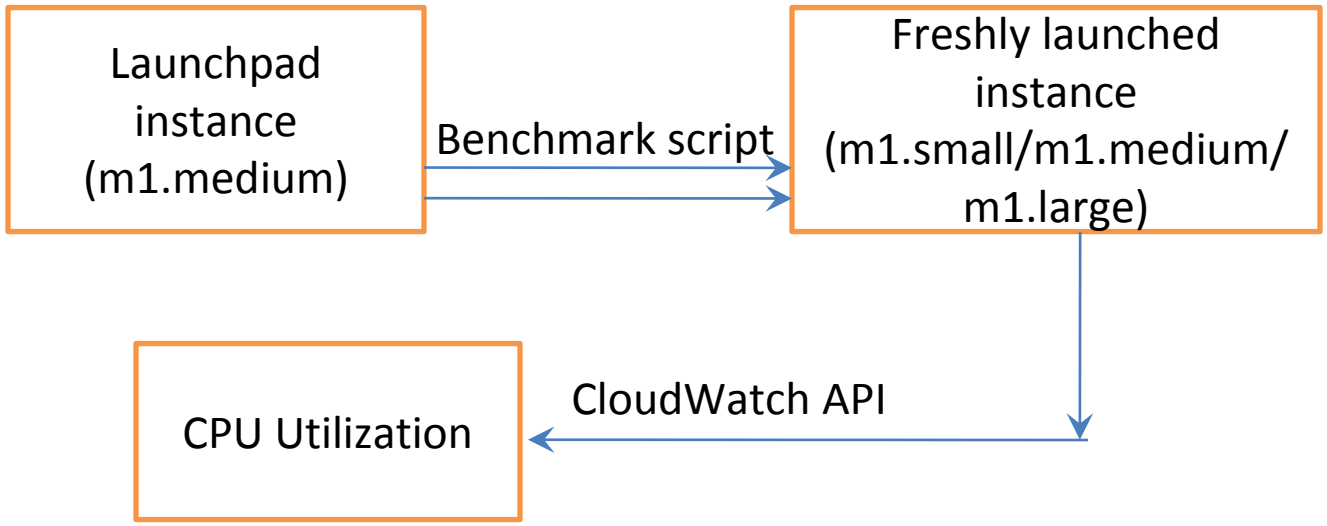
- You can set up an alert to be notified automatically via e-mail when estimated charges reach a threshold that you specify
- You can use up to 10 alarms and 1,000 e-mail notifications free each month
- Demo
- Billing Alarm HOWTO: https://piazza.com/class/hq77w6lddnb6wd?cid=502

Last Week

- Content
  - Unit 2: Data Centers
  - Quiz 2 completed
- EC2 and CloudWatch APIs
  - Amazon Command Line
  - AWS SDK for Java
  - AWS SDK for Python
- Vertical Scaling
  - Instance Capacity

Reflection on Last Week

Metric: bandwidth (avg responses/sec), CPU Utilization

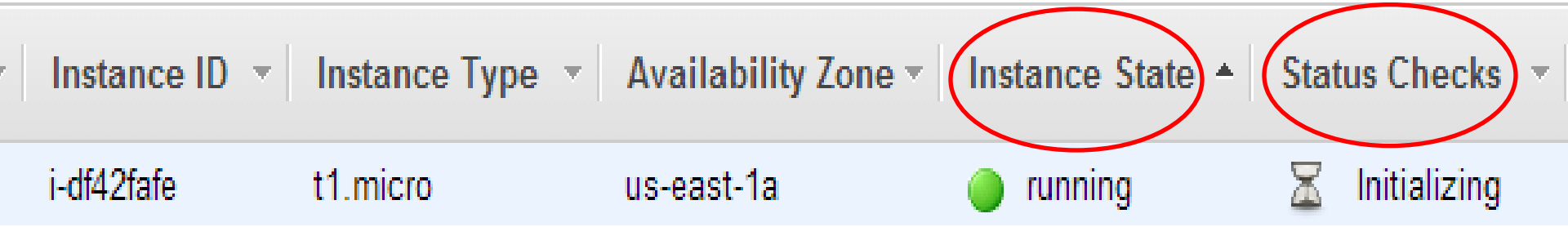


Piazza Questions

- mon-get-stats
  - Refused: The security token included in the request is invalid
- Solution:
  - Each time you use the CloudWatch tools (or Amazon EC2 CLI tools) with your instance, you must provide your identity
- How to make sure the instance is running?
  - DescribeInstaceRequest correct
  - instance.getState().getName() wrong

Piazza Questions

- And…you still need to wait for it to initialize



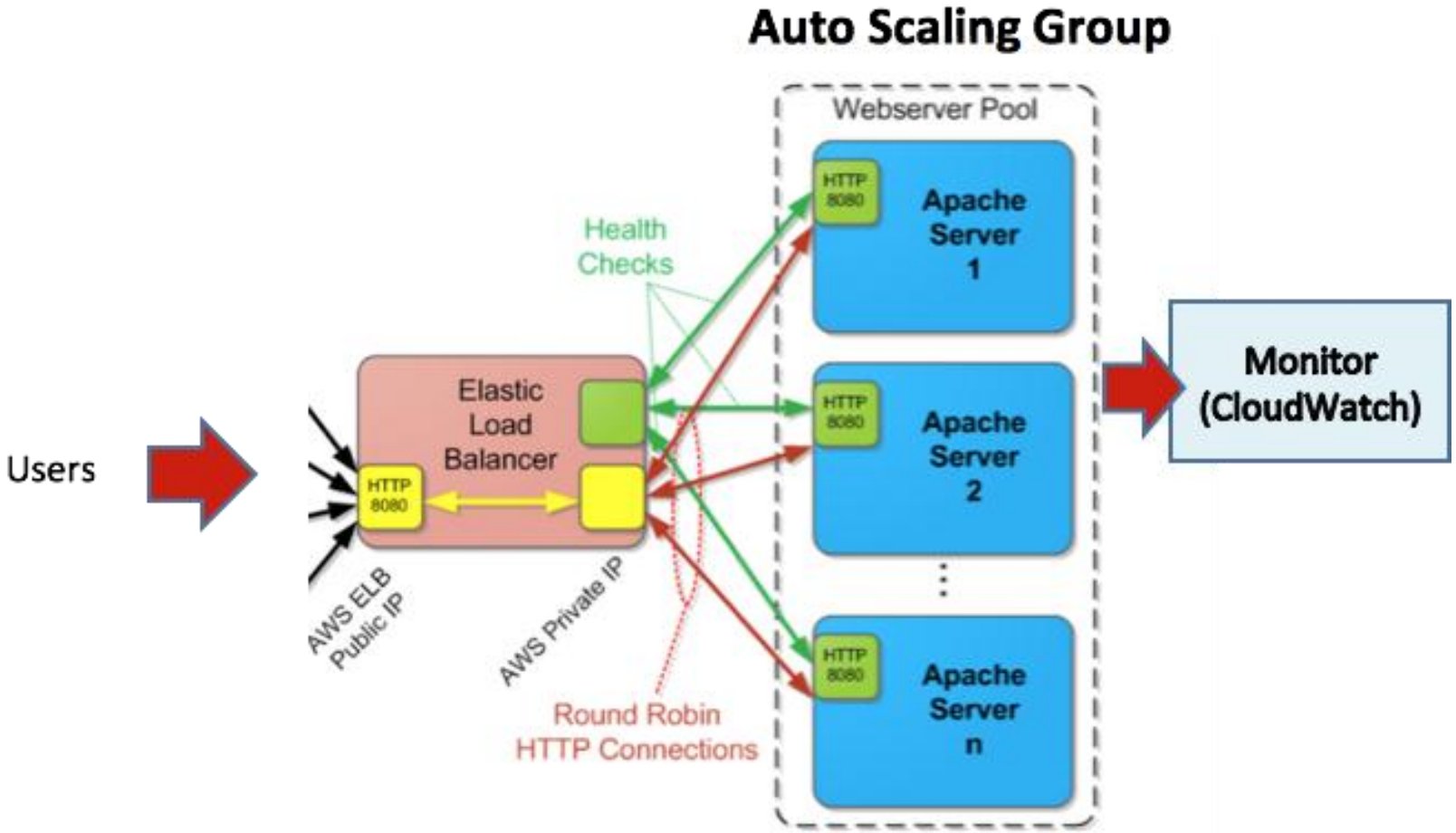
This Week

- UNIT 3: Virtualizing Resources for the Cloud
  - Module 6: Introduction and Motivation
  - Module 7: Virtualization
  - Module 8: Resource Virtualization - CPU
  - Module 9: Resource Virtualization - Memory
  - Module 10: Resource Virtualization – I/O
  - Module 11: Case Study
  - Quiz 3: Virtualizing Resources for the Cloud

This Week

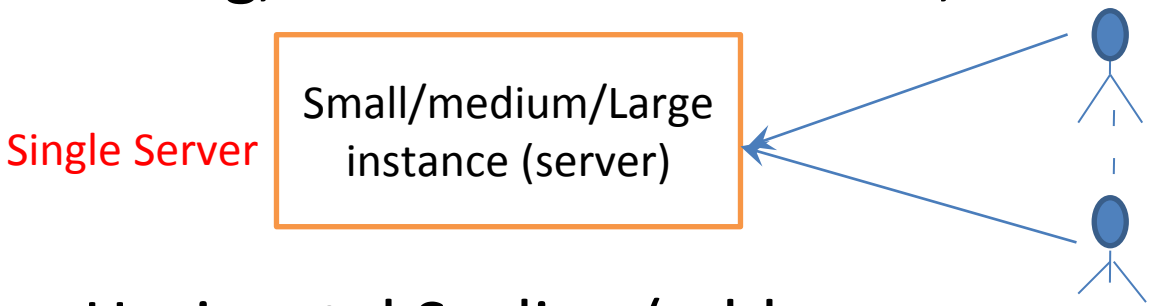
- Introduction and APIs
  - Single Instance Benchmarks
- Elastic Load Balancing (2 modules)
  - Elastic Load Balancer
  - Static Load Benchmarking

Project Module

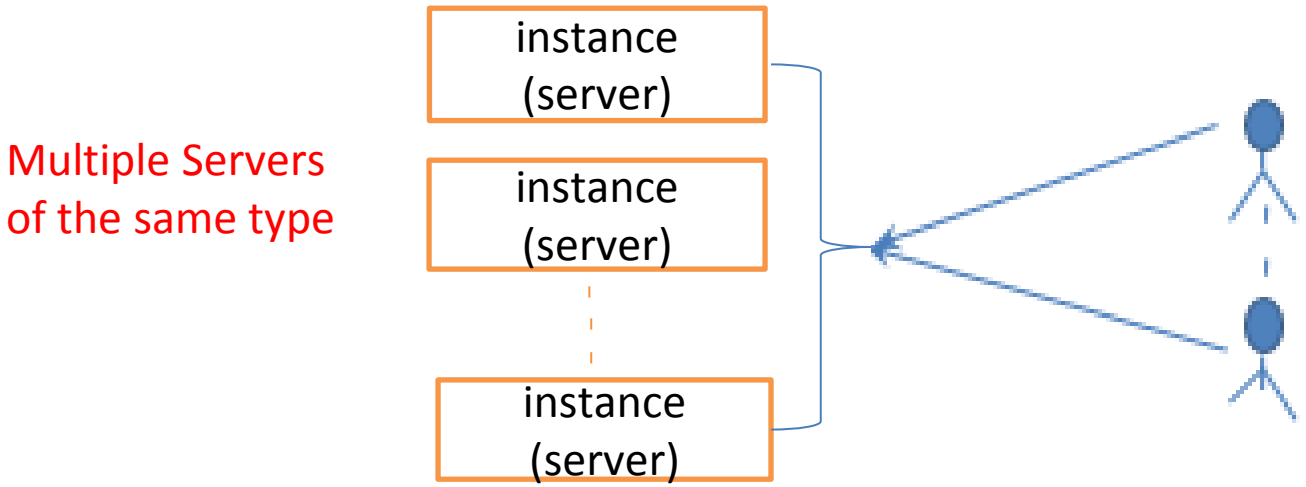


Vertical Scaling vs. Horizontal Scaling

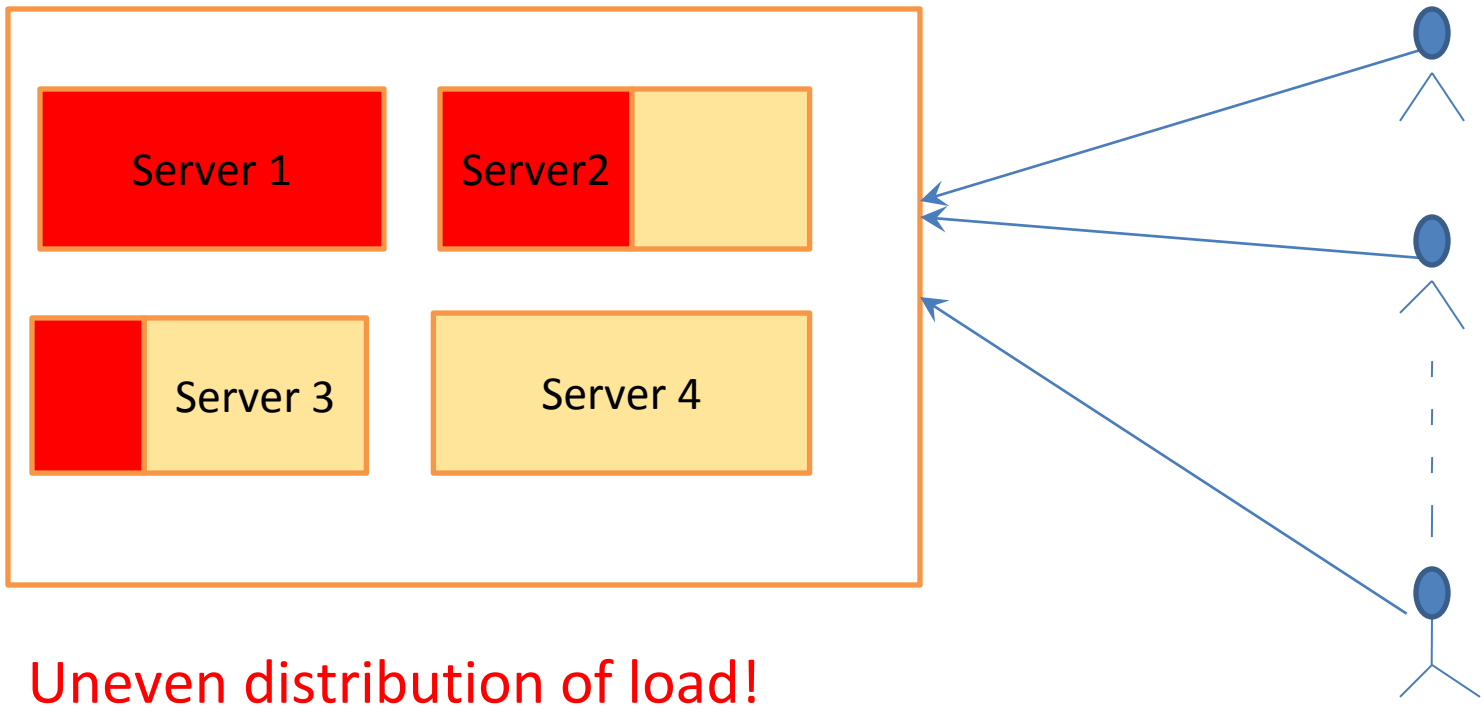
- Vertical Scaling Limitations
  - Can only increase the capacity to a limit
  - When scaling, need to transfer data, have to reboot



- Solution: Horizontal Scaling (add more resources)



Horizontal Scaling

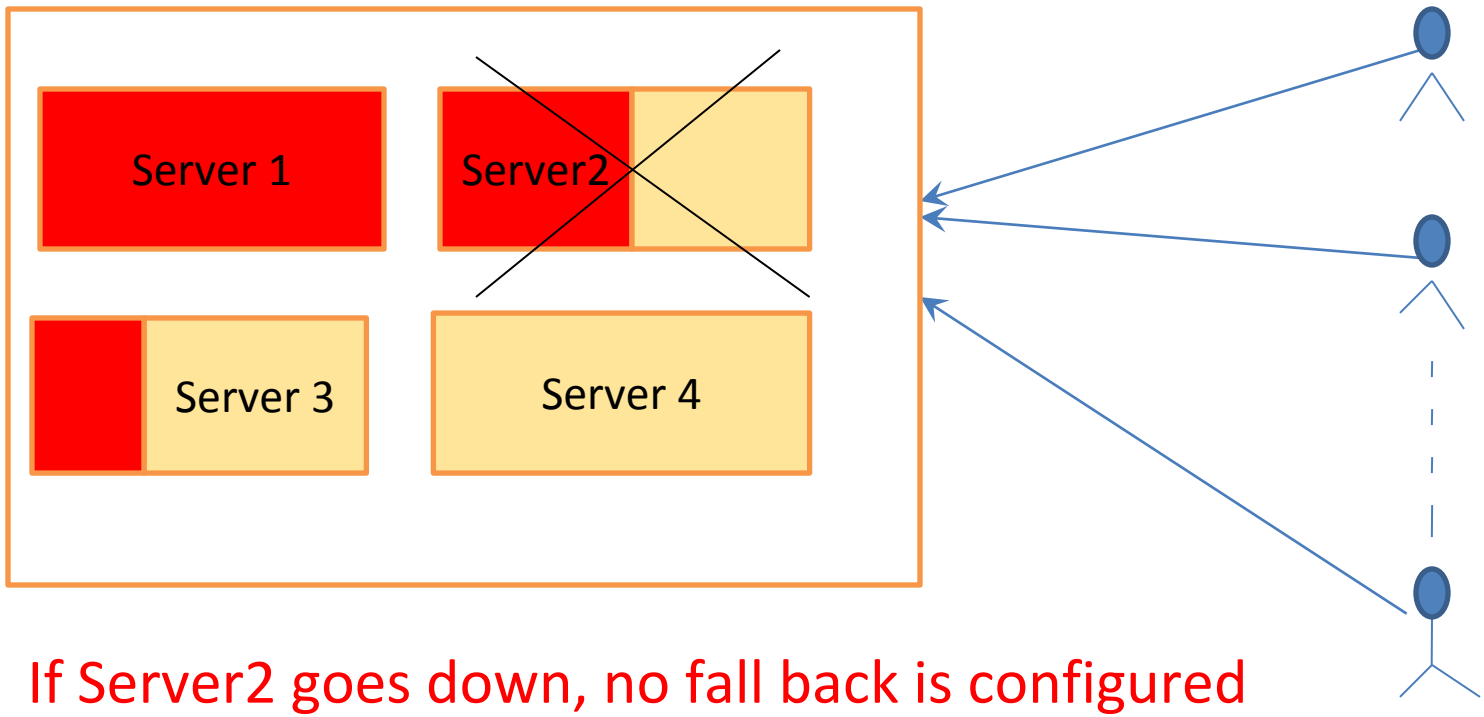


Uneven distribution of load!

CPU utilization, memory utilization…

Available capacity

Horizontal Scaling

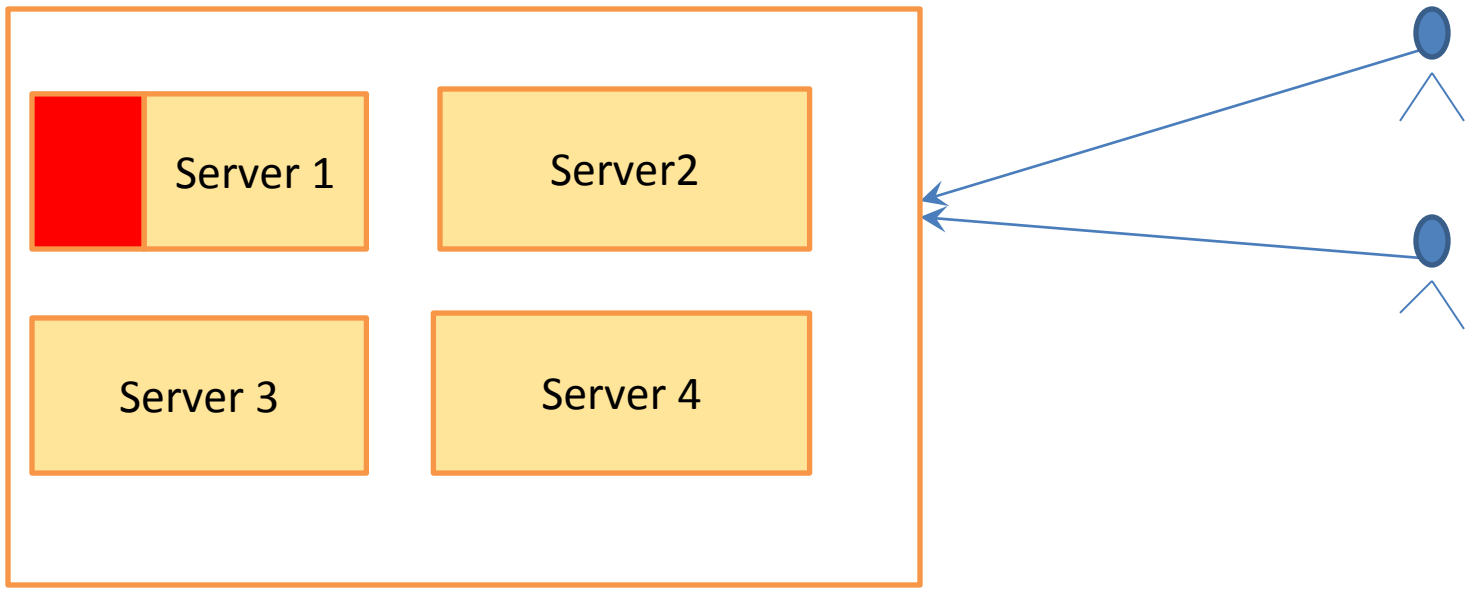


If Server2 goes down, no fall back is configured

CPU utilization, memory utilization…

Available capacity

Horizontal Scaling



If load goes down, we need to change the number of servers

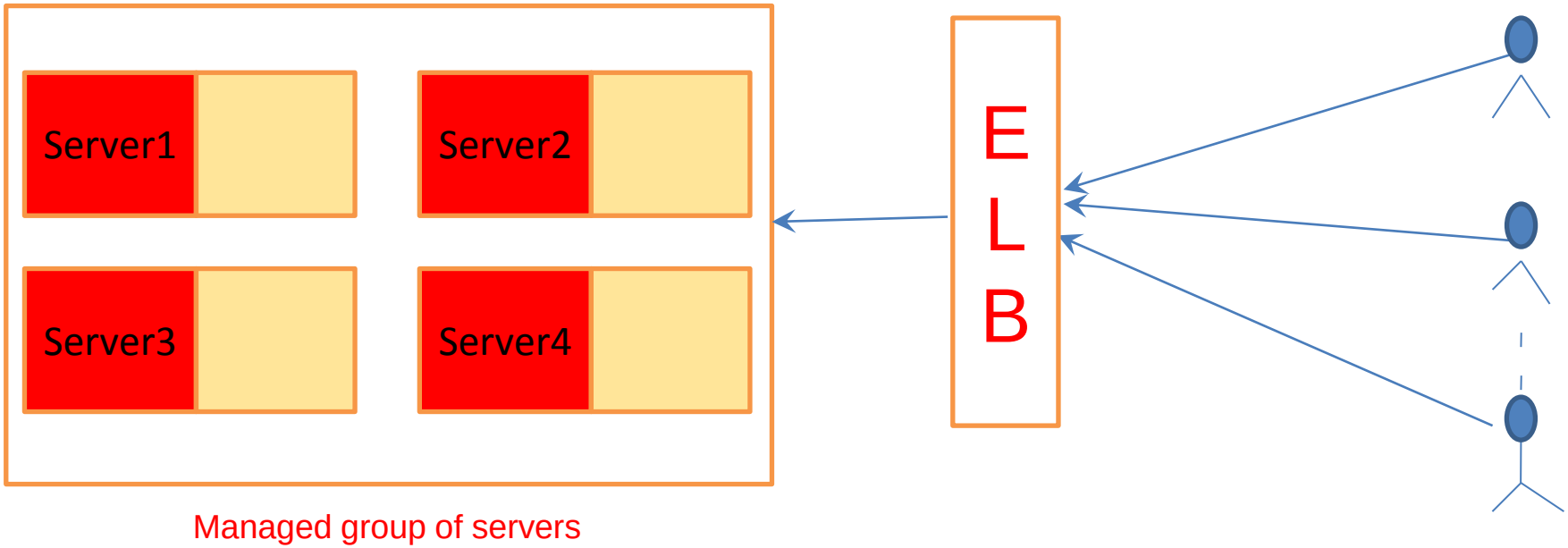
CPU utilization, memory utilization…

Available capacity

What You Need

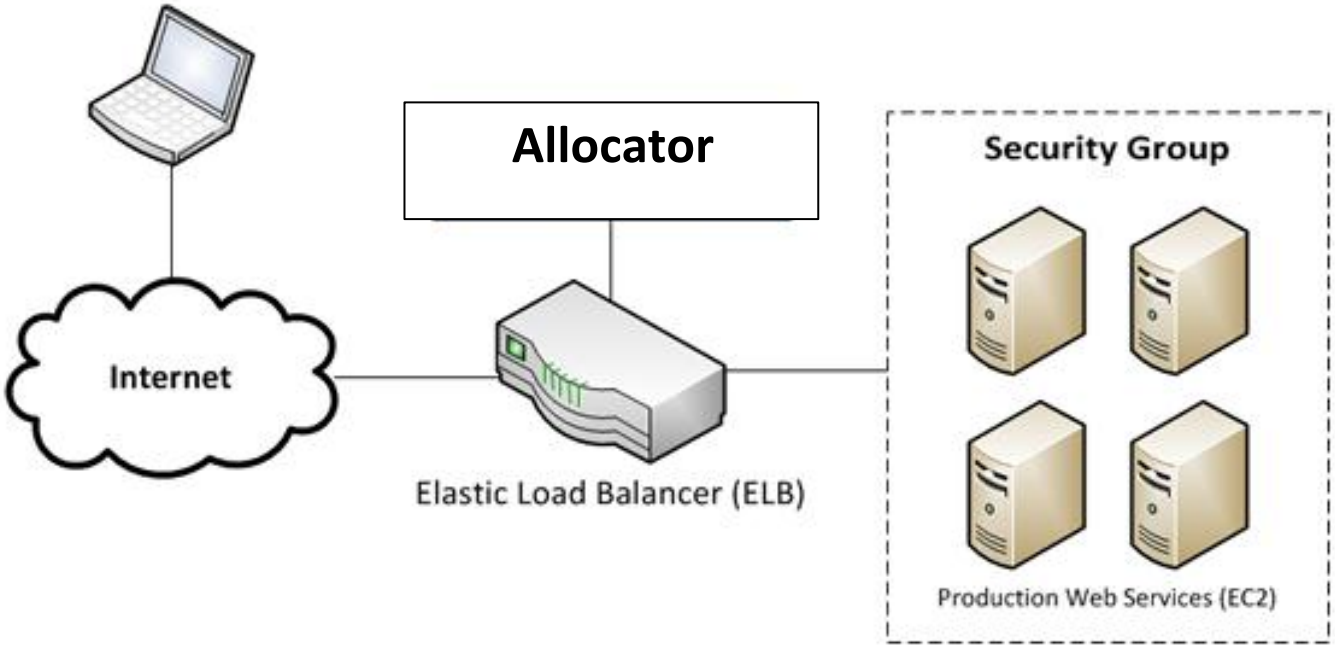
- Make sure that workload is even on each server
- Do not assign load to servers that are down
- Increase/Remove servers according to the changing load

How does AWS help solve these problems?



AWS Elastic Load Balancer (ELB)

- ELB is a gateway that acts as a router interface and sends incoming requests to multiple EC2 Instances sitting behind it
- Distribute requests from clients to all servers equally



modified from: http://brianpeddle.com/wp-content/uploads/2011/02/EC2ELB.png

ELB Features

- Using ELB, you can distribute incoming traffic across your Amazon EC2 instances in multiple Availability Zones (redundancy within the same region)

- ELB can detect the health of Amazon EC2 instances. When it detects unhealthy instances, it spreads the load to other healthy instances

- ELB can offer integration with Auto Scaling to ensure that you can meet varying levels of traffic levels without requiring manual intervention

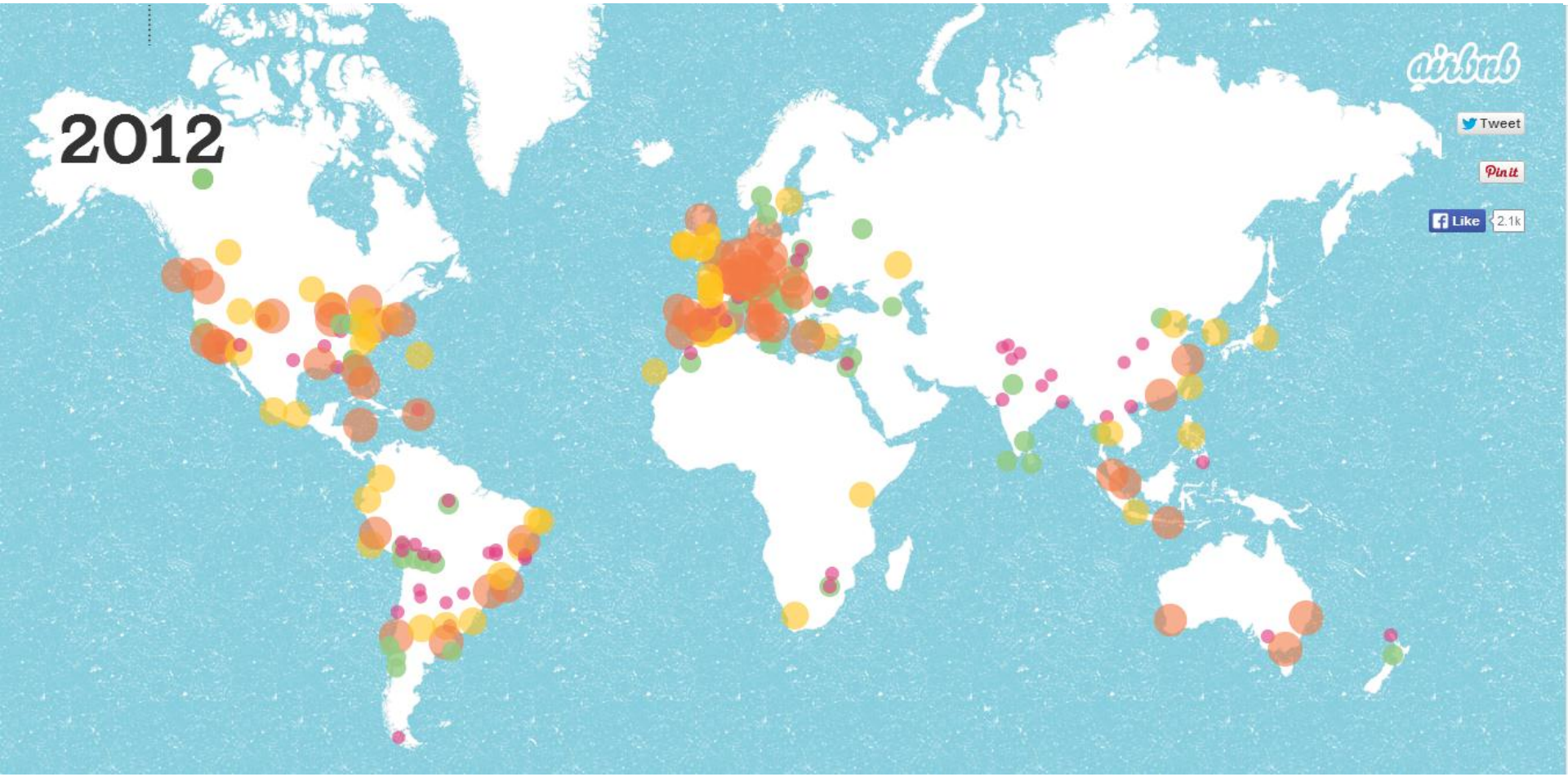
ELB Case

- Airbnb is a community that allows property owners and travelers to connect with each other

- Airbnb is using Elastic Load Balancing, which automatically distributes incoming traffic between multiple Amazon EC2 instances. As a result, it saves a lot of operation expenditures

- Within 4 years, it now has hundreds of employees in nearly 25,000 cities in 192 countries

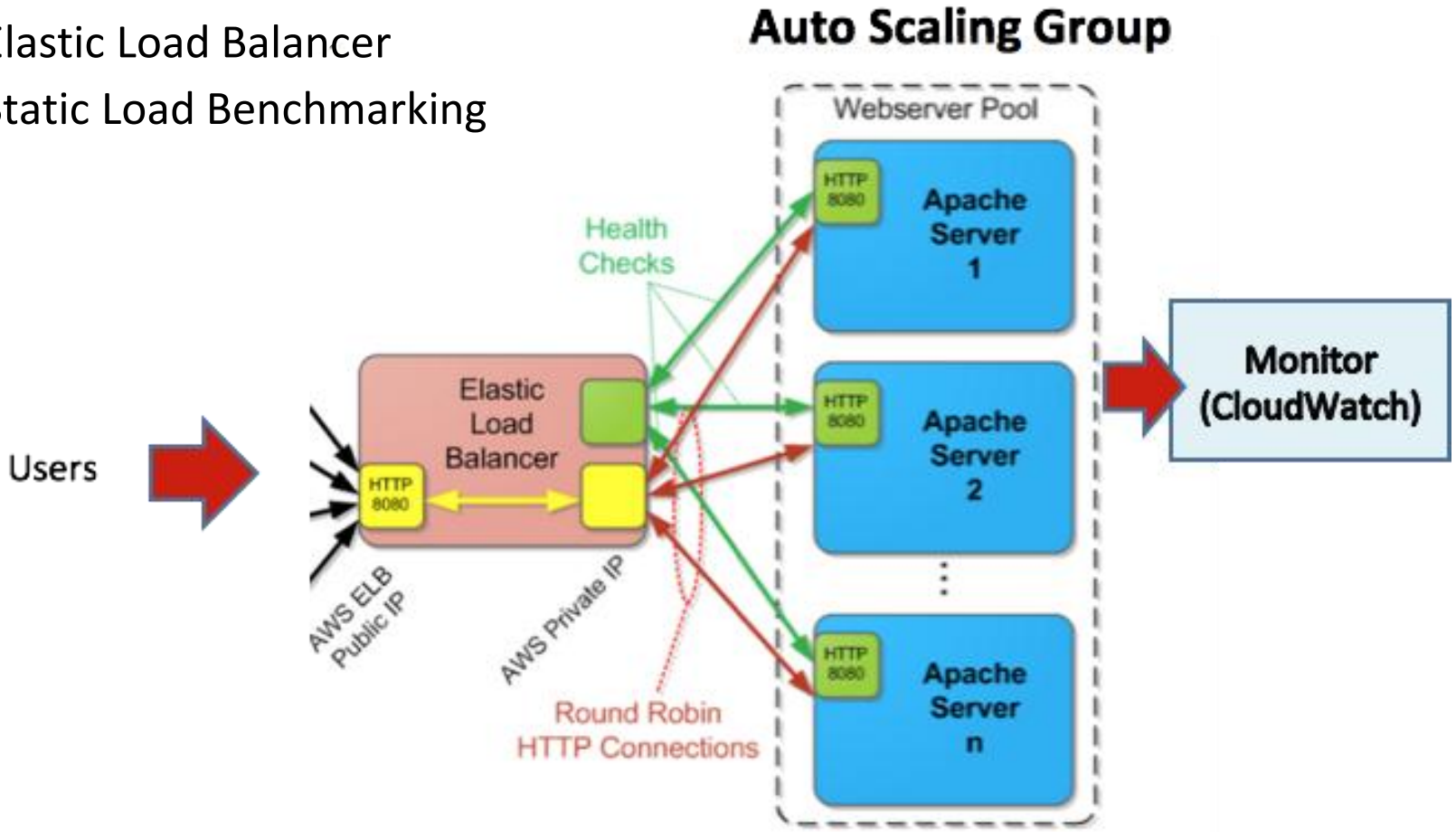
ELB Case



When scaling and load balancing, Airbnb employs ELB!

Project Module

- Elastic Load Balancing
  - Elastic Load Balancer
  - Static Load Benchmarking



Upcoming Deadlines

- Project 2:

Project 2			
	Introduction and APIs		
	Single Instance Benchmarks	Checkpoint	Available Now Due 2/9/14 11:59 PM
	Elastic Load Balancing		
	Elastic Load Balancer	Checkpoint	Due 2/16/14 11:59 PM
	Static Load Benchmarking	Checkpoint	Due 2/16/14 11:59 PM

- Unit 3:

UNIT 3: Virtualizing Resources for the Cloud		
	Module 6: Introduction and Motivation	
	Module 7: Virtualization	

Discussions

- Questions and Comments

Demo

- Create a new Load Balancer manually. (Project 2 Module 2)

- Create a new Load Balancer and attach the instance to ELB programmatically using Python or any other programming language of your own choice. (Project 2 Module 3)
    - Create an ELB
    - Provision an instance & attach it to ELB
    - Wait some time for it to be in service