CS15-319 / 15-619
Cloud Computing

Recitation 5

September 23th & 25th, 2014

Announcements

- Project
  - Tag your instances
  - Start early
- OLI
  - Save your work as you go
  - Lots of students submitting at the same time
- Provide feedback on OLI
- Post on Piazza:
  - Private: grading questions
  - Public: general questions
    - Search Piazza and the web before posting

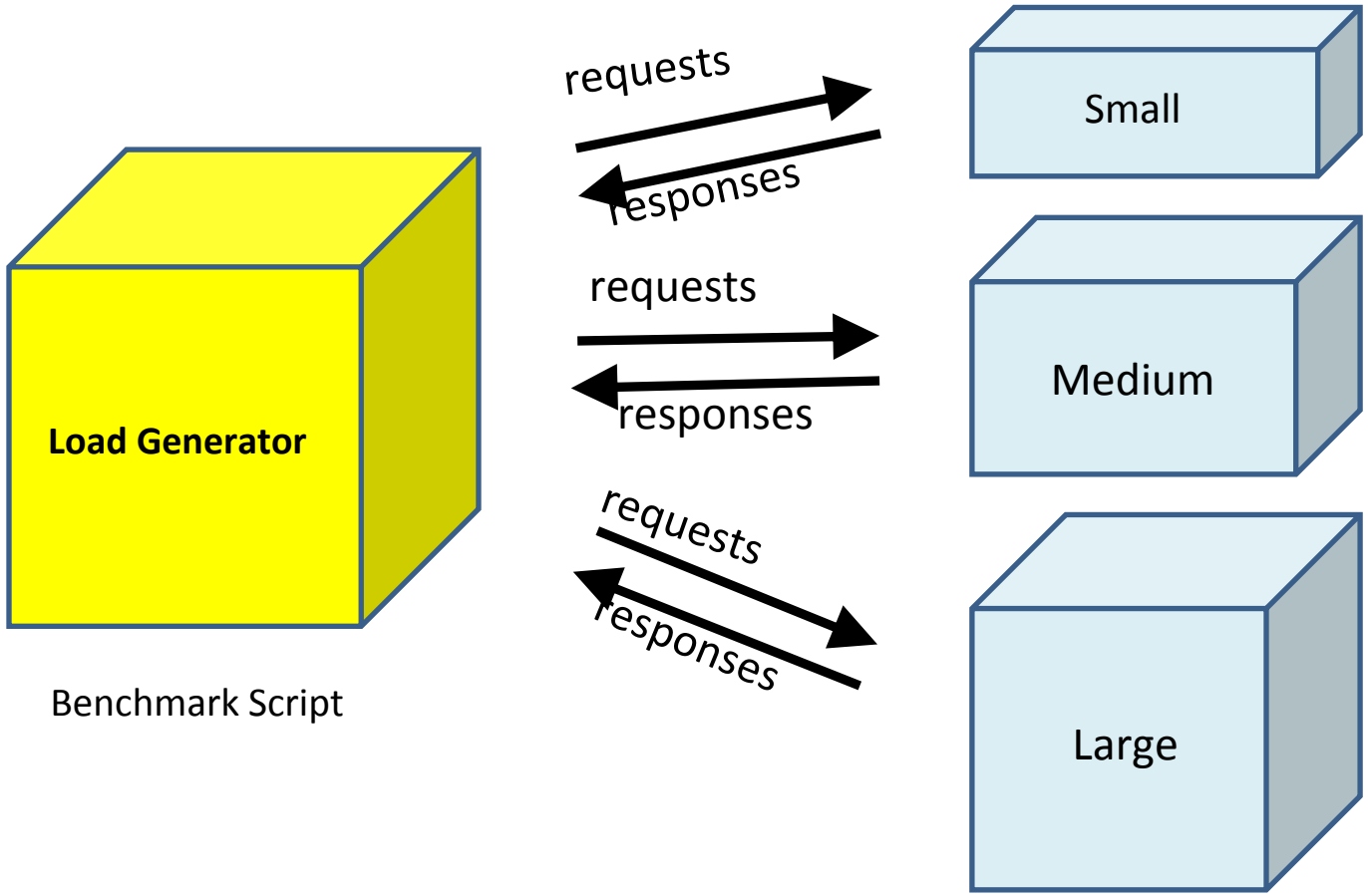
Announcements

- Monitor AWS expenses regularly
  - Suggestions
    - Terminate your instance when not in use
      - stop still costs some money
      - check other availability zones
    - Use smaller instances to test your code
    - Check your bill frequently

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 and Horizontal Scaling
  - Instance Capacity

Reflection on Last Week

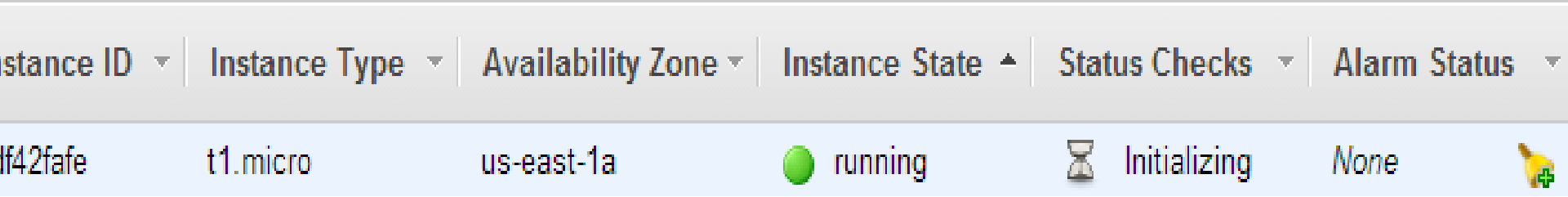


Piazza Questions

- Spot instance & On-demand instance
  - Use On-demand instance if spot price is higher than On-demand price
- Cannot connect to instance
  - Check your SecurityGroup
- How to make sure the instance is running?
  - DescribeInstaceRequest correct
  - instance.getState() wrong

Piazza Questions

- And… you still need time to initialize
  - Otherwise you cannot access it by URL



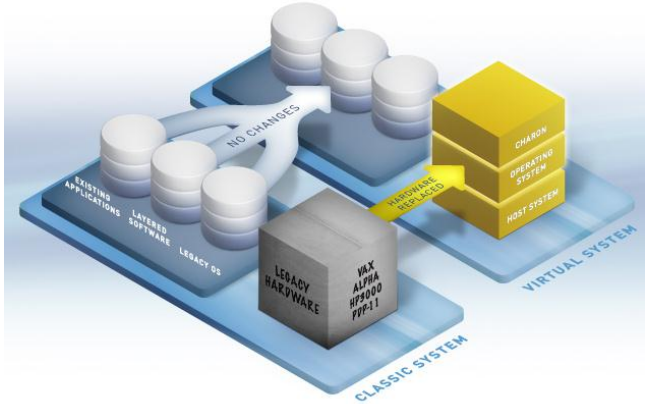
- ELB warming-up
  - ELB will be warmed up after the first run, and will have better performance than before

This Week: Content

- 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

Module 6: Introduction and Motivation

- Why Virtualization
  - Enabling the cloud computing system model
  - Elasticity
  - Resource sandboxing
  - Improved system utilization and reduce costs
  - Mixed-OS environment
- Limitations of General-Purpose OS
- Managing System Complexity
- Resource Sharing

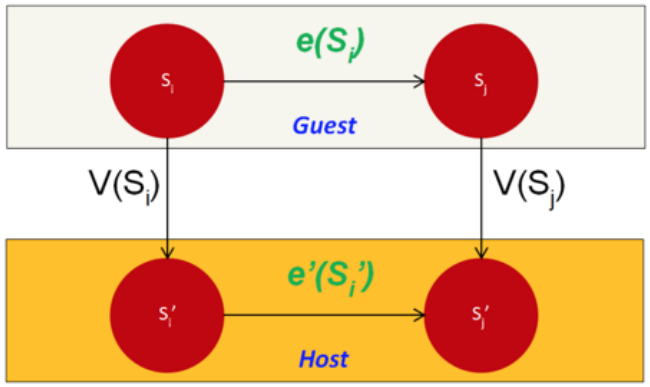


virtualization

Module 7: Virtualization

- What is Virtualization
  - Involves the construction of an isomorphism that maps a virtual guest system to a real (or physical) host system

- Virtual Machine Types
  - Process Virtual Machines
  - System Virtual Machines

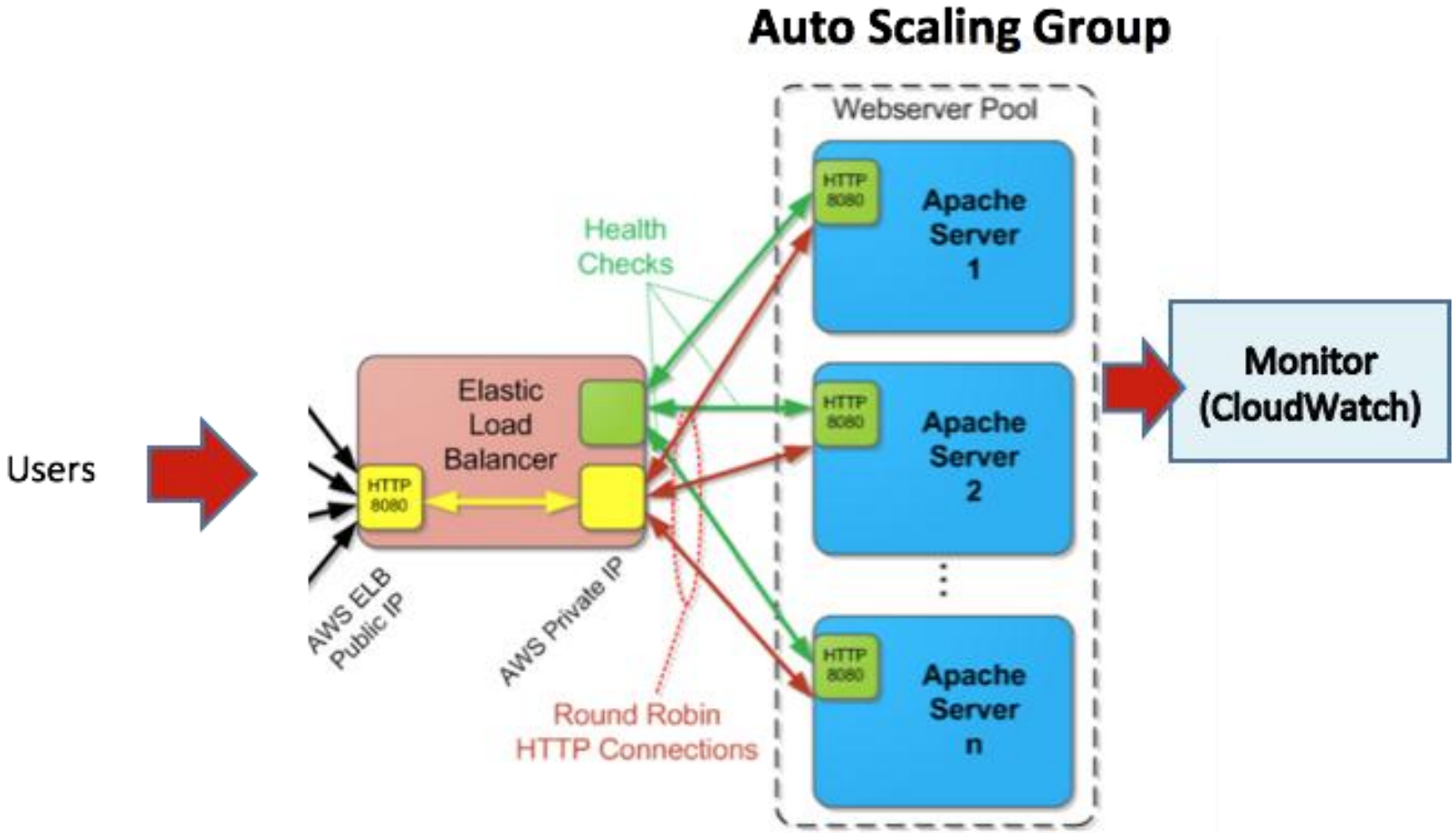


Virtualization

This Week: Project

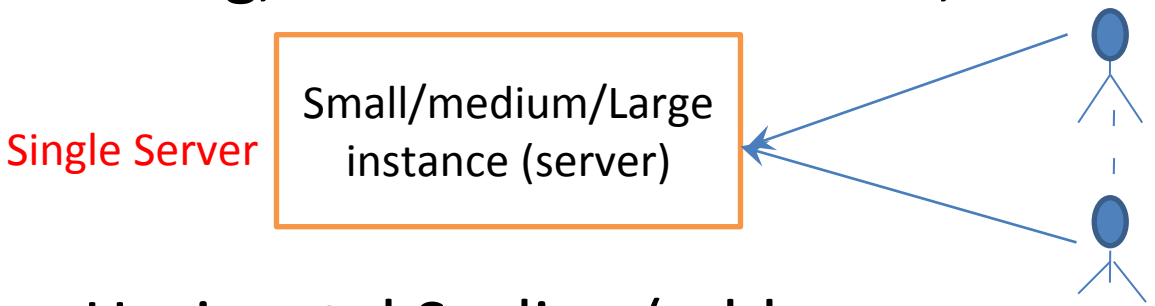
- Introduction and APIs
  - MSB Recruitment Exam
- Elastic Load Balancing
  - Junior System Architect at the MSB

Project Module

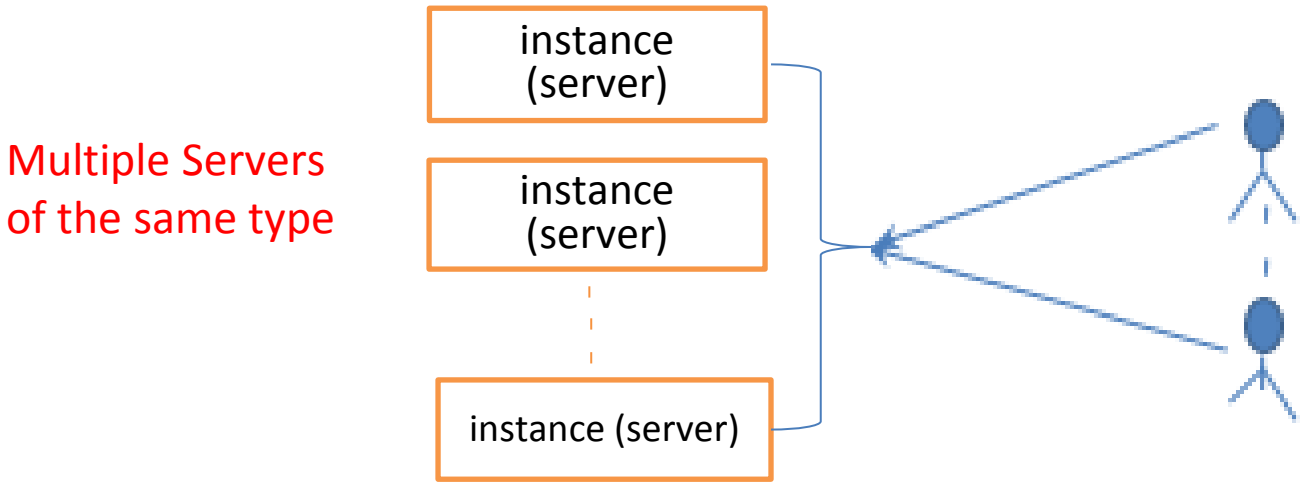


Vertical Scaling vs. Horizontal Scaling

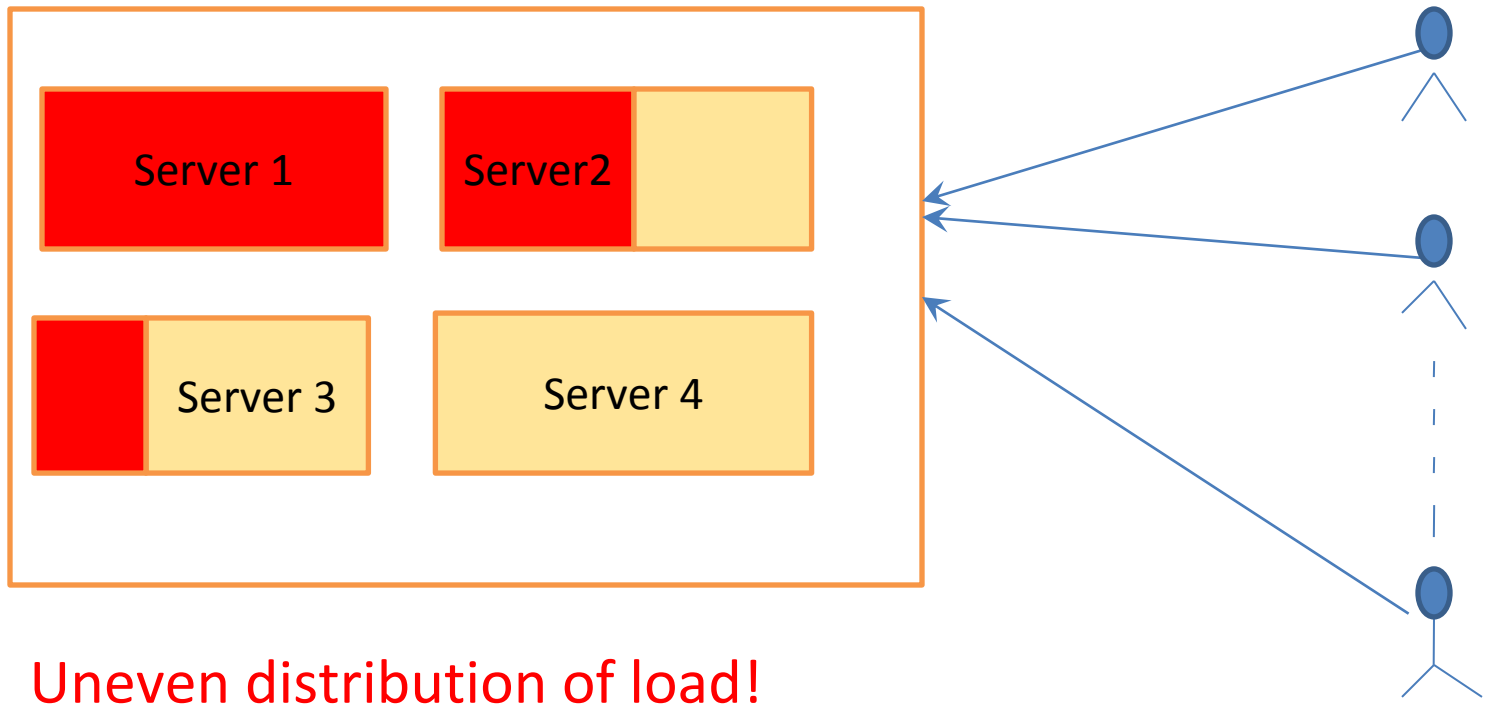
- Vertical Scaling Limitations
  - Can only increase the capacity to a limit
  - When scaling, need to transfer data, have to reboot



- Solution: Horizontal Scaling (add more resources)



Load Balancing in Horizontal Scaling

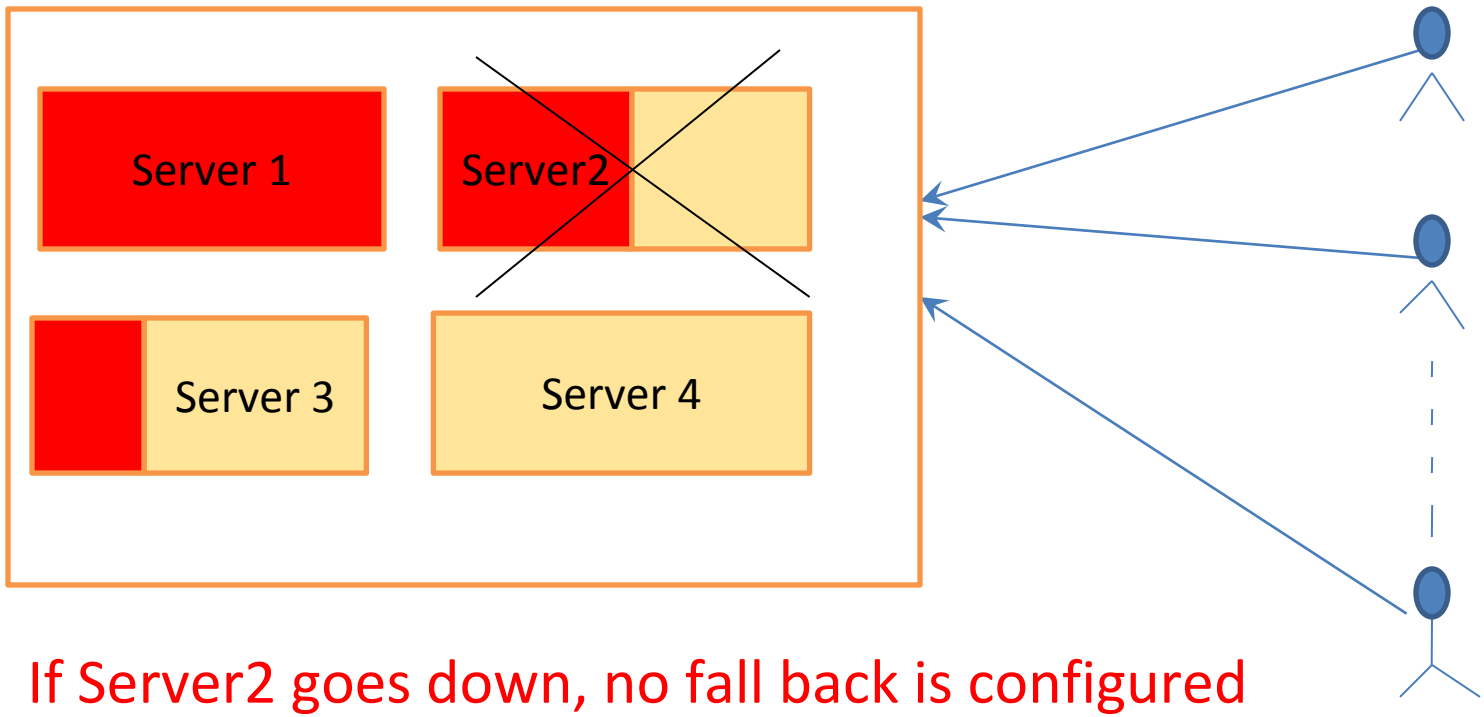


Uneven distribution of load!

CPU utilization, memory utilization…

Available capacity

Health Check in Horizontal Scaling

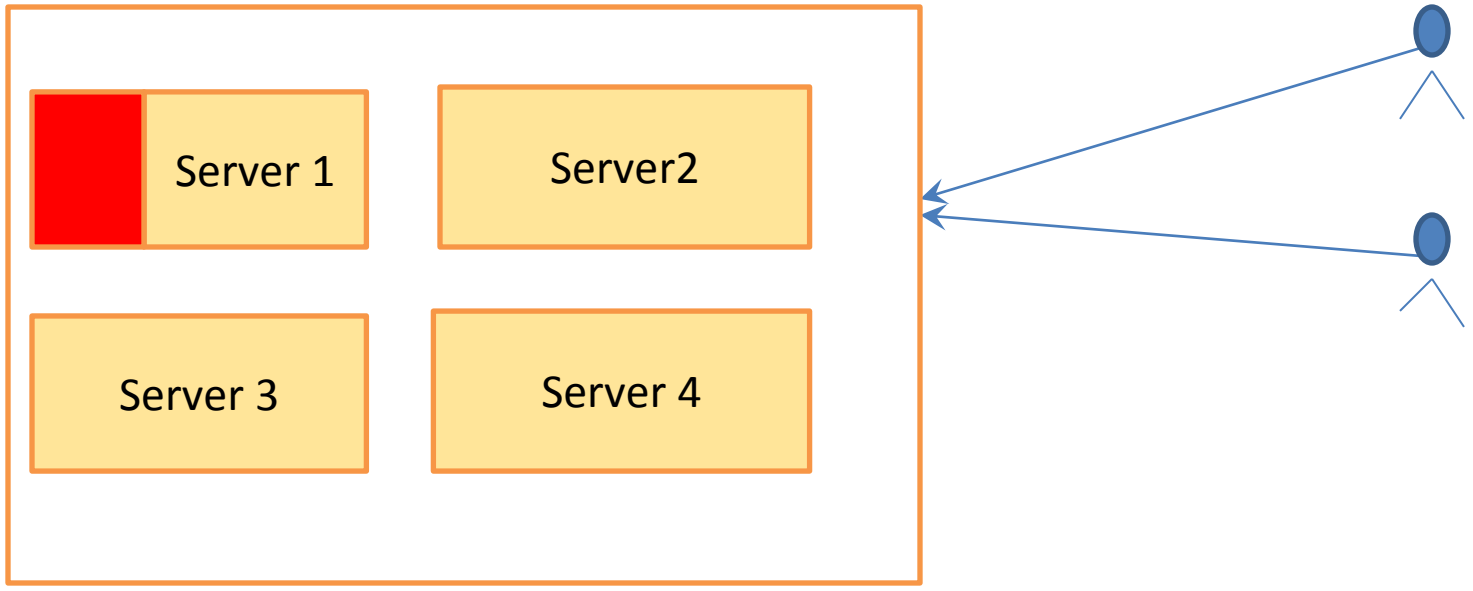


If Server2 goes down, no fall back is configured

CPU utilization, memory utilization…

Available capacity

Utilization in Horizontal Scaling



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

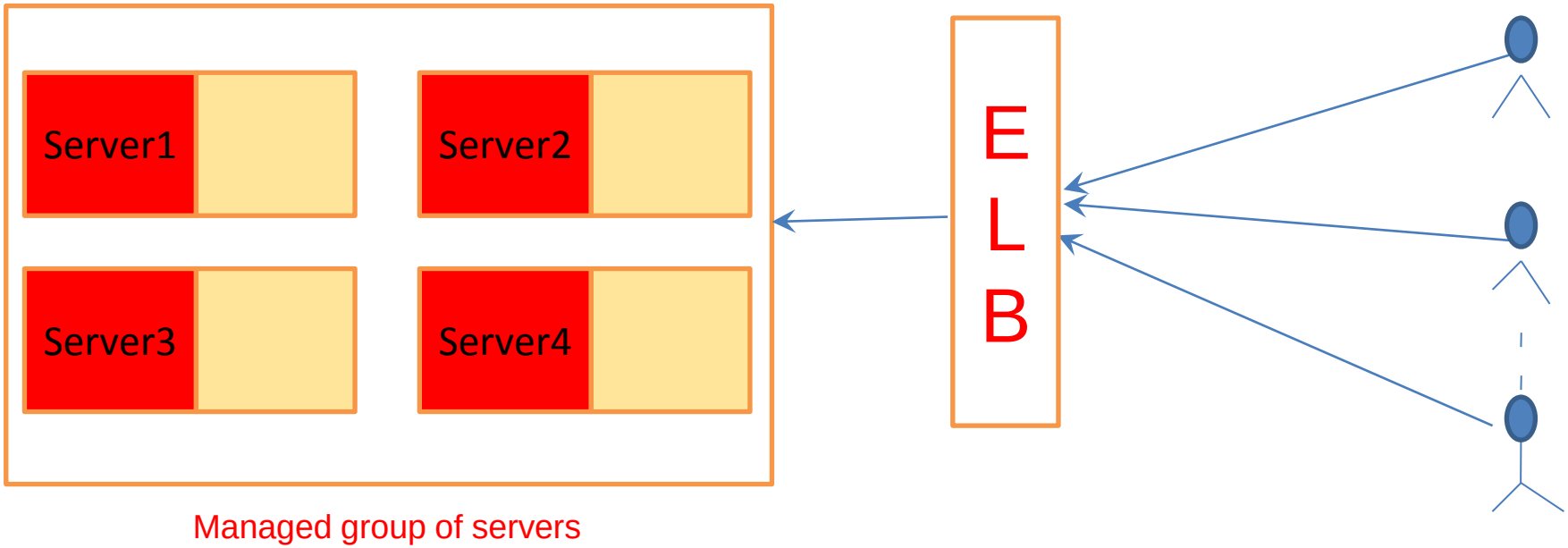
CPU utilization, memory utilization…

Available capacity

What You Need

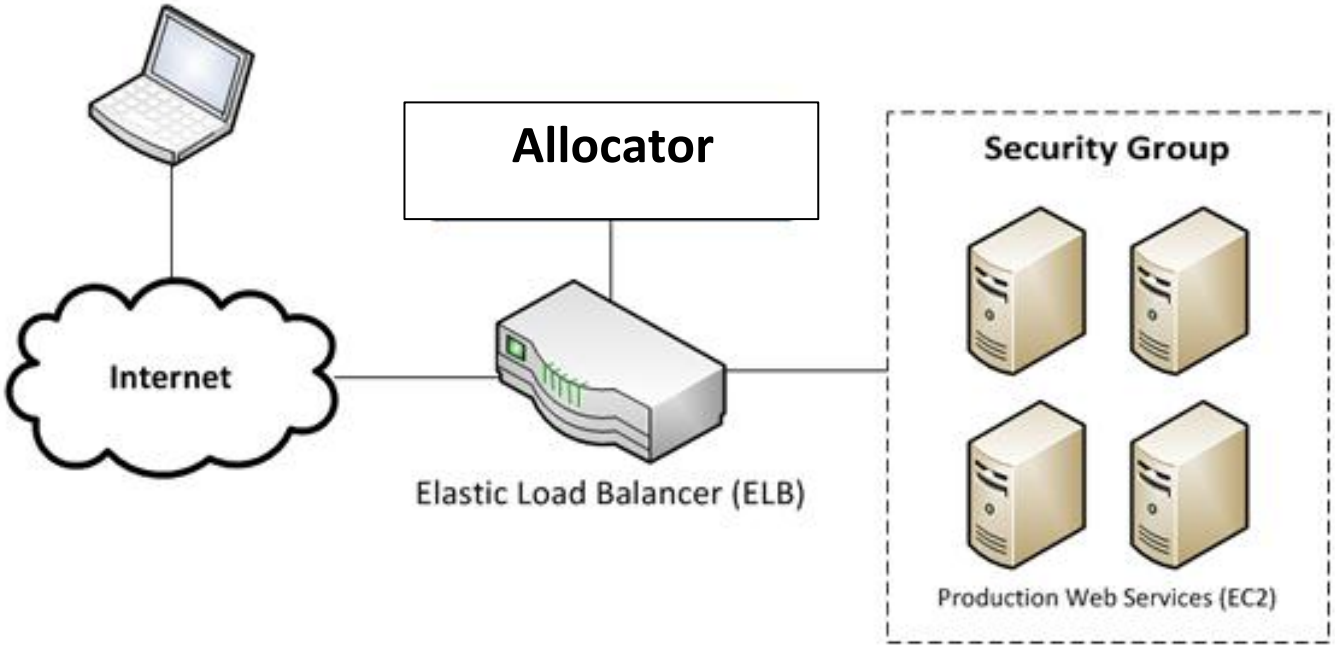
- Make sure that workload is even on each server
- Do not assign load to servers that are down
- Increase/Remove servers according to changing load

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

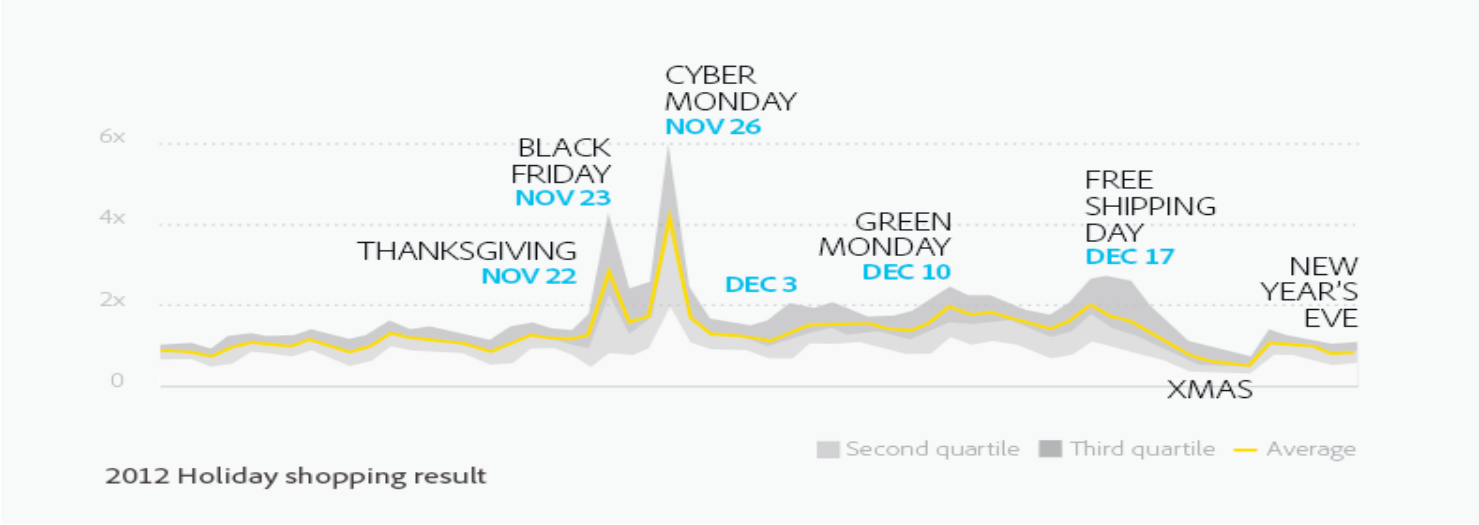
- ELB can offer integration with Auto Scaling to ensure that you can meet varying levels of traffic levels without requiring manual intervention

ELB Needs Warming Up

- ELB has a starting point for its initial capacity, and it will scale up or down based on traffic
- It struggles with high traffic spikes in shorter periods
- It is recommended that the load is increased at a rate of no more than 50 percent every five minutes

Changing Load

- Different network traffic throughout the year
  - There is a burst in the holiday season
  - If performance suffers, you are losing customers
  - Should vary the system size for different seasons

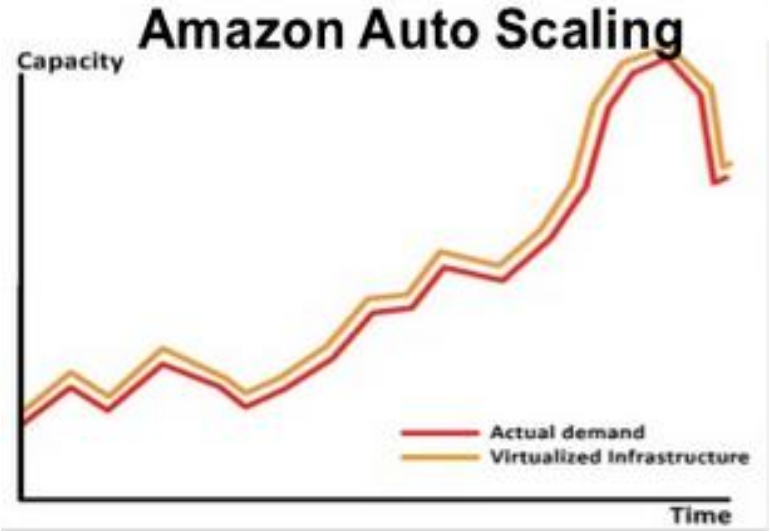


sapient.com

Why Auto Scaling?

- Traditional Scaling:
  - Manually control the size
  - Under utilization of resources
  - Lose customers

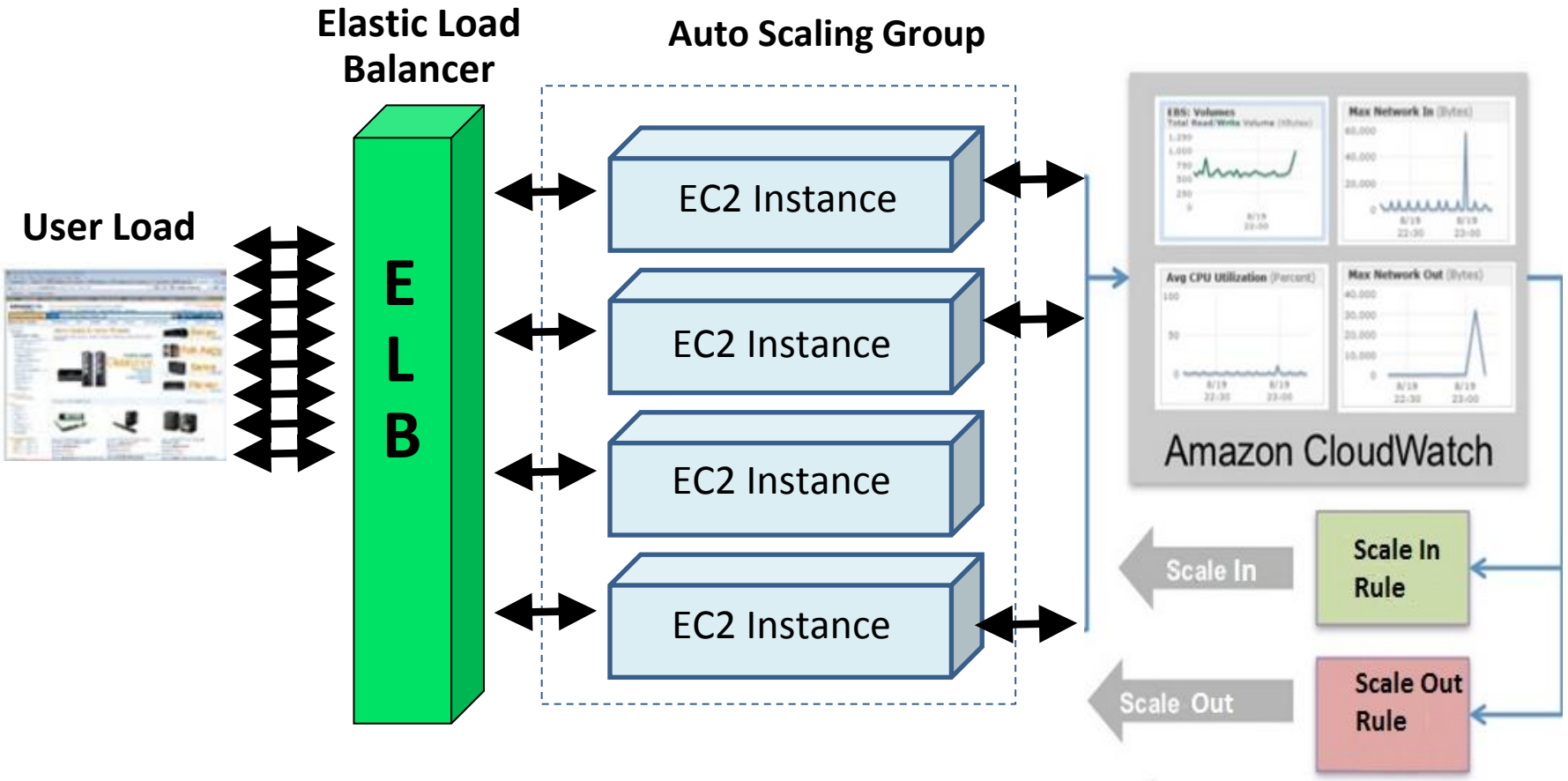
- Auto Scaling:
  - Adjust the size automatically based on demand
  - Flexible capacities and scaling sizes
  - Save cost



Picture From Amazon

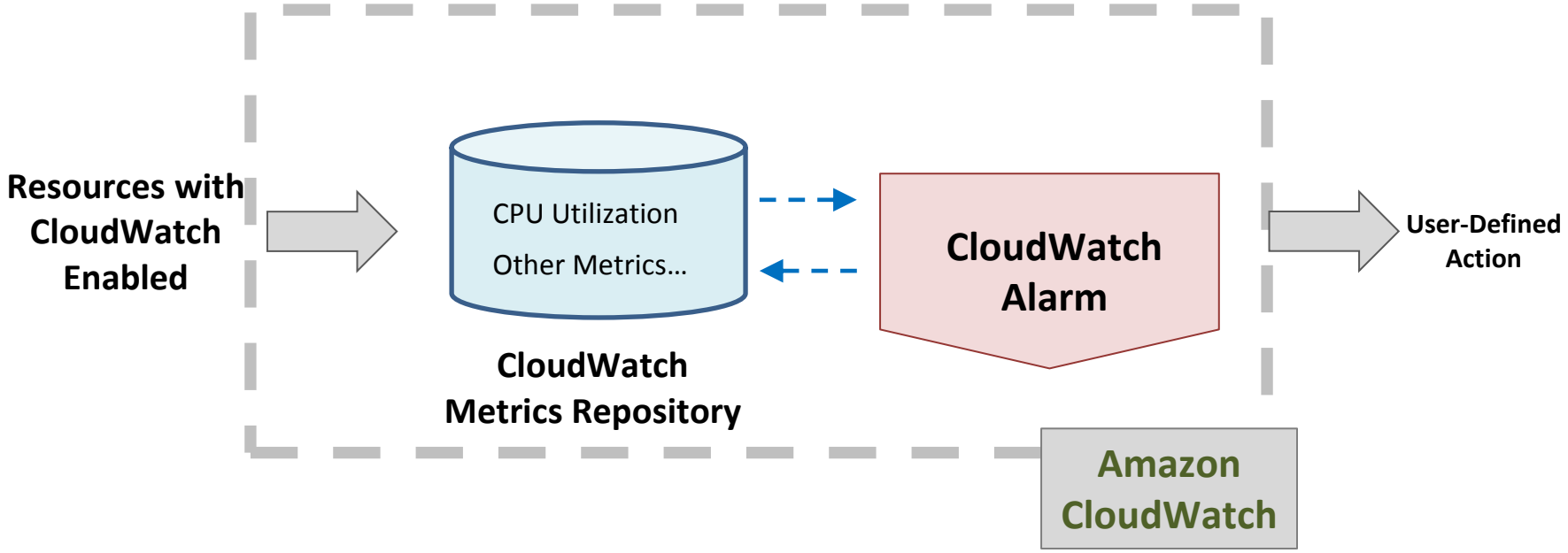
Amazon Auto Scaling Group

- Scale Amazon EC2 capacity automatically according to policies you define



Amazon's CloudWatch Alarm

- Monitor CloudWatch metrics for some specified alarm conditions
- Take automated action when the condition is met

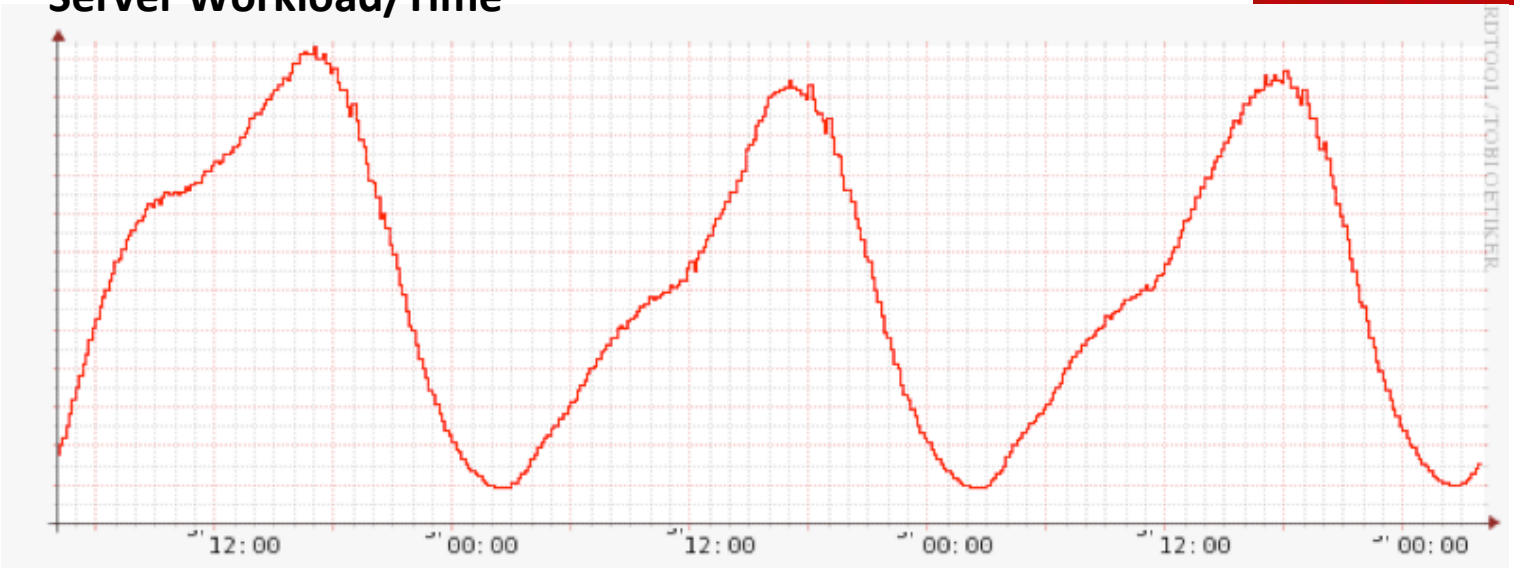


Case Study

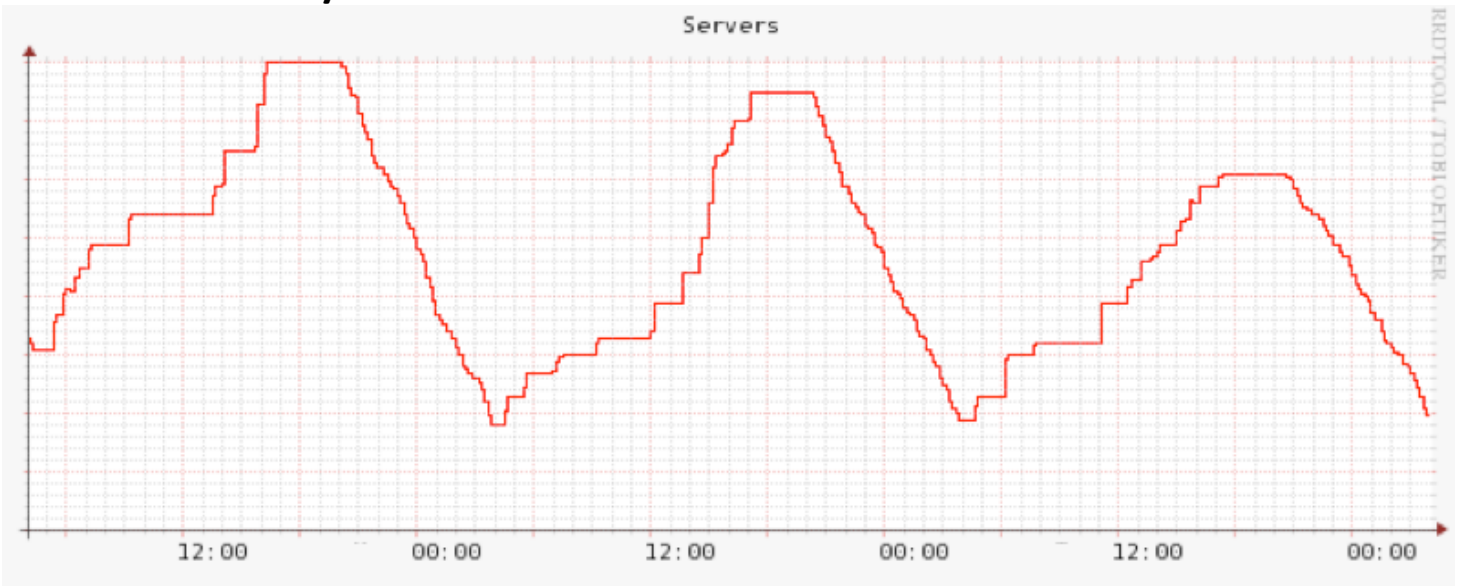
- Netflix is one of the most popular provider of on-demand Internet streaming media
- Netflix has been using Amazon Auto Scaling Group for about 3 years.
- Netflix takes advantage of ASG features to manage running a pool of servers, including the capability to replace failed instances and automatically grow and shrink the size of the pool.
- Data shows that use of ASG greatly improves the availability of Netflix services and provides an excellent means of optimizing cloud costs. http://techblog.netflix.com/2012/01/auto-scaling-in-amazon-cloud.html

Case Study

Server Workload/Time



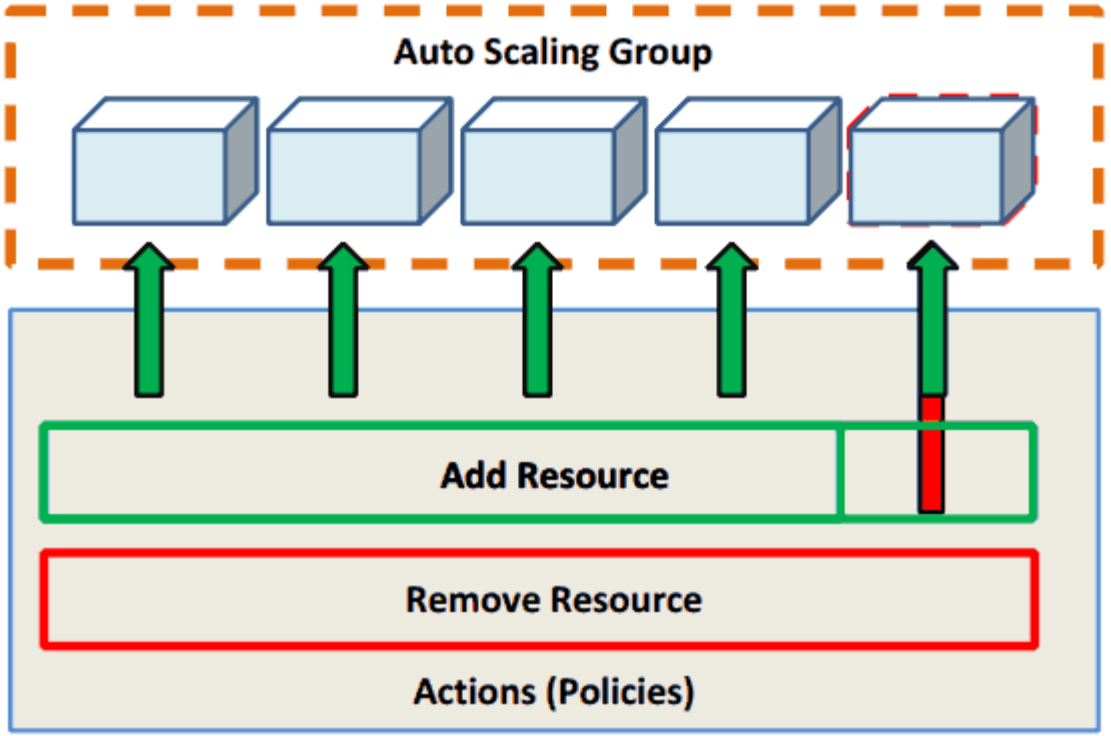
Server Number/Time

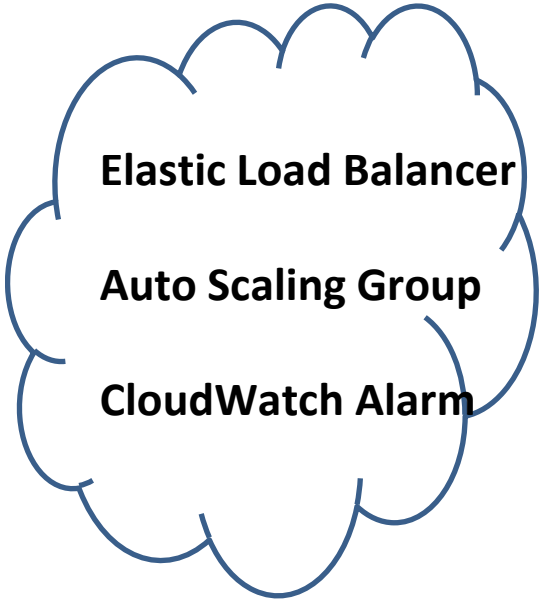


http://techblog.netflix.com/2012/01/auto-scaling-in-amazon-cloud.html

Your Task

- Programmatically create an Elastic Load Balancer (ELB) and an Auto-Scaling Group (ASG) linked to ELB.
- Test by submitting a URL request and observe changes
- Decide on Scale-Out and Scale-In policy





Resources

- Amazon's Auto Scaling Service
  - http://aws.amazon.com/autoscaling/
- Amazon's CloudWatch Alarm
  - http://aws.amazon.com/cloudwatch/
- Amazon's Scaling Developer
  - http://aws.amazon.com/autoscaling/developer-resources/

Project 2.2 Change

- AMI of Load Generator has changed
  - new AMI: “ami-562d853e”
  - posted as a P2.2 Change-Log on Piazza

Project 2 Reminders

- Read project description more than once
- Think about workflow before starting
- Look up API references
  - Read overview first, then details
  - Use samples to go over simple APIs
  - Use the Internet to debug
- Check every step carefully
  - Debug with AWS console

Other Reminders (cont')

- Terminate instances vs. Stop instances
  - Stop will still charge for VM storage (EBS volumes)
    - Stop is a good idea when you need a break

- S3 URL Validation and Naming Convention
  - Always validate your link before submitting the link
    - http://ec2-54-225-106-182.compute-1.amazonaws.com/
  - Follow submission guidelines
  - DO NOT add your credentials in your code

Project 2 Reminders

- Tag ELB and ASG
  - Also tag instances in the ASG
- Check ELB Instance Status
  - Wait until “InService”
- Think about each step
  - Impact of adding a new instance on rps
  - How much ELB warm up is needed
  - How to calculate instance-hours from the ELB Healthy Hosts graph
- Do not design your system hoping for the same pattern.
  - Traffic patterns (location of DDoS spikes) change in every run.
- The solution is not simple. Use all the instance and ELB logs once with a fixed number of instances to understand the traffic pattern.
- Think about how to ensure that you do not overprovision instances.

Upcoming Deadlines

- Project 2:

Project 2			
	Introduction and APIs		
	MSB Recruitment Exam	Checkpoint	Available Now Due 9/21/14 11:59 PM
	Elastic Load Balancing		
	Junior System Architect at the MSB	Checkpoint	Due 9/28/14 11:59 PM Pittsburgh

- Unit 3:

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