03/20/2025 Quiz 1
6:00pm-8:15pm

- **This is an open book exam. Only PDFs can be used as a reference. Internet access is strictly prohibited. Handwritten / Physical Reference notes are allowed**
- **Set your laptop to maximum brightness and have no tab/applications open other than PDFs.**
- **Marks are not awarded based on the length of the answer, the more precise/concise/bullet-pointed it is, the better.**
- **Any AI Tools, messaging applications, collaborative editing apps are strictly prohibited, please make sure they're not present on the dock/taskbar**

Section A: [Based on Learnings from Assignment 1 and Assignment 2] 3x10 points

1. What would happen if your Docker container for MongoDB does not specify a persistent volume?

2. In Amazon Lex, which component is responsible for handling user input before Lex responds?
    a. Bot
    b. Intent
    c. Lambda code hook
    d. Slot type

3. Your serverless chatbot API experiences high response times after a period of inactivity. What is the most likely cause?
    a. The API Gateway caching mechanism is misconfigured.
    b. The Lambda function is experiencing a cold start delay.
    c. The Lex bot is consuming too many Lex slots per request.
    d. The Lambda function is running in a VPC without internet access.

4. Your chatbot API deployed on API Gateway + AWS Lambda is experiencing a sudden surge in traffic, leading to *immediate* failed requests. What is the most likely cause?
    a. The Lambda function's memory allocation is too low
    b. Lambda has reached its concurrent execution limit
    c. The DynamoDB table is not optimized for query performance
    d. The API Gateway is not configured with a usage plan

5. While testing your chatbot, you notice that Lex always returns the fallback intent, even when the user input clearly matches an intent. What is the most probable reason?
    a. The Lex bot is not properly trained with enough utterances
    b. The Lex bot has not been published and deployed to a specific version
    c. The Lambda code hook is failing and Lex is defaulting to fallback
    d. Any of the above can be a cause

6. Your replication controller in Kubernetes is set to maintain 5 replicas of a pod, but you notice that only 3 are running. What could be the reason?
    a. The remaining nodes in the cluster do not have enough resources
    b. Kubernetes does not support replication for multi-container pods
    c. The pod deployment is using an unsupported Docker image
    d. The replication controller is waiting for approval to create new pods

7. Your Flask application on AWS EKS uses MongoDB for data storage. After a pod restart, all previously stored data is lost. What is the most likely issue?
    a. The MongoDB deployment is using an emptyDir volume instead of a Persistent Volume Claim (PVC).
    b. The Persistent Volume Claim (PVC) is not mounted under /data/db, MongoDB's default data directory.
    c. The Kubernetes cluster does not support stateful applications.
    d. The MongoDB container does not expose the correct port.

8. If your Minikube cluster fails to expose the To-Do app, which of the following is most likely the issue?
    a. The service is created as ClusterIP instead of NodePort.
    b. The Kubernetes API server is down.
    c. The pod is in Running state but has no logs.
    d. The To-Do app does not have enough replicas.

9. If your liveness probe is failing continuously, what action will Kubernetes take?
    a. Kubernetes will scale down the pod
    b. Kubernetes will restart the failing pod
    c. Kubernetes will log the failure but take no action
    d. Kubernetes will redirect traffic to a different service

10. If your Flask application is failing after being deployed on AWS EKS with an error indicating 'ImagePullBackOff,' what should you check first?
    a. Ensure Docker is installed on your local machine
    b. Verify that the correct image is pushed to Docker Hub
    c. Increase the number of replicas in the deployment
    d. Restart the Minikube cluster

Section B: System Design[30]

Real-time Restaurant Recommendation System

Background:

Extend assignment 1 to contain the features for:

- Calculate ratings for restaurants
- Trending restaurants in the user's area

Assume trending restaurants in the area can be obtained through an external api call that has already been provided to you.

As user traffic increases, performance bottlenecks emerge. You must design a scalable, low-latency system that efficiently delivers recommendations while handling high traffic loads.

For both the above, implement the following:

- State all APIs created and assumptions made.
- Devise a way to figure out the user's location
- Design an architecture diagram for your backend system with all services. You need not draw the architecture provided in the assignment, just mention the connecting services properly.
- Design the schema for any Databases used for this.
- How would you scale this system to handle millions of users while maintaining a low-latency experience?

Section C: [Lecture Notes] 4x5

1. Explain the differences between the three primary cloud service models—Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) and provide 3 points for each and two example products for each
2. Your company is deciding between serverless (AWS Lambda) and containerized (ECS/EKS) architectures for a high-traffic web application.Explain:
    - The key differences in scalability, cold start latency between serverless functions and containers.
    - A scenario where serverless would be the better choice, and another where containers would be preferable.
3. In a distributed system with unpredictable traffic, how does a message queue system like SQS improve resilience and scalability?
   Explain:
    - How message queues prevent system overload and failures.
    - How they enable scalability across microservices.
    - A real-world use case demonstrating these benefits.

4. What is the use of Secondary Indexes in DynamoDB tables? List and explain the different types of secondary indexes available in AWS DynamoDB.

Section D: [Papers] 4x5

1. How does DynamoDB use data partitioning to handle scaling?

2. What is the key problem that Dremel handles? How is it different or compared to BigQuery?

3. Borg employs a priority and quota system for job scheduling. Describe how this system works and explain how it contributes to balancing resource allocation across production and non-production tasks.

4. If the master node in GFS faces a sudden increase in workload due to high-frequency read requests, what steps could be taken to prevent bottlenecks?