Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
Second Semester 2024-2025
End-Semester Test
(EC-3 Regular)

Course No.	: SE ZG583
Course Title	: Scalable Services
Nature of Exam	: Open Book
Weightage	: 40%
Duration	: 2 Hours 30 minutes
Date of Exam	:

No. of Pages = 2

Note to Students:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Ques1. Define scalability in the context of system architecture. How is it different from performance and availability? **[3Marks]**

Ques2. A large logistics company is planning to deploy **IoT-enabled sensors** across its fleet of delivery vehicles to achieve two key goals:

1. **Real-time tracking** of vehicle location, speed, and route deviation.
2. **Predictive maintenance** based on engine health, vibration analysis, fuel efficiency, and temperature data.

Each vehicle sends data to a central system every few seconds. The firm expects to scale from 100 to 10,000+ vehicles within the next 2 years. The system needs to support real-time decision-making (e.g., rerouting), immediate alert generation (e.g., engine overheating), and historical trend analysis for long-term planning.

To meet these goals, the firm is considering a **hybrid architecture** involving **edge computing**, **cloud scalability features**, **real-time stream processing tools (Kafka/Spark)**, and **distributed storage**.

a) Which data should be processed at the edge, and which should be sent to the cloud? Justify your choices with examples. **[3Marks]**
b) Identify the scalability challenges in collecting and processing data from 10,000+ devices in real time. **[2Marks]**
c) Discuss trade-offs in consistency vs. availability in the context of this system. When would you prioritize one over the other? **[2Marks]**

Ques3. Your organization handles credit card transactions in real time and must detect fraud within milliseconds.

a) Discuss what are the risks of using async messaging in fraud detection? **[2Marks]**
b) What types of caching techniques (local, distributed, or global) could be applied in a fraud detection system? How would you keep cache fresh while minimizing latency? **[3Marks]**

Ques4. **National Digital Health Ecosystem:** The Government of India is rolling out a digital health mission to enable citizens to manage their medical data across hospitals, clinics, diagnostic labs, and insurance providers. The system enables:

- Secure storage of health records (prescriptions, test results, discharge summaries)
- Booking appointments with doctors and diagnostics
- Initiating and tracking insurance claims
- Consent-based data sharing between stakeholders
- Real-time notifications to patients (e.g., lab result availability)

The platform is being re-architected using microservices for modularity, scalability, and maintainability. Each core functionality (e.g., patient profile, diagnostics, claims) will be handled by an independent service. The system must support millions of concurrent users, real-time updates, and ensure data privacy and security.

a) Identify at least four microservices for this application and explain how each one aligns with a distinct business function. **[3Marks]**
b) Which services would communicate synchronously, and which asynchronously? Justify your choices. **[3Marks]**
c) Explain how the Saga pattern could be used to manage the insurance claim approval workflow. Include at least one compensation scenario. **[4Marks]**
d) Should the team adopt a monorepo or multi-repo approach for this system? If multiple teams need to coordinate releases across services, how does your chosen repo pattern affect the build and deployment process? **[3Marks]**
e) Identify one potential sources of failure in the system and explain how these can cause cascading failures. **[2Marks]**

Ques5. Answer Briefly:

a) What is a service mesh, and how can it help manage observability, resiliency, and communication policies in this application? **[3Marks]**
b) Differentiate between unit testing, integration testing, and load testing in the context of Microservices. **[3Marks]**

Ques6. You are leading the DevOps team for an EdTech startup that offers live online courses. The platform includes services for user registration, live session scheduling, payment processing, and real-time chat. Due to growing demand and peak-time loads, the company has decided to containerize and deploy all services using **Kubernetes** for better scalability and manageability. The team also plans to implement a basic CI/CD pipeline and ensure role-based access to manage deployments securely.

a) How does **horizontal pod autoscaling** work in Kubernetes, and how would you use it to manage fluctuating loads on the **LiveSessionService**? **[2Marks]**
b) What commands are required in order to achieve the following tasks on a **local machine**
    i. Run Kubernetes cluster **[1Mark]**
    ii. Check the recent events for the resources in the Kubernetes cluster **[1Mark]**