**Microservices Communication and Load Balancing: Solutions and Challenges**

In a microservices architecture, how do services communicate with one another? Specifically, how does one microservice send a request to another? A straightforward yet impractical approach would be to store and hard-code IP addresses for each microservice. In real-world cloud environments, however, multiple instances of each microservice are typically deployed to support scalability and load balancing, with each instance having a unique IP address. This raises two important questions:

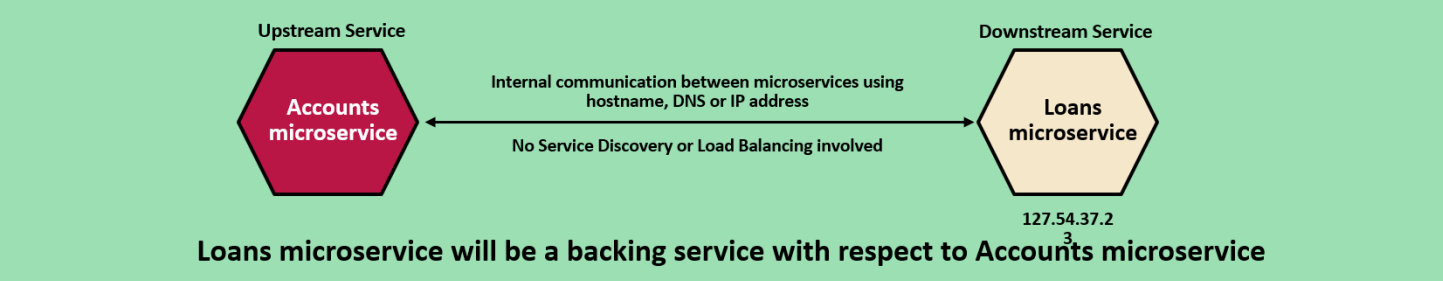
* How can microservices communicate efficiently in this dynamic environment?
* How do we ensure effective load balancing across instances?

To address these challenges, three key concepts are essential:

1. **Service Discovery**
2. **Service Registry**
3. **Load Balancing**

**Traditional Monolithic Communication**

In traditional monolithic applications within a network, if one service or app needs to communicate with another, it locates it through a specific IP address or DNS name.



For example, consider two services, *Accounts* and *Loans*. If there’s only one instance of the *Loans* service, the *Accounts* service can simply use a DNS name or IP address to communicate.

When a single instance of *Loans* is running, hardcoding the DNS name and IP address works fine. However, in a cloud-based microservices architecture, deploying multiple instances of a service is standard, each with its unique IP address.

One way to handle this is by using round-robin DNS, which associates multiple IP addresses with the service’s DNS name and rotates them for load balancing.

However, in microservices, where services and containers frequently change, this approach becomes inefficient due to the need for constant IP address updates. Therefore, dynamically managing service communication and load balancing requires Service Discovery and Registry mechanisms, which automate the process and adapt to changing instances.

**Core Concepts**

1. **Service Registry**
   * When a service instance starts, it registers itself with the service registry, providing essential details like IP address, port, and metadata. The registry helps maintain an updated catalog of all available service instances.
   * In simpler terms, whenever a new instance of a microservice is created, it registers itself with the service registry, supplying basic information such as IP address and port.
2. **Service Discovery**

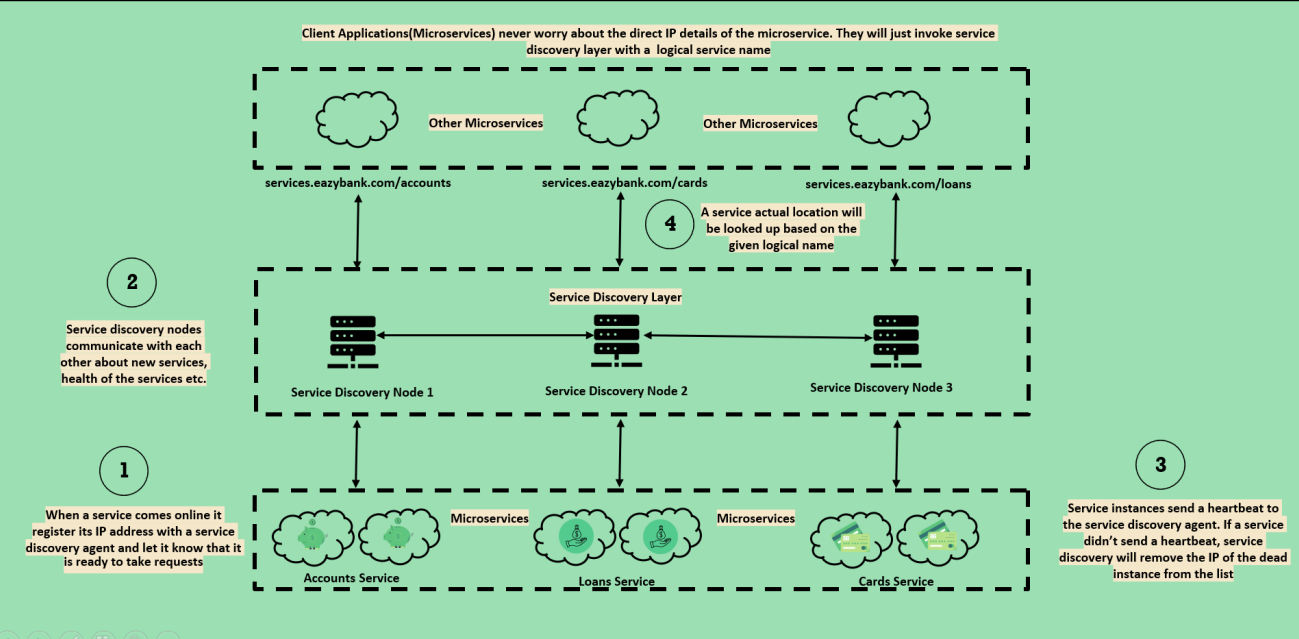
* When a client application needs to communicate with a specific service, it queries the service registry for available instances. The registry responds with the necessary details (e.g., IP addresses) to establish a connection.
* This enables the client to dynamically locate service instances as they start or stop, ensuring uninterrupted communication.

1. **Load Balancing**

* When a registry returns multiple IP addresses for a service request, the client application or a dedicated load balancer determines which instance to use. Load balancing can be based on algorithms such as round-robin or weighted distribution, ensuring an even distribution of requests among service instances.

**Client-Side Service Discovery Workflow**

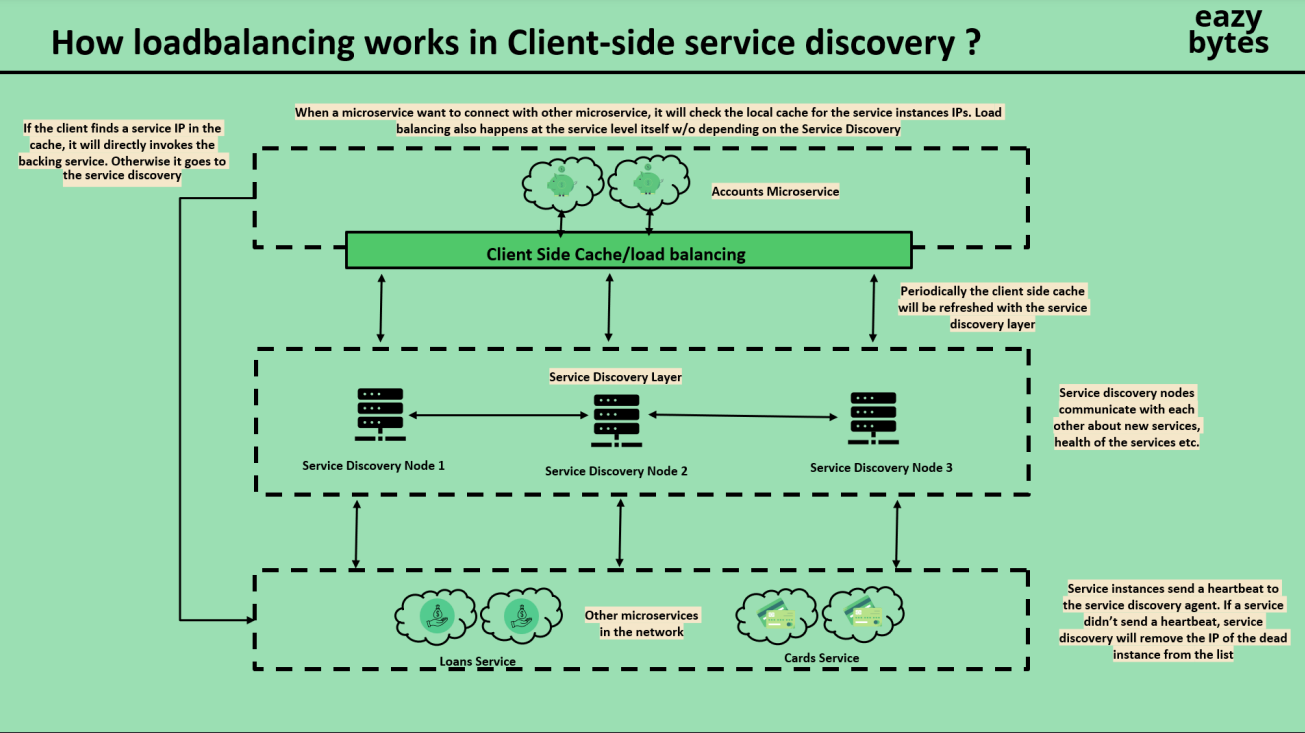
In client-side discovery, the client is responsible for selecting an available instance from the list of service instances returned by the service discovery mechanism.



1. A new service instance comes online and registers its IP address with the service discovery agent.
2. Service discovery nodes share updates about available services.
3. Service instances periodically send heartbeats to confirm their availability.
4. When a client application needs a service, it sends a request to the service discovery layer using the logical service name.
5. The service discovery layer looks up the service in its registry and returns the IP address of an available instance.
6. The client then connects directly to the provided IP address to interact with the service.

**Load Balancing in Client-Side Service Discovery**

In client-side load balancing, the client manages the selection of service instances based on a load-balancing strategy.



1. When a microservice needs to connect to another, it first checks its local cache for an IP address.
2. If the IP is cached, the service instance is directly invoked.
3. If the cache is empty or outdated, the client queries the service discovery layer.
4. The service discovery layer returns a list of healthy service instances.
5. The client selects an instance based on its load balancing algorithm, such as round-robin or least connections.
6. The client then connects to the selected instance and sends its request.

Create Service discovery agent microservices and for that add below dependencies:

* 1. eureka server
  2. Config Client
  3. Spring boot actuator

Now in EurekaServerApplication add annotation called @EnableEurekaServer to convert normal spring boot project to act as an service discovery agent

Go to application.yml file