Quick recap

The instructor discussed challenges in scaling a social media application, focusing on data storage and infrastructure management issues. They explored cloud computing solutions, particularly AWS services like EC2, and demonstrated how to set up and manage remote instances using SSH and elastic IPs. The discussion concluded with an overview of scaling approaches for distributed systems, comparing stateful and stateless systems, and introduced concepts of auto-scaling and load balancing for handling traffic spikes.

Next steps

- AlgoCamp: Upload remaining networking lecture videos by 2-3 PM
- Students: Watch the uploaded networking lectures in the prerequisite section by tonight/midnight
- Students: Review the shared Hotstar case study video about system scaling challenges
- Students: Review the previous class recording about vertical and horizontal scaling
- Students: Read about virtual machines and containers to understand security concerns in cloud computing
- AlgoCamp: Send Discord link to students who haven't received it within 24 hours
- Students: Join the Discord channel for further discussions and questions
- AlgoCamp: Post a message on Discord once all lecture videos are fully uploaded

Summary

Social Media Scaling Infrastructure Challenges

The instructor discusses challenges in scaling a social media application. They highlight two main issues: data storage and manual management of infrastructure. The first problem is determining how to store and retrieve data across multiple servers, as it's unclear whether each server should have its own database or if data should be distributed. The second issue is the manual effort required to add new servers and maintain existing ones as the application scales, which can be time-consuming and costly.

Cloud Computing Benefits for Businesses

The discussion focuses on cloud computing and its benefits for businesses. Cloud computing companies like AWS, GCP, and Azure provide data centers with numerous machines that can be rented by other businesses. This allows companies to focus on their core business logic without managing physical infrastructure. Cloud providers offer both bare machines and managed services, where pre-configured setups are available for common needs like databases or load balancers. The speaker demonstrates how to access the AWS console and introduces EC2 (Elastic Compute Cloud) as a service for renting computing resources from AWS.

AWS Global Infrastructure Overview

AlgoCamp explains the concept of AWS global infrastructure, including regions, availability zones, and edge servers. He demonstrates how to rent a machine in a different location, such as London, while being in India. AlgoCamp discusses the potential latency issues with cross-continent connections but highlights the benefits of deploying servers in multiple locations for reliability. He then guides through the process of launching an EC2 instance, selecting an operating system (Ubuntu), and choosing the machine configuration, emphasizing the importance of selecting a free tier option like t2.micro to avoid charges.

Understanding AWS Key Pair Access

AlgoCamp explains the concept of key pairs for accessing AWS instances remotely. He describes how a key pair file acts as a security credential, similar to a password for a laptop. The instructor demonstrates creating a new key pair, saving it locally, and launching an EC2 instance. He then shows how to start, stop, and restart the instance, noting that the public IP address disappears when the instance is stopped and a new one is assigned when it's restarted.

Elastic IPs in AWS Explained

AlgoCamp explains the concept of elastic IPs in AWS, which are static IP addresses that can be associated with EC2 instances. He demonstrates how to allocate, associate, and release an elastic IP, showing that it remains constant even when an instance is restarted. AlgoCamp also clarifies the difference between public and private IPs, noting that private IPs are for internal AWS network communication. He emphasizes that elastic IPs provide a solution for maintaining a consistent IP address for services like websites, addressing the problem of changing IPs with dynamic addressing.

AWS EC2 Remote Access Tutorial

AlgoCamp discusses how to connect to an AWS EC2 machine from a local machine. He mentions that there are various options available in AWS, including IAM roles, subnets, and auto-scaling groups, which will be covered in future lectures. AlgoCamp plans to demonstrate how to connect to the AWS machine after a short break, emphasizing the need for some form of wireless network access to control the EC2 instance remotely.

Understanding SSH and Network Protocols

AlgoCamp explains that different types of communication in computer networks require different protocols, similar to how languages have different rules. He introduces SSH (Secure Shell) as a protocol for remote access to servers, allowing command-line execution from one machine to another. AlgoCamp demonstrates how to use a terminal and mentions that protocols are implemented in operating systems or applications. He prepares to show how to SSH into a machine using a previously downloaded key.

SSH and AWS Remote Access

The instructor demonstrates how to access and manage an AWS Ubuntu machine remotely using SSH. They explain the importance of changing the key permissions to read-only for security reasons, using the chmod command in Linux. The instructor then shows how to connect to the remote machine using SSH, providing the syntax for the command including the pem key, username, and IP address. Once connected, they demonstrate running basic commands on the remote machine, such as updating packages and running Python. The instructor clarifies that the remote machine uses its own internet connection provided by the AWS data center, independent of the user's local internet.

AWS Auto-Scaling and Load Balancing

The speaker discusses AWS's auto-scaling feature, which automatically adjusts the number of EC2 machines based on load, and the use of elastic load balancers to prevent single points of failure. They explain how companies like Ticketmaster and Hotstar use cloud infrastructure to focus on their core business logic rather than managing infrastructure. The speaker also shares a video about Hotstar's engineering challenges, highlighting the need to consider unexpected traffic spikes, such as when millions of users simultaneously leave a live stream and return to the home screen.

Distributed Data Storage System Approaches

The discussion focuses on different approaches to data storage in a distributed system. Initially, the idea of replicating data across all machines is considered but deemed inefficient due to the overhead of manual data copying. Instead, a stateless system is proposed, where a separate database machine handles all data storage and retrieval for the application servers. The conversation then shifts to the challenges of database scaling and the potential need for a stateful system in certain applications, such as AI chatbots. An example of Chat GPT is used to illustrate how a stateful system with sticky load balancing and local data storage can improve performance by keeping user context readily available and reducing database queries.

Stateful vs Stateless Systems Overview

The instructor discusses stateful and stateless systems, explaining that stateful systems maintain user-specific information across machines, while stateless systems are more common and easily scalable. They provide examples of vertical and horizontal scaling, comparing them to improving one friend's strength versus having a group of friends. The instructor also clarifies the differences between SSH and VPN, and mentions that the next class will cover HTTP and APIs. They inform students that networking lectures are being uploaded and should be available soon, encouraging them to watch these for better understanding of networking concepts.