Cloud-based Data Management

Auto-scaling

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

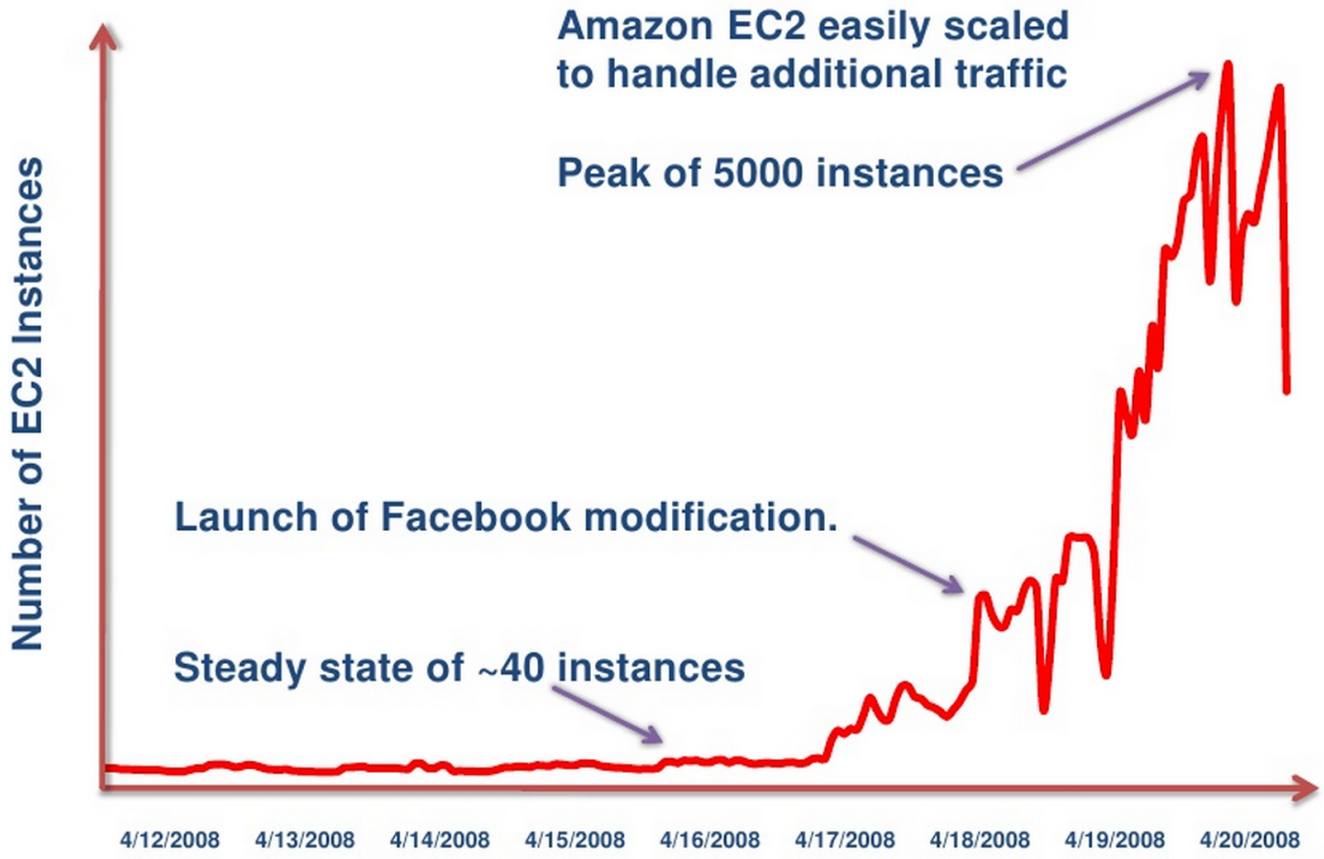
Objective

Explain AWS programming interfaces

Build an Elastic Application on IaaS

Elastic cloud application

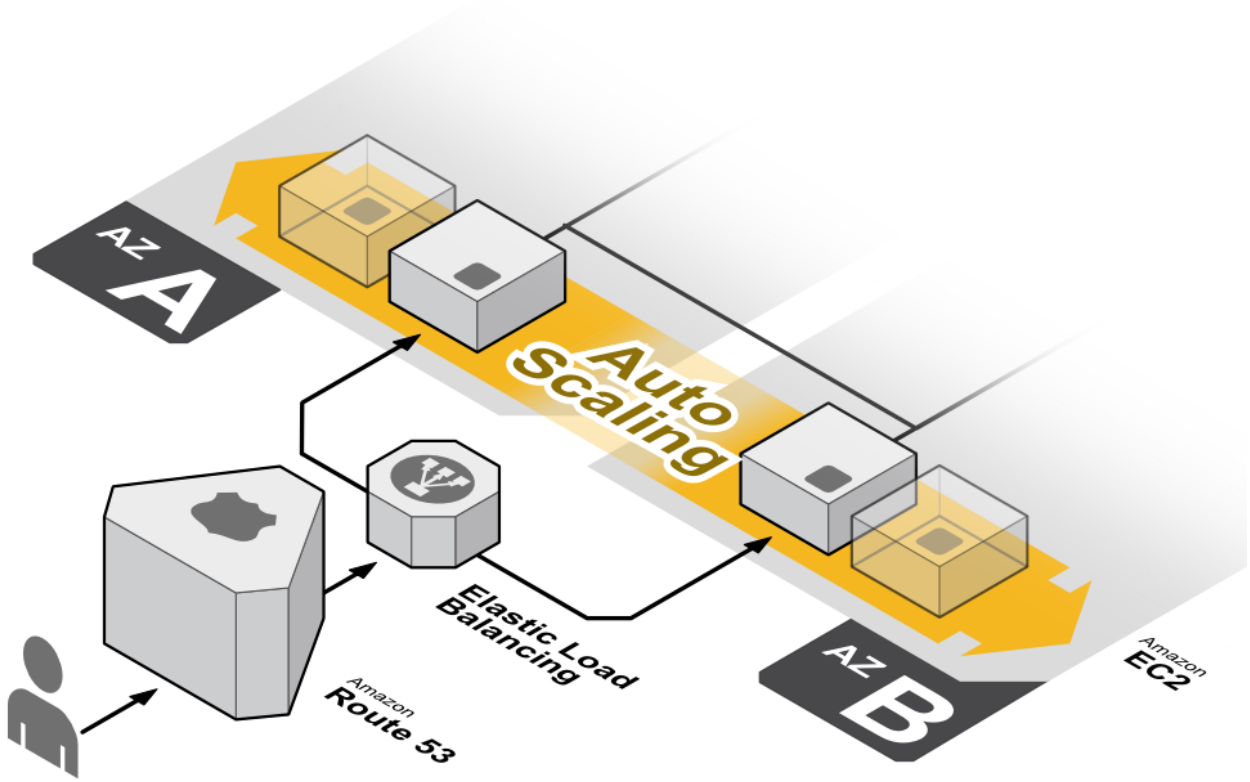
- Acquire cloud resources on-demand (compute and storage)
- Make efficient use of the rented resources
- Relinquish unneeded resources timely



Auto Scaling

AWS has limited support for auto scaling

- Use CloudWatch to monitor utilization of instances
- Trigger scaling when utilization is high
- Create new instances to handle the load
- Use Elastic Load Balancing to distributed the load across instances
- And vice versa



Scaling based on SQS

- SQS queues the requests received by the web server and to be processed by the processing servers
- Use the length of the queue to determine the load and decide auto scaling