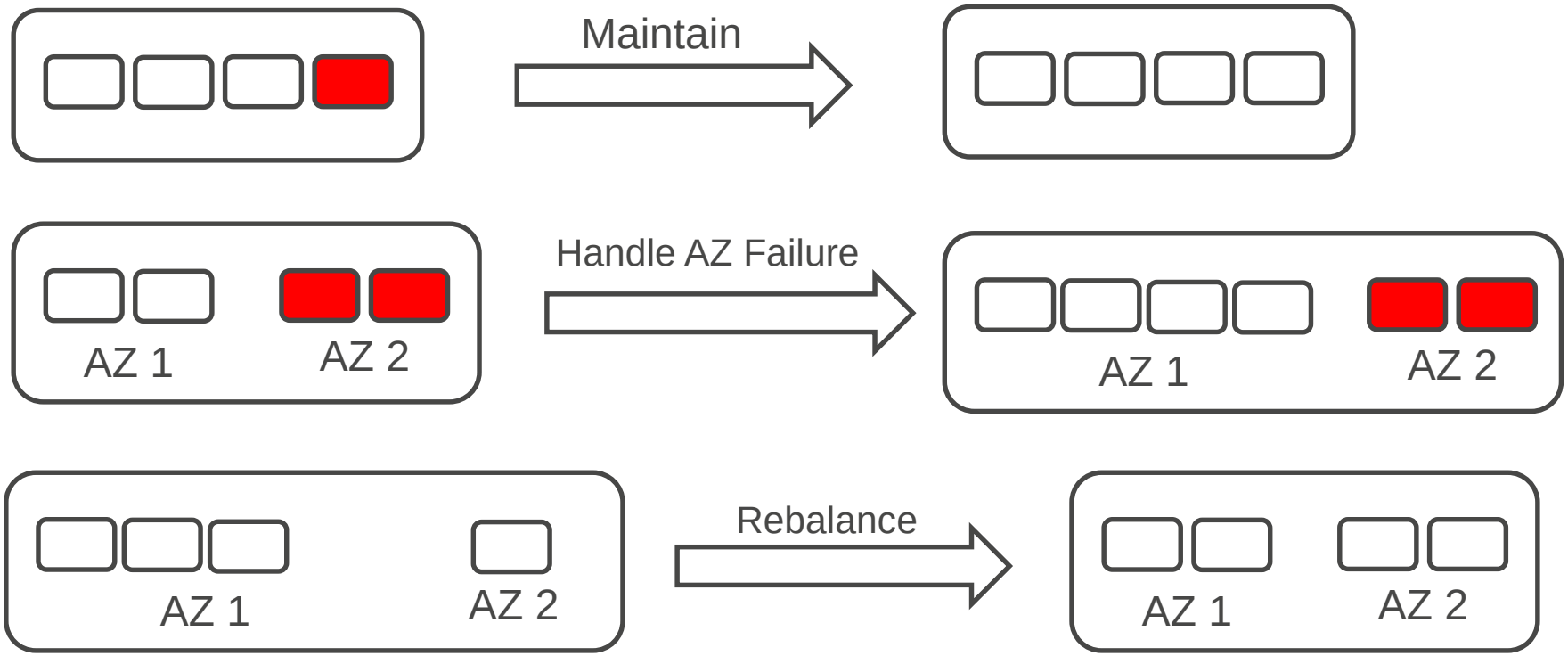
Auto Scaling

Maintain Right Compute Capacity

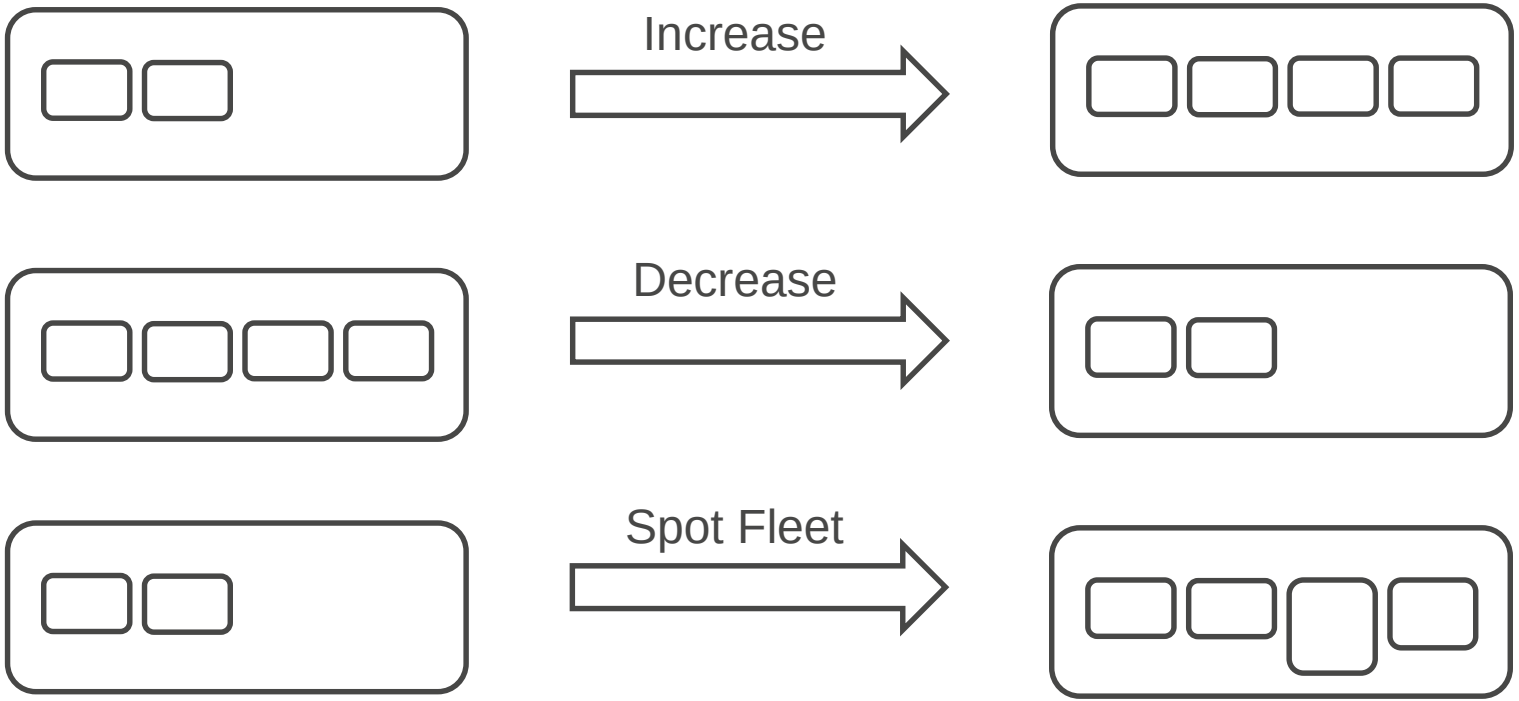
Chandra Lingam

Cloud Wave LLC

Auto Scaling



Auto Scaling



Lifecycle Hooks

"Lifecycle hooks let you take action before an instance goes into service or before it gets terminated."

"For example, launch hooks can perform software configuration on an instance"

"Terminate hooks can be useful for collecting important data from an instance before it goes away. For example, you could use a terminate hook to preserve your fleet’s log files by copying them to an Amazon S3 bucket when instances go out of service"

https://aws.amazon.com/ec2/autoscaling/faqs/

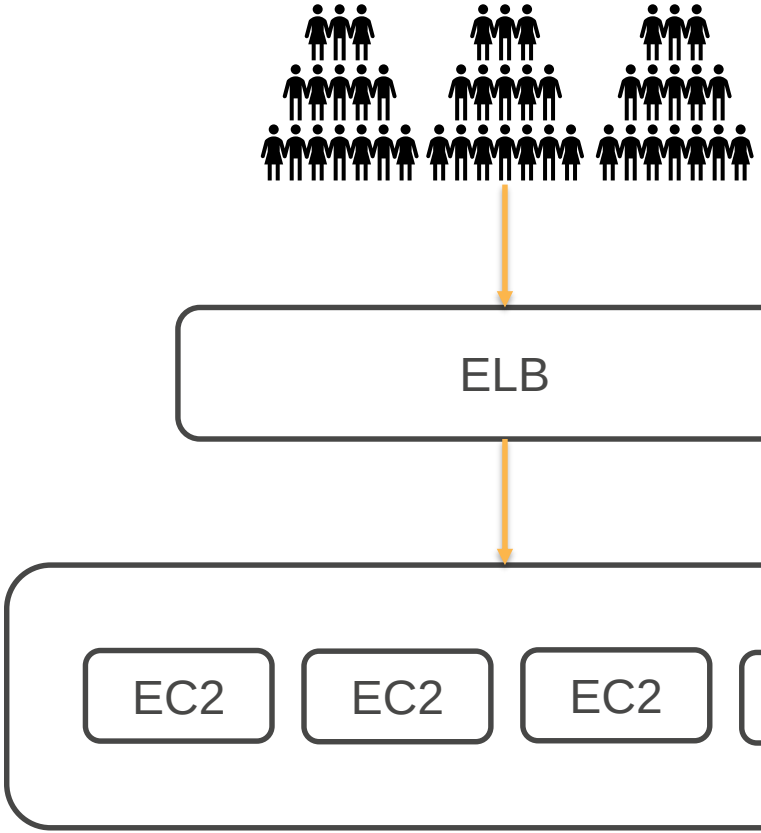
Auto Scaling Group

Criteria	Description
Minimum	Minimum number of instances that are always needed
Maximum	Maximum number of instances that are allowed
Desired	• Slider moves between minimum and maximum • Number of instances that are required for current traffic

Scaling Types

Scaling Type	Description
Maintain	Maintain a constant fleet size
Scheduled	Time based – add, remove
Dynamic	Metrics based – respond to changes to traffic
Predictive	Machine Learning based – adjust based on predicted traffic

Auto Scaling



- Request Count

Services with Auto Scaling Support

- EC2 – Adjust EC2 instance capacity based on need
- EC2 Spot Fleets – optimize cost, replace instances that are interrupted due to price or capacity reasons
- Elastic Container Service – adjust tasks based on load, adjust instances in the cluster
- DynamoDB – Adjust provisioned read and write capacity
- Aurora – Adjust the number of read-replicas based on workload

Auto Scaling Benefits

- Improve Fault Tolerance
- Increase Application availability
- Lowers Cost
- Support for multiple purchase models, instance types, Availability Zones
- Balance across Availability Zones
- Customer needs to check for account service limits

Lab – Auto Scaling and Load Balancer

Launch an Application Load Balancer

Define Launch Configuration for Web Servers

Configure Auto Scaling to launch and register instances with load balancer

Lab – Maintain Fleet Size

Simulate server and application failures

Verify how Auto Scaling and Load Balancer responds to errors

Lab – Manual Scaling and Dynamic Scaling

Add additional capacity with manual scaling

Use Dynamic Scaling to adjust capacity

Chandra Lingam

50,000+ Students

Up-to-date Content