zhgeeks meetup

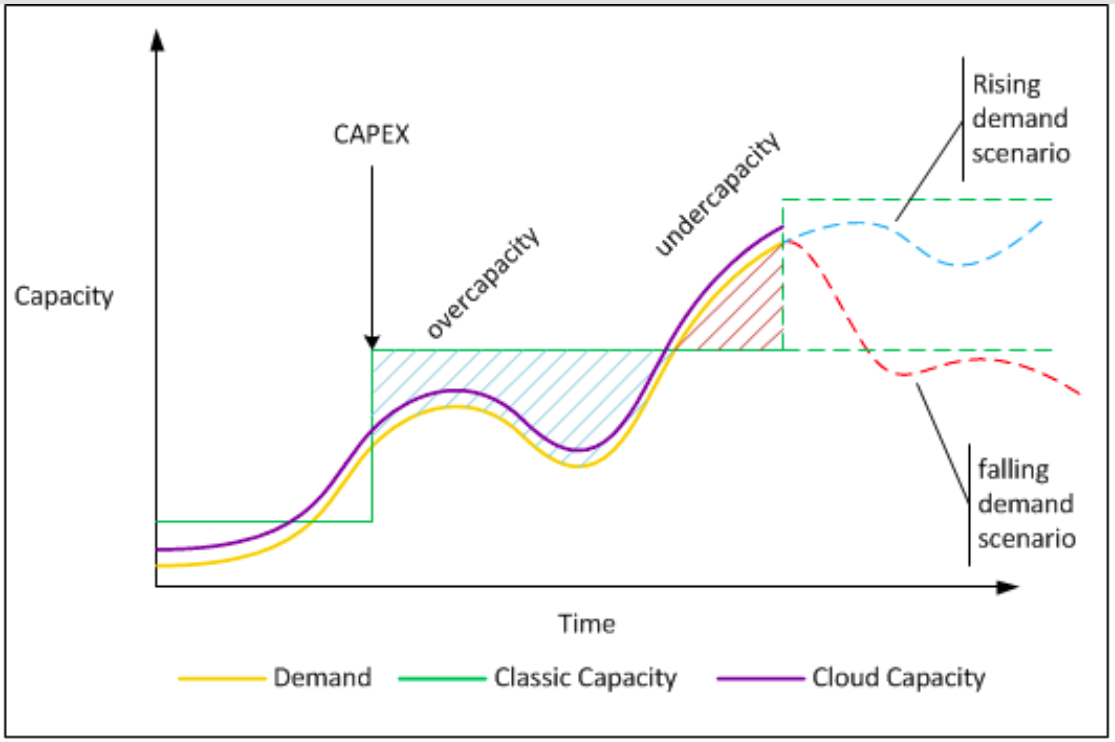
15-May-2012

Elastic load balancing and auto-scaling

https://github.com/zhgeeks/presentations/

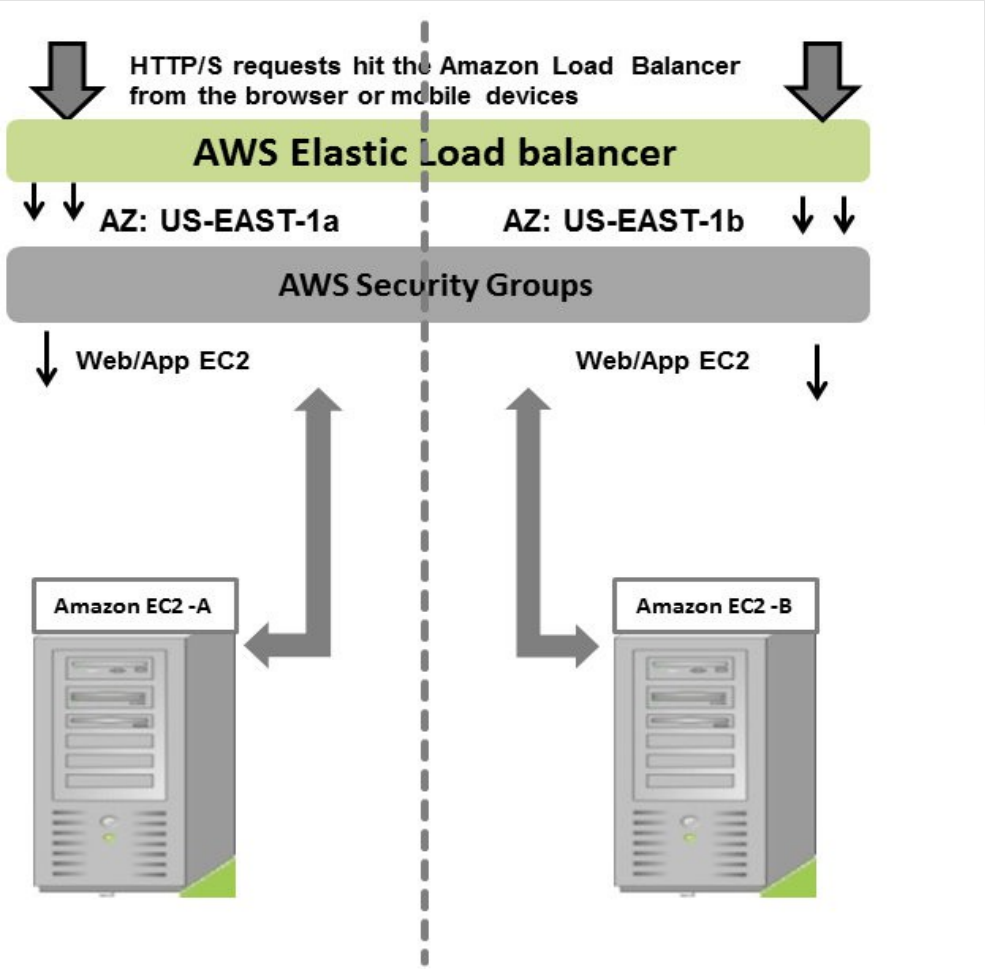
Muharem Hrnjadovic <mh@star.io>
@al_maisan

Motivation



http://www.chades.net/wp-content/uploads/2010/05/capacity-vs-demand.png

“Agenda”



http://harish11g.blogspot.com/2012/02/elastic-load-balancing-aws-deployment.html

Questions

- How do I make sure my
  - system scales (up|down)?
    - at given times (predictable traffic)
    - with more|less load (elastic behaviour)
  - ec2 instance fleet stays healthy?

Ingredients

- CloudWatch

  “monitor, manage, and publish various metrics, as well as configure alarm actions based on data from metrics”

- Auto-scaling

  “launch or terminate EC2 instances automatically based on user-defined policies, schedules, and health checks”

- Elastic Compute Cloud (EC2)

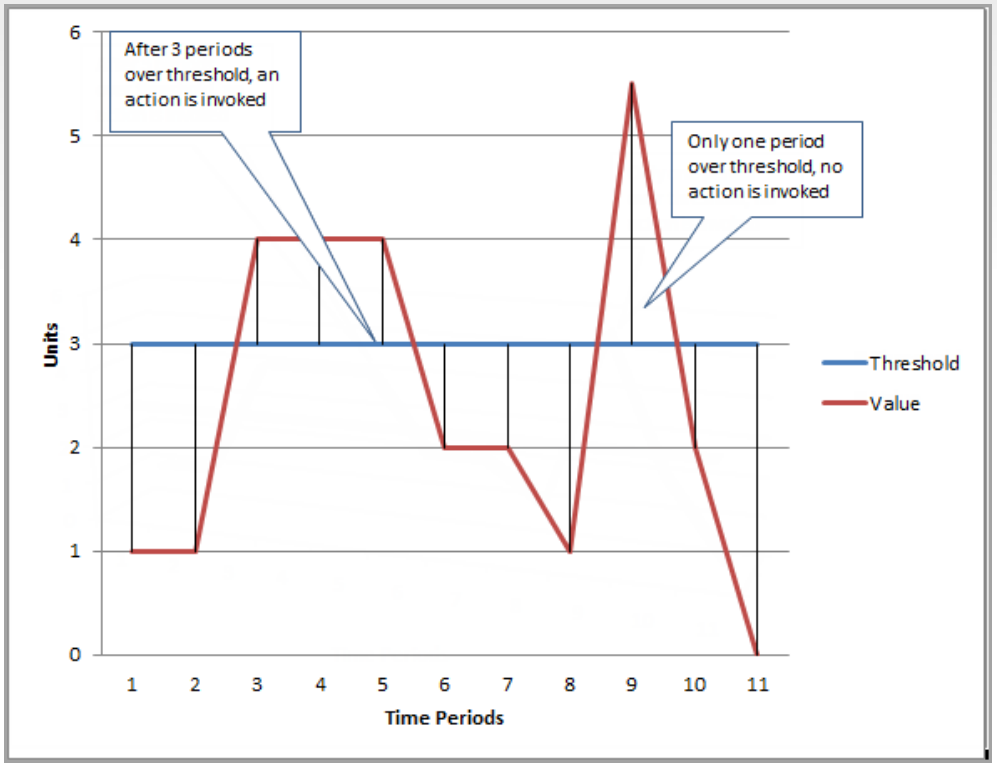
  “provides resizeable computing capacity, use and pay for only the capacity that you need”

CloudWatch

- Available metrics vary per EC2, EBS, ELB etc.
- EC2 instances
  - CPUUtilization
  - Disk reads/writes
  - Network in/out
- ELB
  - Latency
  - Request count
  - Healthy/unhealthy host count

Alarms

- States: `OK, ALARM, INSUFFICIENT_DATA`



http://docs.amazonwebservices.com/AmazonCloudWatch/latest/DeveloperGuide/images/AlarmsGraph.png

Auto-scaling

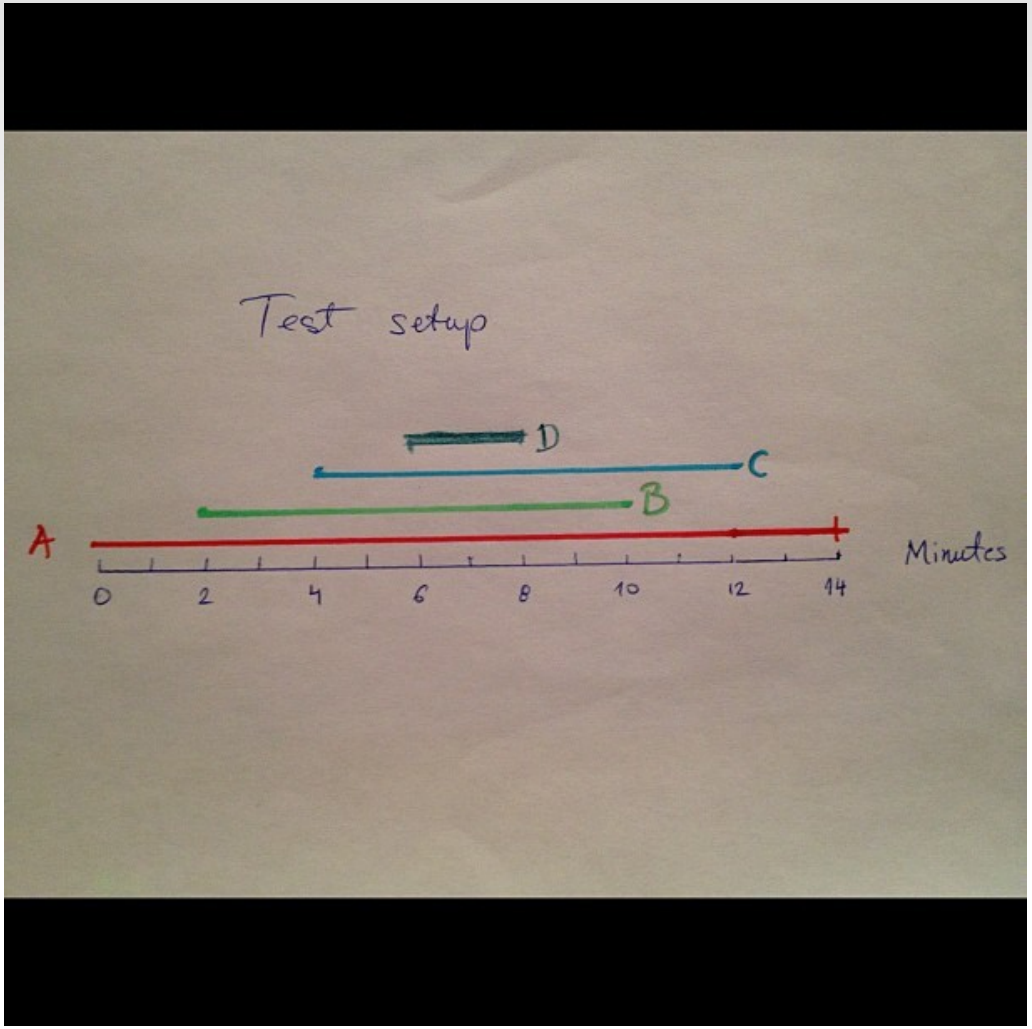
- Launch configuration +
- Load balancer(s) +
- Scaling policies +
- Alarms

- Can be turned off (“big red button”)

Scaling policies

- Adjustment type
  - ChangeInCapacity, ExactCapacity or PercentChangeInCapacity
- ScalingAdjustment: number of instances by which to scale
- Cool down: number of seconds to hold off with next scaling action

Demo time !!



References

- Auto Scaling Documentation
- Amazon CloudWatch Documentation
- Amazon EC2 Documentation