

Cloud Cost Optimization at Scale

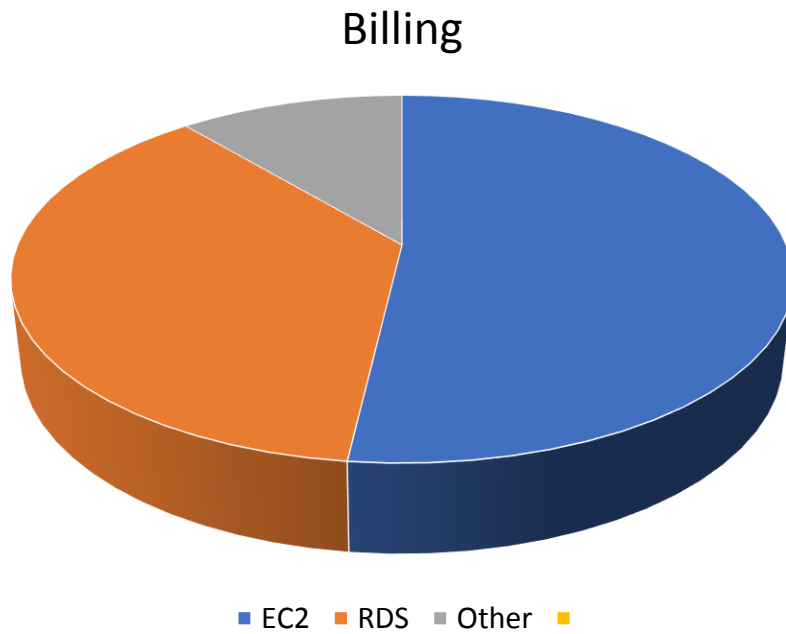
How we use Kubernetes and spot instances
to reduce EC2 billing up to 80%

Tuan Anh Tran – Engineering Manager @ Mytour
anhtht@mytour.vn / github.com/tuananh

Scalability

- Performance scalability
- Functional scalability
- Programming scalability
- **Cost scalability** ? ? ?

Monitoring your usage & cost






Cost optimization

- We will focus on the biggest pieces of the pie.
- We will take care of the low hanging fruit first.

Types of EC2 instances

- On-demand instances
- Reserved instances
- Spot instances

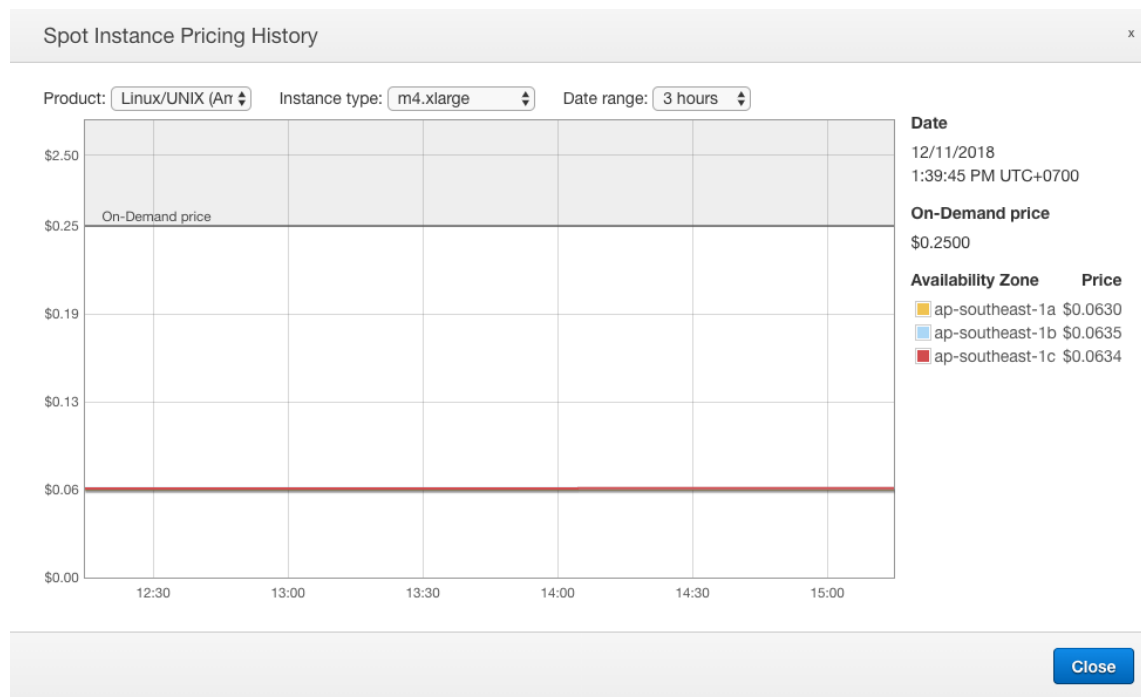
Reserved Instances

- Commitment.
- 30-60% off On-demand instances. 
- Capacity reservation.
- Cannot be cancelled. 
- Can be resale via Reserved Instance marketplace. 

Spot instances

- Average 70-90% lower than On-Demand instance ? ? ?
- Prices are volatile. ?
- Prices are varied between instance types and availability zones.
- Can be terminated anytime with 2 mins notice ☠☠☠

Spot instance pricing



m4.xlarge: 76% saving
m5.xlarge: 70% saving
m5.4xlarge: 85% saving
etc...

Spot instances are no-go



- Re-deploy applications when instances got shutdown.
- Application are stateful.
- What if all spot instances got shutdown.

Kubernetes

> Kubernetes is an open-source system for automating deployment, scaling, and management of containerized applications.

<https://kubernetes.io>

Kubernetes features

- Orchestrate workload 
- Self-healing 
- Horizontal scaling
- Automated rollout and rollback
- Service discovery
- Load balancing
- Secret management
- etc...

When k8s meets spot instances (1)

- Multiple autoscaling groups with different instance type and zone.
- Older instance types are less volatile.
- Mixed use of reserved instances and spot instances.
- Watching k8s API to scale on-demand autoscaling group when needed.

When k8s meets spot instances (2)

- Daemonset on spot instances to watch if the server is shutting down, drain it and migrate app to other nodes.

```
> curl http://169.254.169.254/latest/meta-data/spot/termination-time
```

Other tips

- We work 40 hours per week, out of 168 hours. If we can turn off the dev/staging environment, we can instantly save 76% cost.

But we don't use AWS. Similar offerings?

- Google's Preemptible VMs
- Azure's low priority VMs
- Aliyun's ECS spot market

Result

- With k8s and spot instances, we were able to shave off 80% of our EC2 billing cost.
- Maintain high availability of the whole infrastructure.

Btw, we're hiring!

<http://career.mytour.vn/tim-viec-lam/tat-ca-viec-lam/vi>

STT	Chức danh	Nơi làm việc
1	Visual Designer	Hà Nội Hồ Chí Minh
2	UX/UI Designer	Hồ Chí Minh
3	PHP Developer	Hà Nội
4	SENIOR FRONTEND DEVELOPER	Hà Nội
5	Android Developer	Hà Nội
6	iOS Developer	Hà Nội
7	Senior Accountant	Hà Nội
8	Testing Intern	Hà Nội
9	Tester - Attractive salary	Hà Nội

Questions?