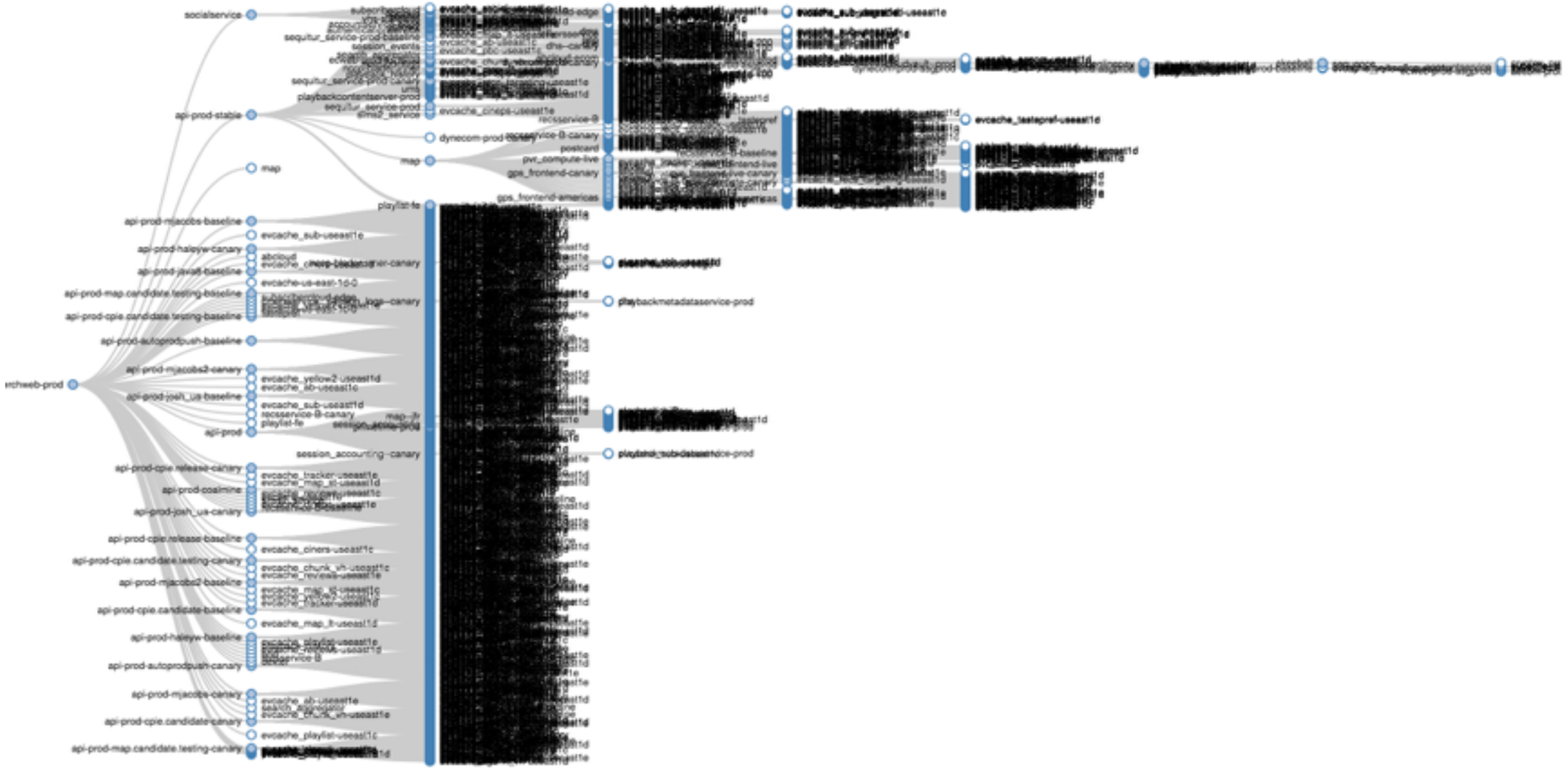
Reliably shipping containers in a resource rich world using Titan

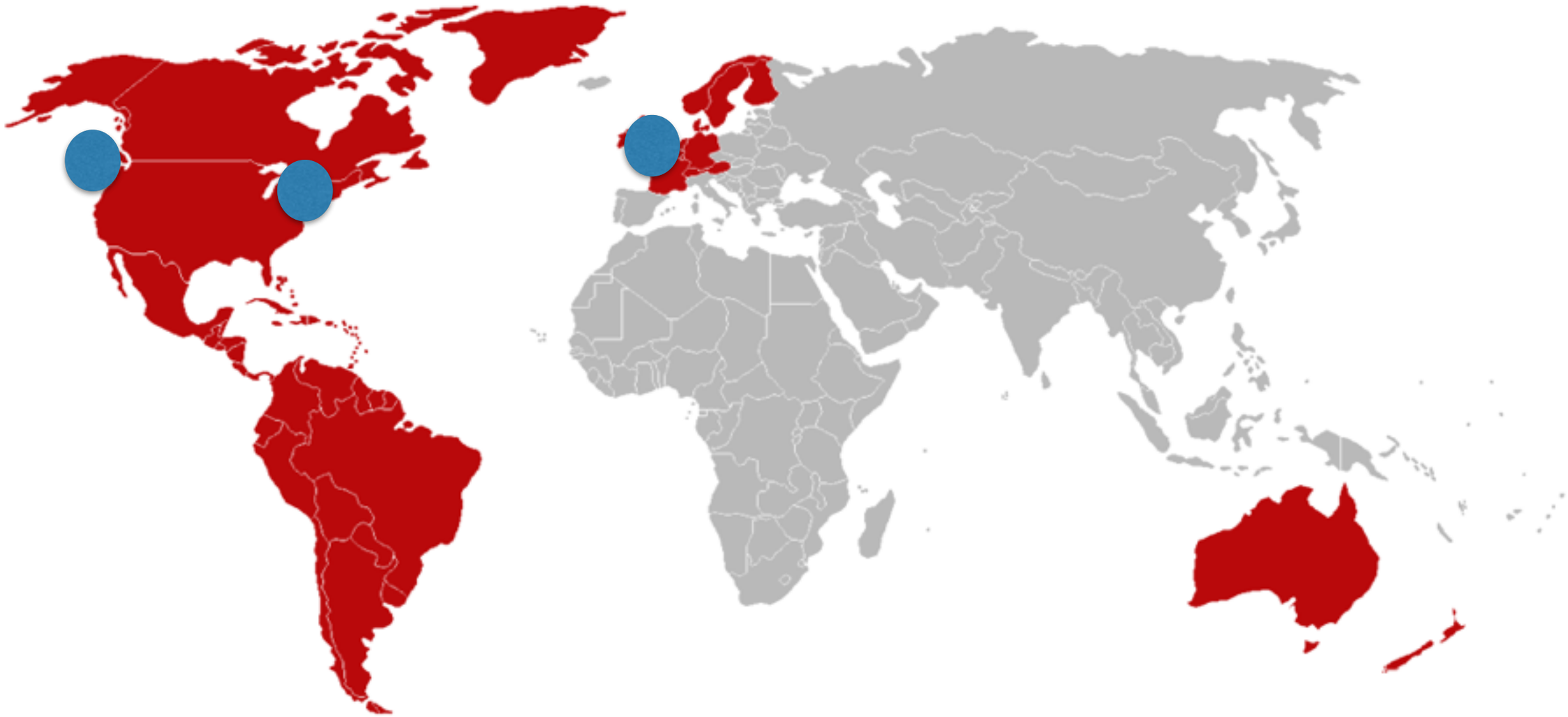
Diptanu Choudhury
Software Engineer, Netflix
@diptanu

dockercon 15

A Cloud Native Application built on Micro Services Architecture


socialservice
api-prod-stable
map
api-prod-mjacobs-baseline
api-prod-haleyw-canary
api-prod-javad-baseline
api-prod-map.candidate.testing-baseline
api-prod-cpie.candidate.testing-baseline
api-prod-autoprodpush-baseline
api-prod-mjacobs2-canary
mcweb-prod
api-prod-josh_ua-baseline
api-prod
session_accounting--canary
api-prod-cpie.release-canary
api-prod-coalmine
api-prod-josh_ua-canary
api-prod-cpie.release-baseline
api-prod-cpie.candidate.testing-canary
api-prod-mjacobs2-baseline
api-prod-cpie.candidate-baseline
api-prod-haleyw-baseline
api-prod-autoprodpush-canary
api-prod-mjacobs-canary
api-prod-cpie.candidate-canary
api-prod-map.candidate.testing-canary
sequtur_service-prod-baseline
session_events
sequtur_service-prod-canary
playbackcontentserver-prod
sequtur_service-prod
dynecom-prod-canary
postcard
pvr_compute-live
gps_frontend-canary
gps_frontend-americas
playlist-fe
evcache_sub-useast1e
evcache-us-east-1d-0
evcache_yellow2-useast1d
evcache_sb-useast1c
evcache_sub-useast1d
recsservice-B-canary
playlist-fe
session_accounting--prod
evcache_tracker-useast1e
evcache_map_st-useast1d
evcache_ciners-useast1c
evcache_chunk_sh-useast1c
evcache_reviews-useast1e
evcache_map_st-useast1c
evcache_tracker-useast1d
evcache_map_lt-useast1d
evcache_playlist-useast1e
recsservice-B
evcache_ab-useast1e
search_aggregator
evcache_chunk_sh-useast1e
evcache_playlist-useast1c
evcache_sub-useast1e
evcache_testopref-useast1d
playbackmetadataservice-prod

Multi- Geography



Active-Active

Titan is a Compute service distributed across multiple geographies

A need for a common resource scheduler for domain specific distributed systems

The operational benefits of a PaaS without the dilemmas of sandboxing technologies.

Consistent tooling and operational control plane for SREs across all technology stacks

Faster turn around time from development to production

Auto Scaling Groups are harder to adopt for event based orchestration systems

Increasing density of application processes per server


docker
dockercon
15

Why we chose Docker

- Process isolation
- Immutable deployment artifacts
- Ability to package dependencies of an application in a single binary
- Tooling around the runtime for building and deploying
- Scalable distribution of binaries across clusters

Docker Containers are the deployment artifacts and process runtime for Titan

The Titan API

```
{
  "name": "MixTape",
  "applicationId": "12-120-34-56-89",
  "image": "mixtape-release",
  "numberOfTasks": 250,
  "version": "1.0.1",
  "cpu": 6,
  "memory": 4096,
  "disk": 20,
  "hardDrain": true,
  "numberOfIps": 2,
  "ports": [8080, 8081],
  "numberOfRestarts": 10,
  "maxAge": "48h",
  "numberOfConcurrentJobs": 20,
  "entryPoint": "/apps/mixedTape/bin/start.sh",
  "env": { "foo": "bar", "pipe": "baz" },
  "location": { "dc1": 10, "dc2": 5, "dc3": 5 }
}
```

AutoScaling

- Two Levels of Autoscaling
  - Scaling of underlying compute resources
  - Application Scaling based on business and performance metrics

AutoScaling

- Two Types of Autoscaling
  - Predictive
    - Titan scales up infrastructure based on historical data on statistical modeling.
  - Reactive
    - Scaling activities are triggered based on pre-defined thresholds

Disk

- Titan manages volumes for containers.
- We use ZFS on Linux to create volumes
- Data volumes are mounted within containers

Logging

- Titan allows users to stream logs of a Task from a running container in a location transparent manner
- Logs are archived off-instance and Titan provides API to stream logs of finished tasks

Network

- In EC2 Classic, Titan exposes ports on containers on the host machine.
  - Mesos is used as a broker for port allocation

Network

- In VPC, every container gets its own IP address.
  - Mesos is completely out of the picture
  - We use ENIs and move them into the network namespace of containers
  - Developing a custom network plugin

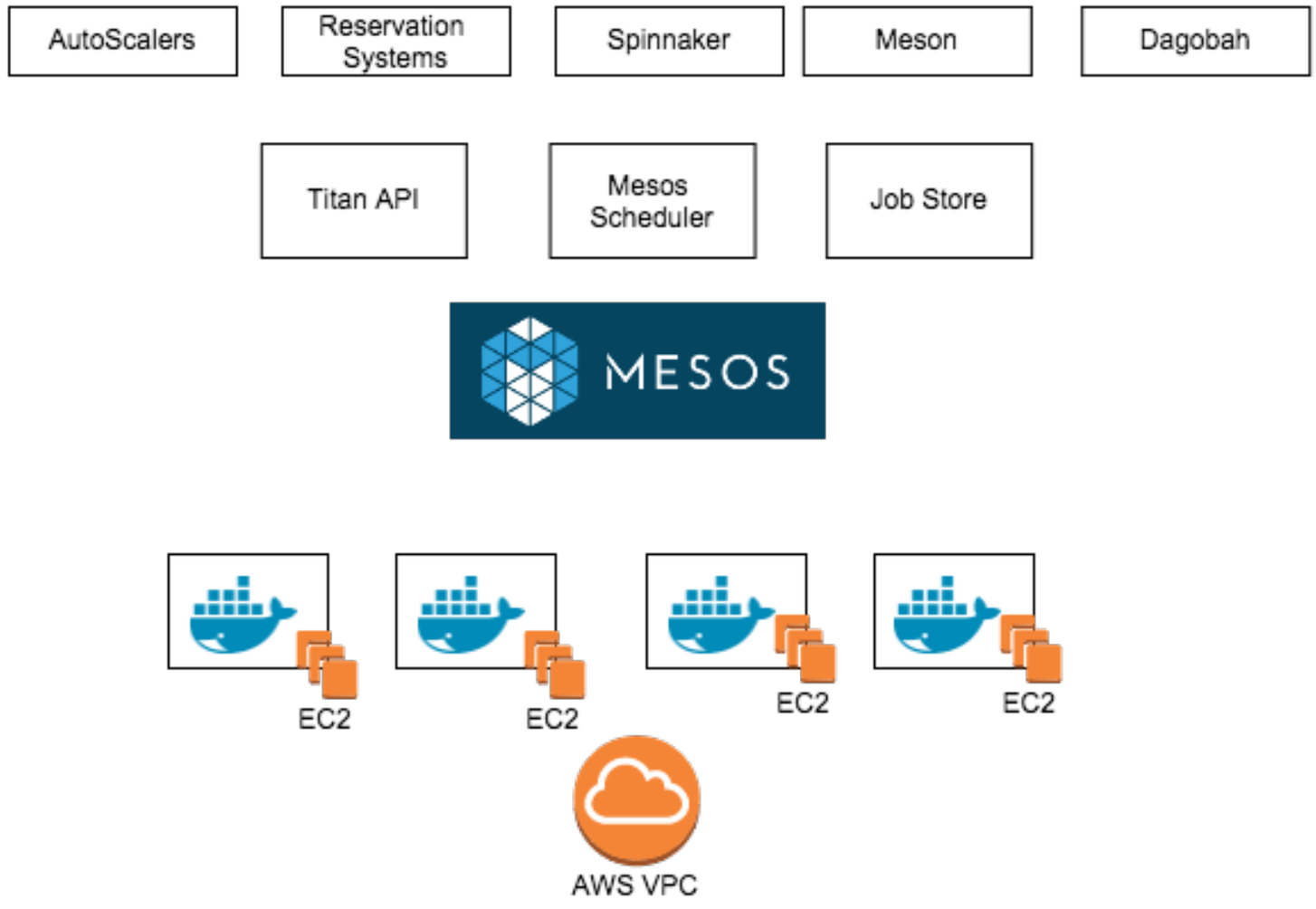
Monitoring

- cgroup metrics published by the kernel are pushed to Atlas.
- Users can see all the cgroup metrics per task.
- cgroup notification API for alerting

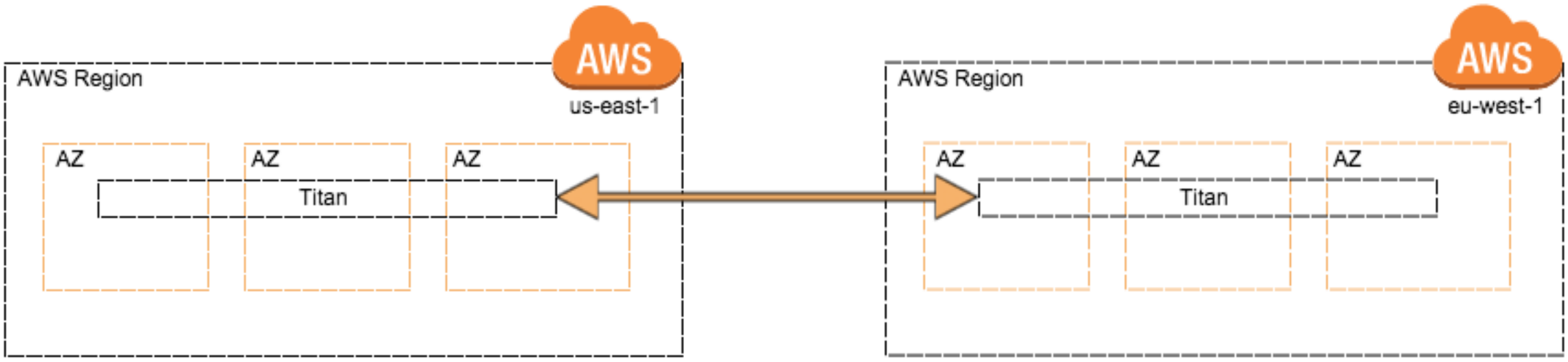
Failover

- Titan allows SREs to drain a cluster of containers into newer compute nodes
- Underlying VMs are automatically terminated when containers crashes for hardware/OS problems
- Allows failover across multiple data centers

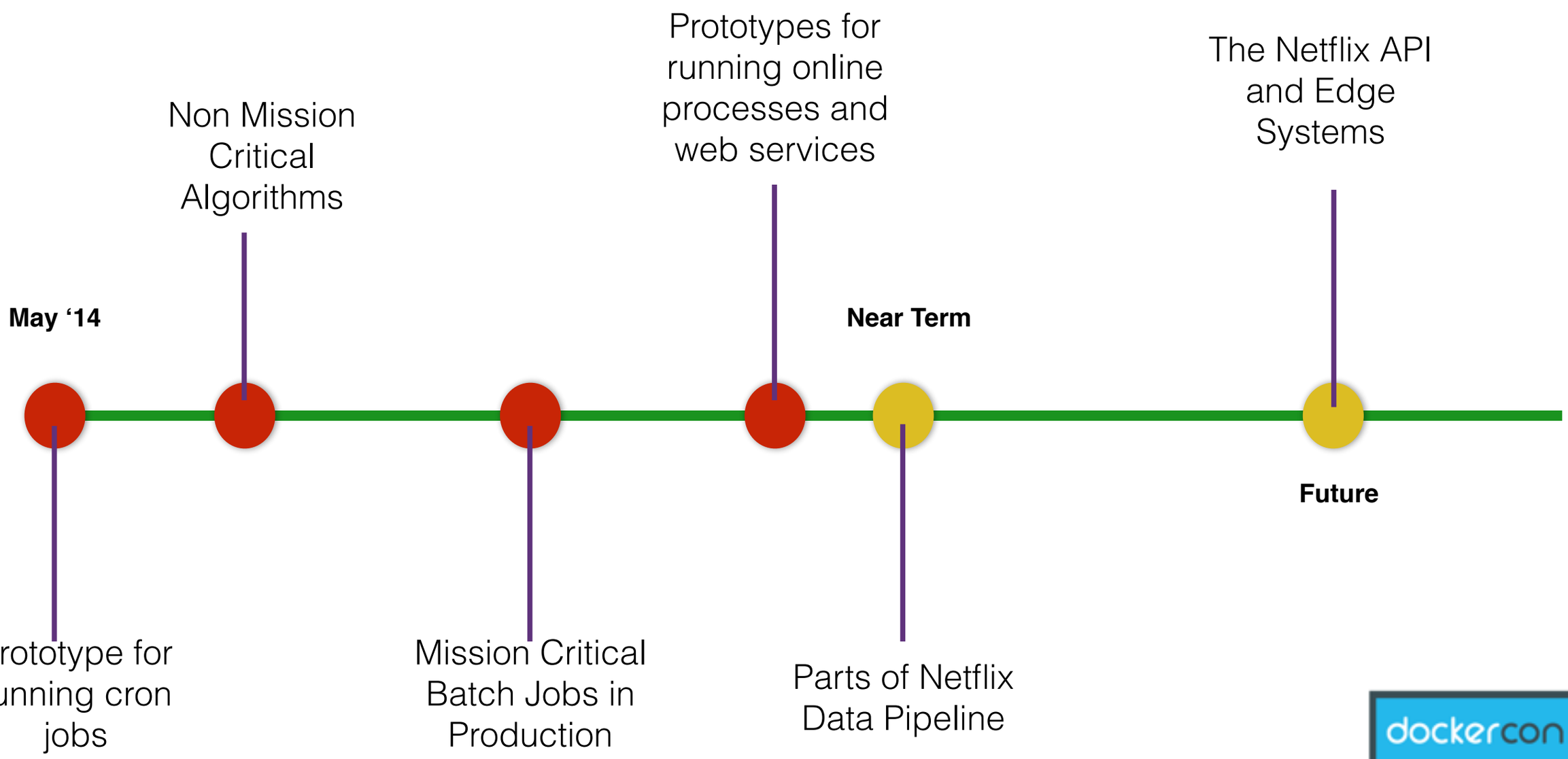
From a 1000 Feet



From a 5000 Feet



Where are we with Titan at Netflix



Thank you

Diptanu Choudhury

@diptanu

www.linkedin.com/in/diptanu