Smart Workloads: Automated Routing, Scaling of K8s and Serverless Functions

Enlin Xu, Director of Advanced Engineering, Turbonomic
• Who are we
• Challenges to deploy and operate applications in Cloud
• Smart Workloads in Edge Computing
• Demo
• Key Takeaways
Who are we?

- Founded in 2009
- Headquartered in Boston, MA, USA
- 500+ employees
- $100m+ Revenue in 2017

Founded on the idea that software should manage IT resources, not people. Our software uses a Common Abstraction with Economic Principles to unleash automation!

Kelsey Hightower
@kelseyhightower

Here is my application, run it for me, when and where I want it, securely. That's the end game.

10:21 AM - 22 May 2018

We’re Hiring!

Join our engineering team as we build for the end game.

IDC Innovator: Multicloud Management, 2017
I want to run some code in cloud.. And?

- Operational complexity
  - VM – Configure machines, storage, network, OS, etc.
  - Container – Configure application, scaling, etc.
  - Serverless – Memory configuration, where to run, etc.

- Capacity management concerns
  - Cloud has infinity capacity – what about your budget?

- Scalability – I want to have infinite scale!!
  - But only pay for what you really need
Serverless looks simple enough! Or is it?
Serverless Use Cases

- An image upload to S3 triggers a Lambda to compute a thumbnail to post on Twitter
- The same upload triggers another Lambda to analyze the image and create metadata such as place, persons, etc.
- A code commit or PR merge triggers build jobs
- An IoT sensor on a garbage can detects its fullness and triggers a function to put the garbage can on a pickup schedule
- An ML scheduler kicks off a series of MapReduce worker functions
- …
Complex Serverless Chains

https://hackernoon.com/capture-and-forward-correlation-ids-through-different-lambda-event-sources-220c227c65f5?gi=48f5de1450fa
Challenges

• Functions aren’t properly sized
  • When undersized high response time manifests
  • When oversized unnecessarily high cost accumulates
  • Solution: Rightsizing

• Cross-region data transfer cost isn’t considered
  • Place Lambda in every region (*stateless, no charge without using*)
  • Solution: Re-route traffic to a different region to minimize overall transfer cost

• A Kubernetes cluster is overloaded
  • Traditional solution: Scale out the cluster by adding more VMs
  • Alternative Solution: Route traffic out to a different cluster => Traffic Engineering
Smart Workloads with Edge Computing!

- Easier to place, size (*short-lived, smaller in size*)
  - No need for continuous placement of function *invocations*
- Traffic Engineering capable (*stateless*)
  - Multiple instances of the same function can be placed across clusters and regions
  - Client workload can be distributed over multiple function instances
- Edge Computing
  - Moving intensive workloads from the Cloud out to the Edge of the network.
Demo: Route Control in Edge Computing

Istio
Pushing Policy to route the traffic

Kong
API Gateway to forward to the correct endpoint

Push routes

Knative
Cloud

AWS Lambda

Turbonomic Decision Making Engine to make real-time routing decision based on application performance

Edge 1
The Serverless Platform

Public Cloud

AWS Lambda

Microsoft Azure

Google Cloud Functions

Kubernetes-based

OpenWhisk

OpenFaaS

riff

fn

Kubeless

Virtual Kubelet
Glue Layer
Service Routing Layer

HTTP/1.1, HTTP/2, gRPC or TCP -- with or without mTLS

Policy checks, telemetry

Config data to proxies

TLS certs to proxies

Control Plane API
Decision Engine

- AppDynamics
- Cisco HyperFlex
- Cisco UCS Manager
- Cisco UCS Director
- Cisco Tetration
- Cisco ACI
- Cisco CloudCenter
- ...
Abstraction: The Supply Chain Market

1. Everything in the data center is abstracted into a supply chain market.

2. Services entities shop for the best overall price for every commodity (resource) they need to perform.

   - **Service Entity**: App, Container, VM, Physical Machine, Fabric, etc.

   - **Commodity**: CPU, Memory, Flow, IO, Storage, IOPS, etc.

3. Within 1 hr. you see these relationships and metrics in Turbonomic.
Analysis: Economic Supply, Demand, and Price

- Utilization (demand/supply) determines price.
- Workloads/service entities make scaling, placement, and capacity decisions based on *all* the resources they need.
Automation: Real-time Action

Continuous Optimization
Real-time actions drive continuous health:
- Placement
- Sizing
- Provisioning

Capacity Management
Quickly & accurately model what-if scenarios:
- Workload growth
- Add/remove hardware
- Cloud costs
KubeCon | CloudNativeCon
-------|-----------------------
China 2018

Demo Time
Key Takeaways

• Deploying in cloud is complicated, even serverless!
• Workload needs to be Smart!
  • Resizable
  • Scalable
  • Routable
• Take advantage of Edge Computing!
  • High bandwidth + Low latency
• Decision-making is critical!
  • Performance
  • Cost Efficiency
  • Compliance
Thank you!!
谢谢😊