KubeCon | CloudNativeCon

China 2018

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

Enlin Xu, Director of Advanced Engineering, Turbonomic

# Agenda



- Who are we
- Challenges to deploy and operate applications in Cloud
- Smart Workloads in Edge Computing
- Demo
- Key Takeaways

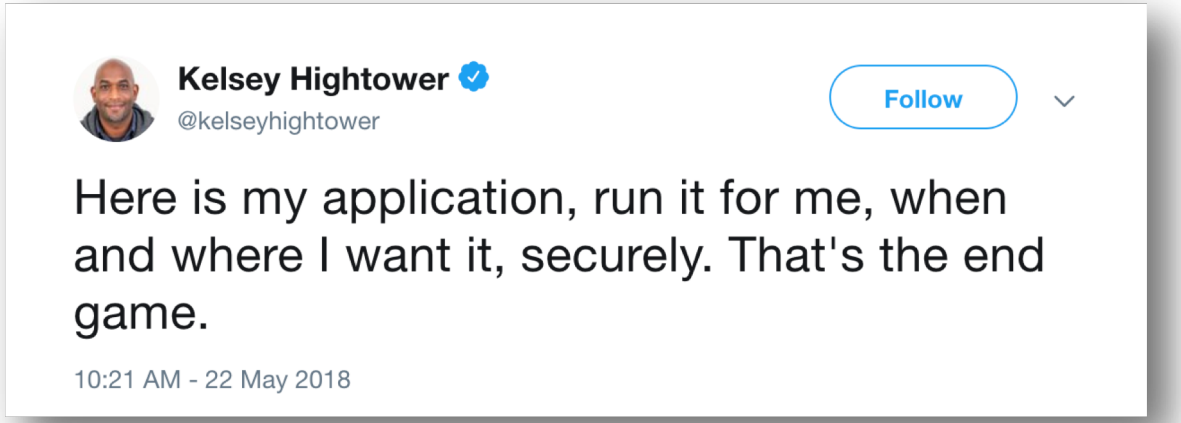
# Who are we?

KubeCon | CloudNativeCon China 2018

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

turbonomic

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




## We're Hiring!

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

Inc. BEST WORKPLACES 2018

IDC

*IDC Innovator: Multicloud Management, 2017*

Inc. 5000 AMERICA'S FASTEST-GROWING PRIVATE COMPANIES

Forbes THE 2018 Cloud100 THE WORLD'S BEST CLOUD COMPANIES

# 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

KubeCon | CloudNativeCon

China 2018

# **Serverless** looks simple enough! Or is it?

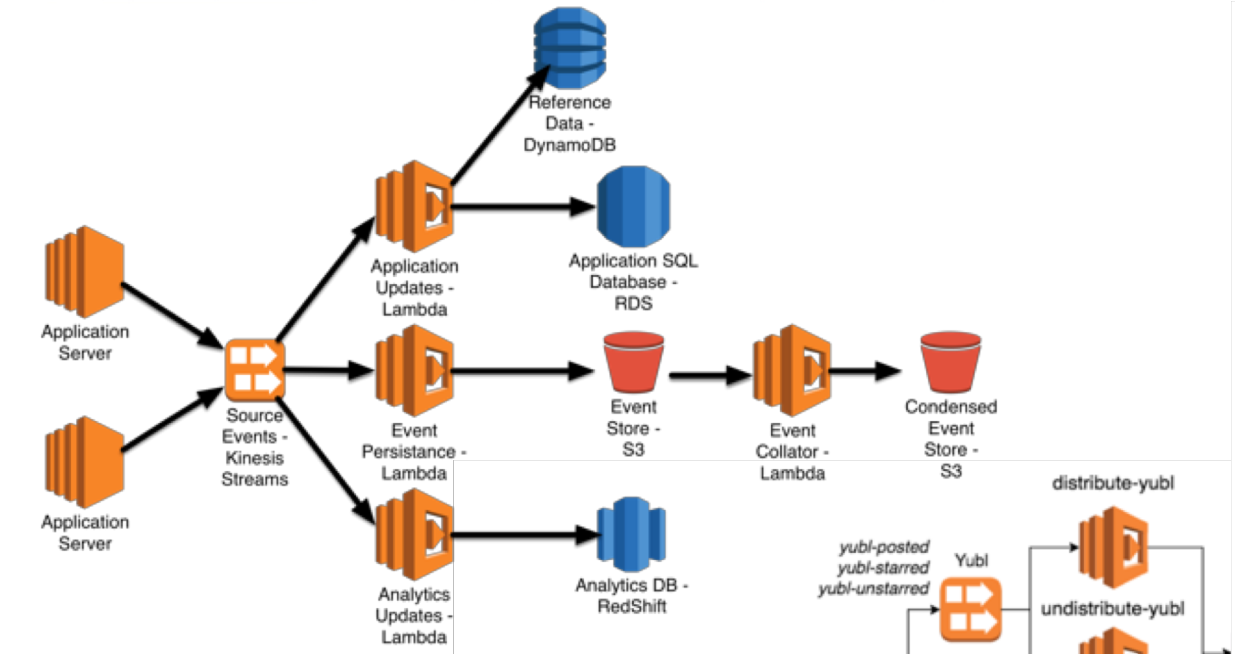
# Serverless Use Cases

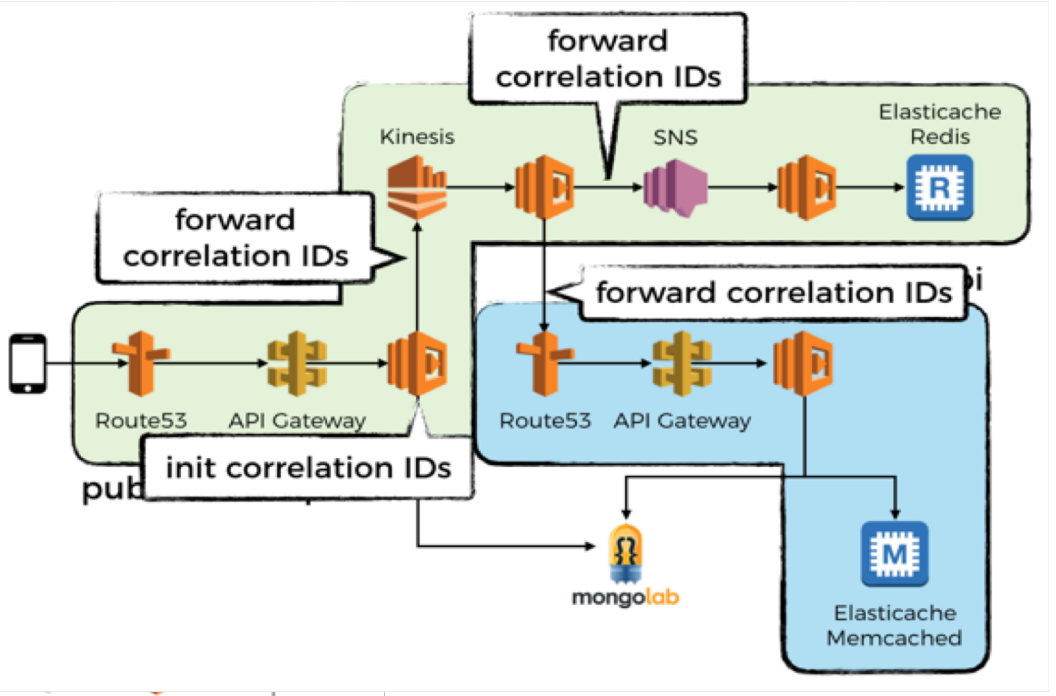


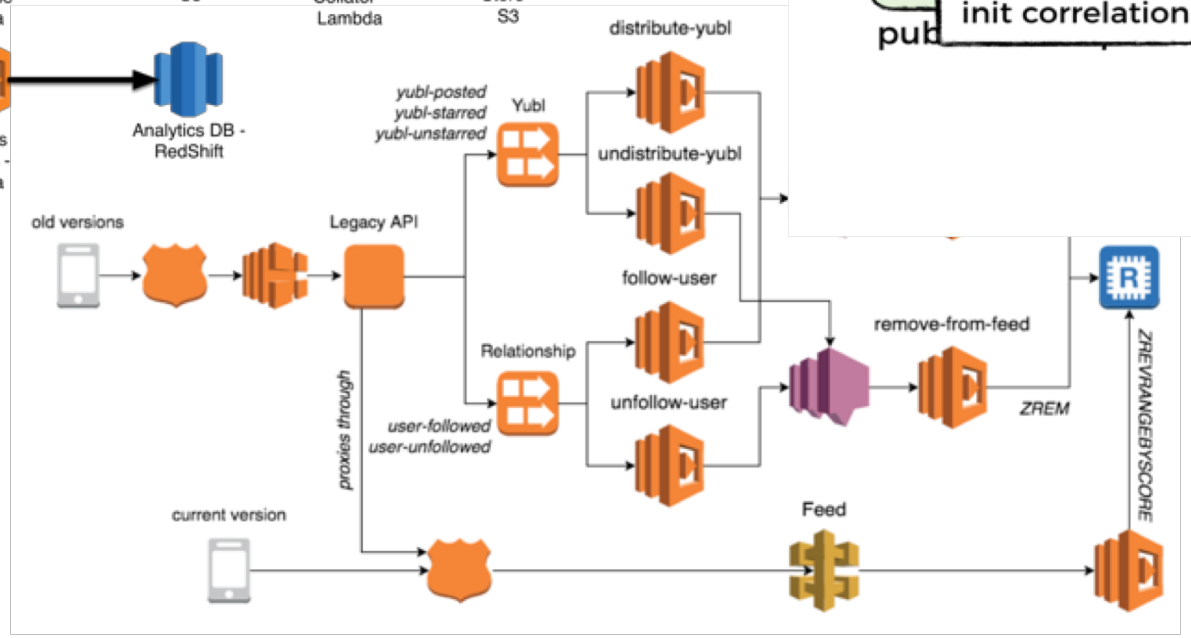
- An image upload to S3 triggers a Lambda to compute a thumbnail to post on Twitter
- The same upload triggers another Lambda to analyze the image and create metadata such as place, persons, etc.
- A code commit or PR merge triggers build jobs
- An IoT sensor on a garbage can detects its fullness and 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

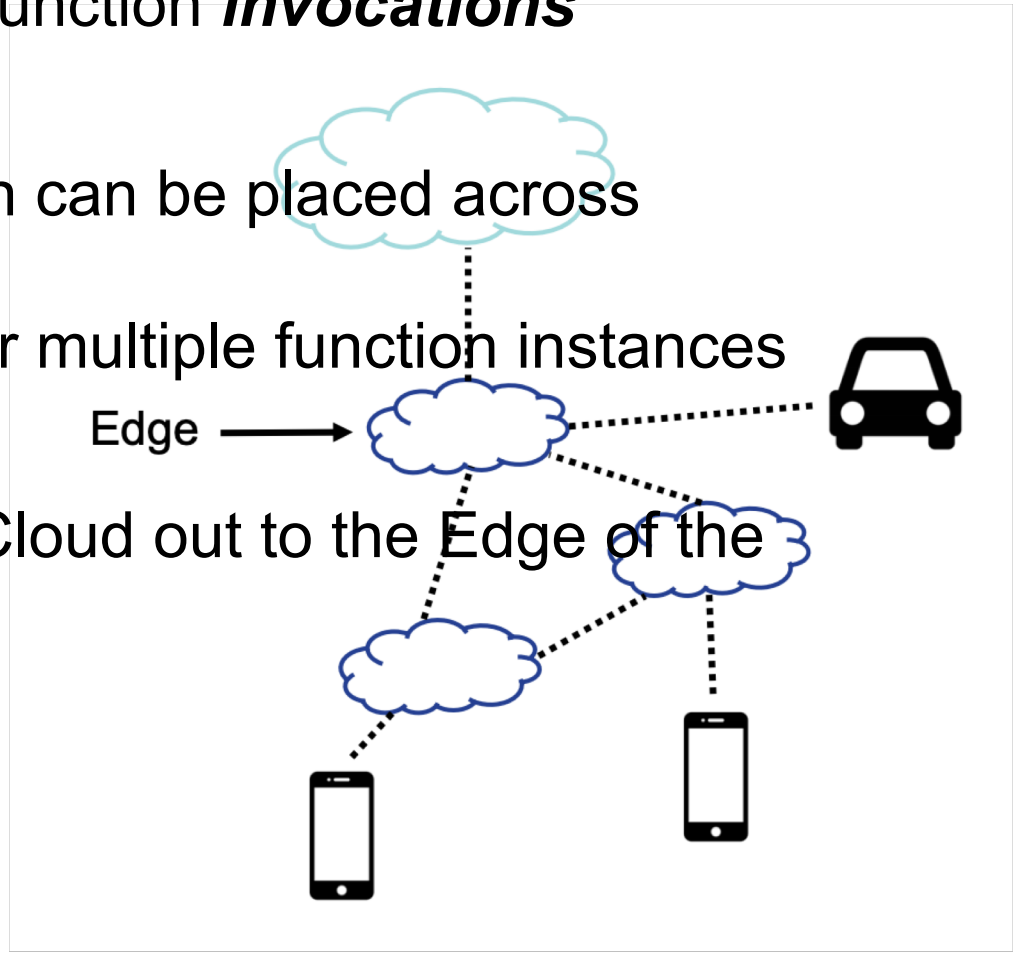


- Functions aren't properly sized
  - When undersized high response time manifests
  - When oversized unnecessarily high cost accumulates
  - **Solution: Rightsizing**
- Cross-region data transfer cost isn't considered
  - Place Lambda in every region (*stateless, no charge without using*)
  - **Solution: 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!

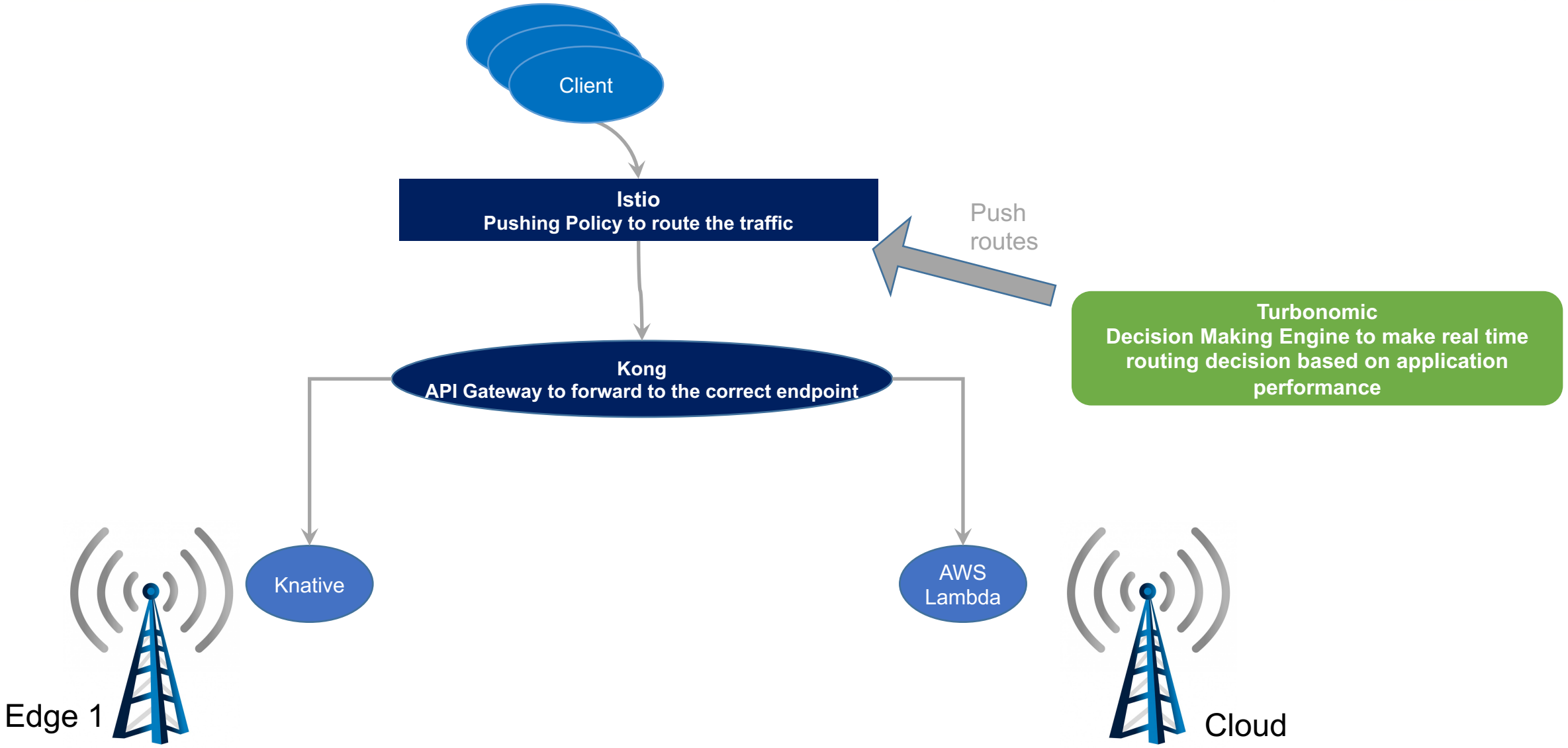


- Easier to place, size (*short-lived, smaller in size*)
  - No need for continuous placement of function ***invocations***
- Traffic Engineering capable (*stateless*)
  - Multiple instances of the same function can be placed across clusters and regions
  - Client workload can be distributed over multiple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



# Demo: Route Control in Edge Computing



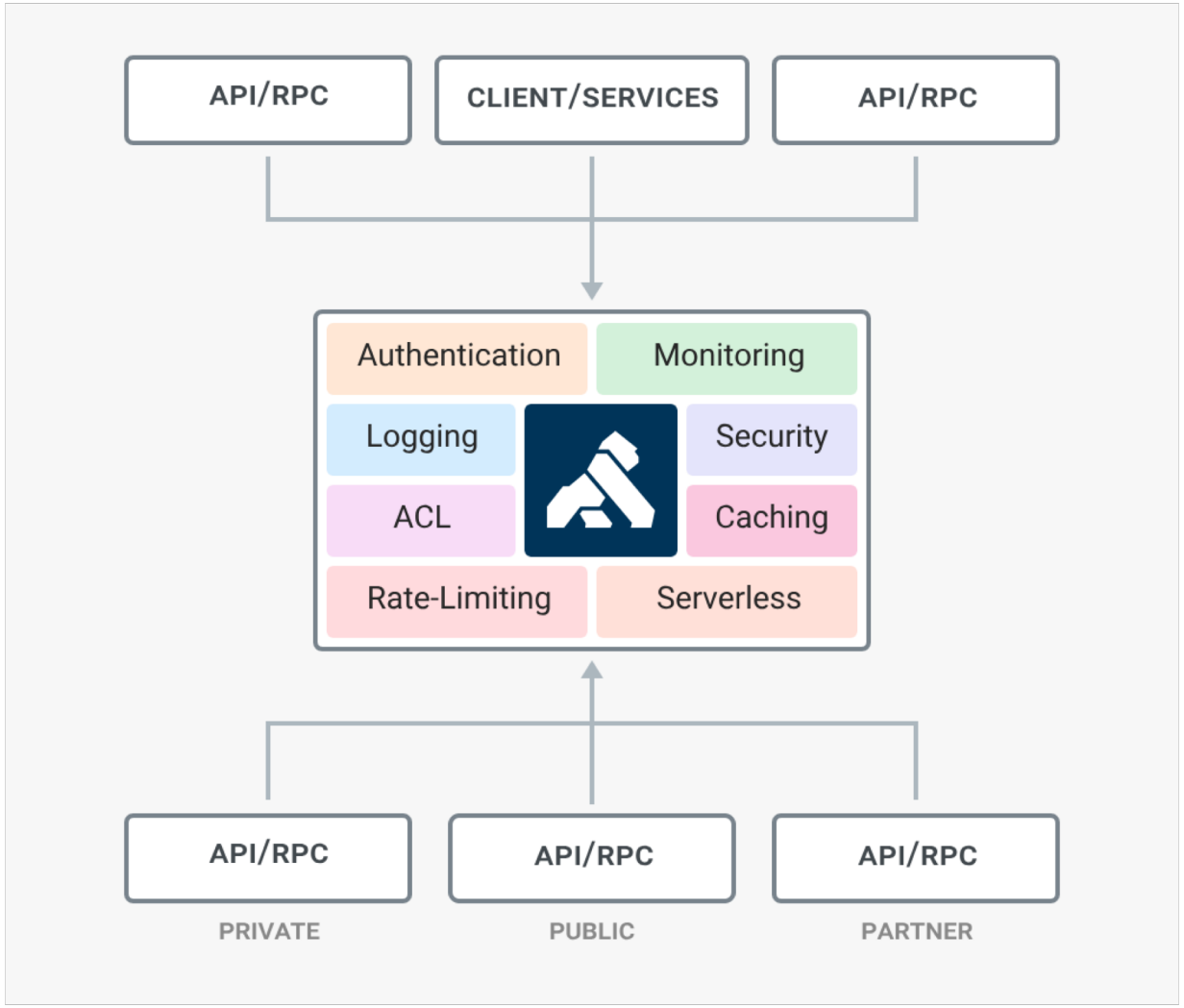


# The Serverless Platform



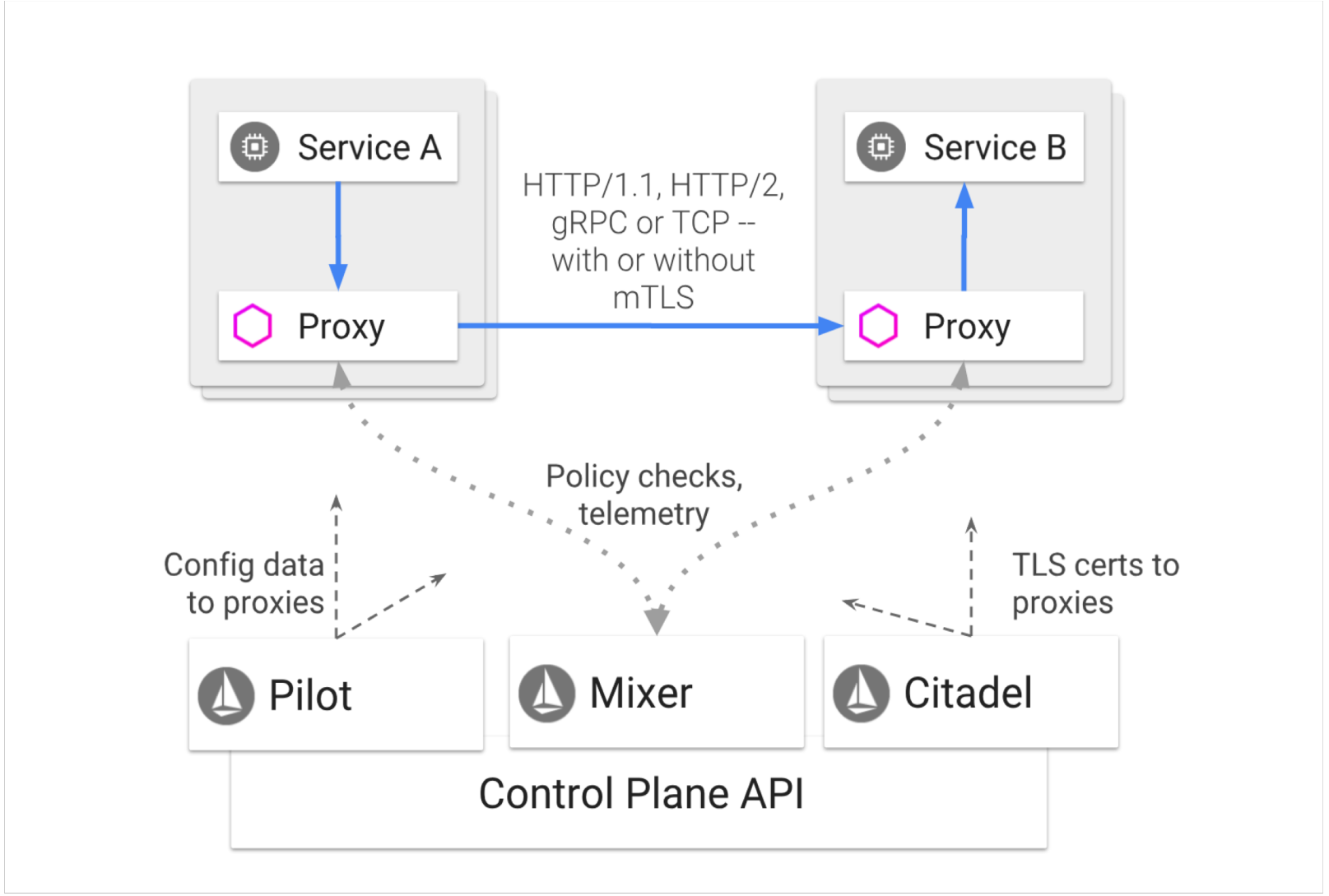

# Glue Layer




API/RPC
CLIENT/SERVICES
API/RPC
Authentication
Monitoring
Logging
Security
ACL
Caching
Rate-Limiting
Serverless
API/RPC
API/RPC
API/RPC
PRIVATE
PUBLIC
PARTNER

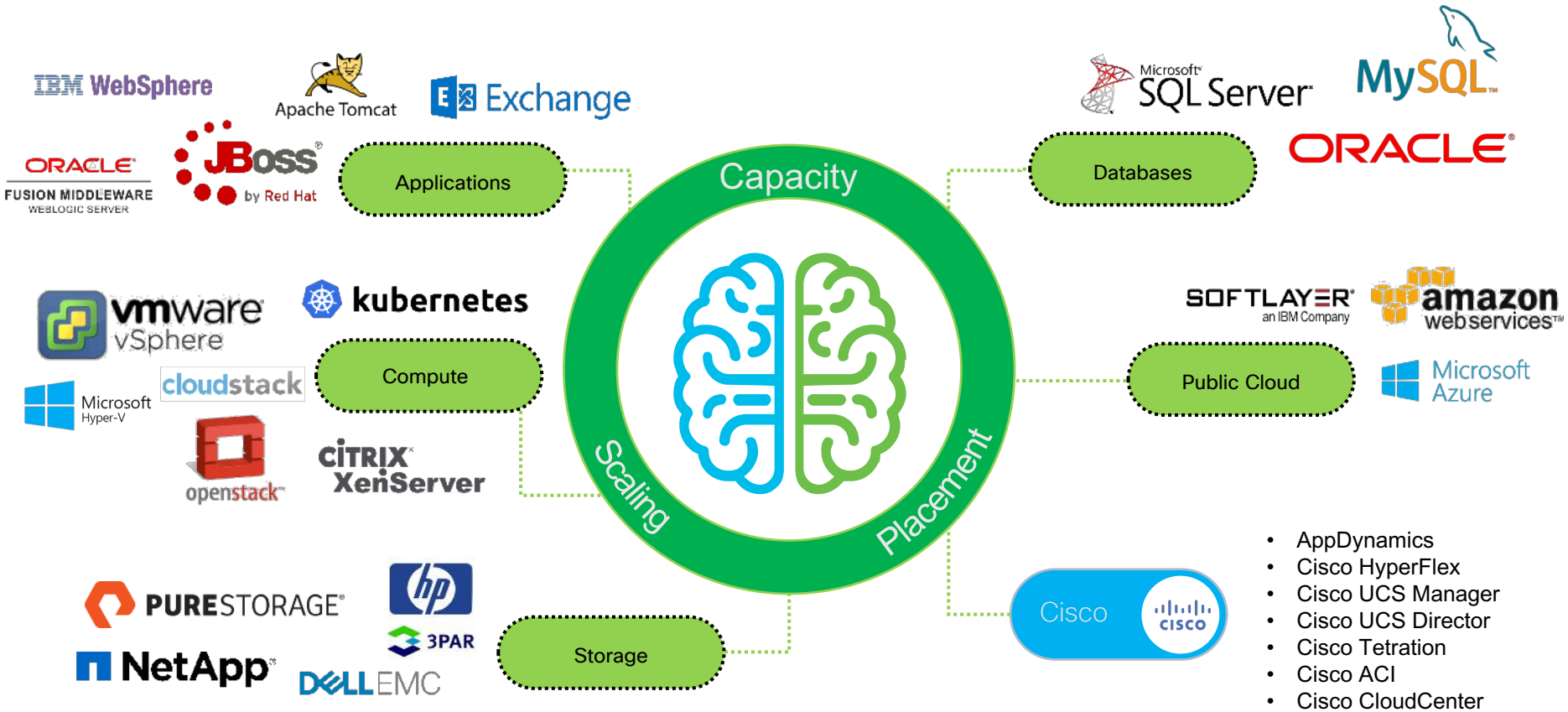
# Service Routing Layer





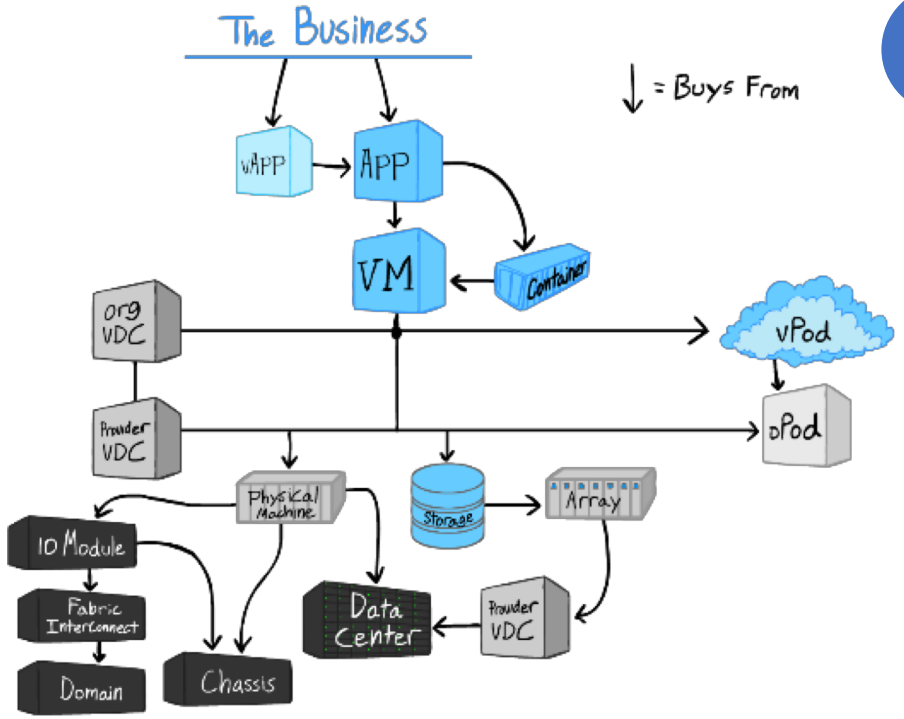
# Decision Engine





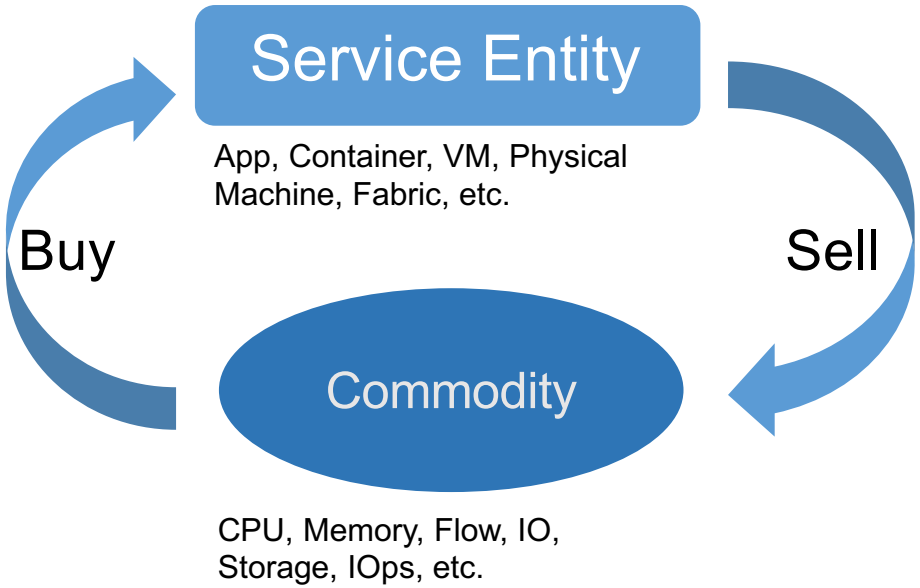

- AppDynamics
- Cisco HyperFlex
- Cisco UCS Manager
- Cisco UCS Director
- Cisco Tetration
- Cisco ACI
- Cisco CloudCenter
- …

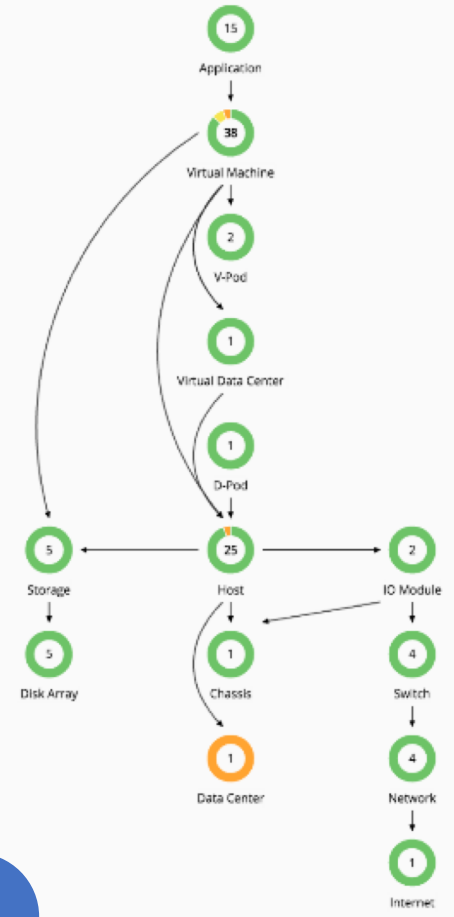
# **Abstraction:** The Supply Chain Market




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.

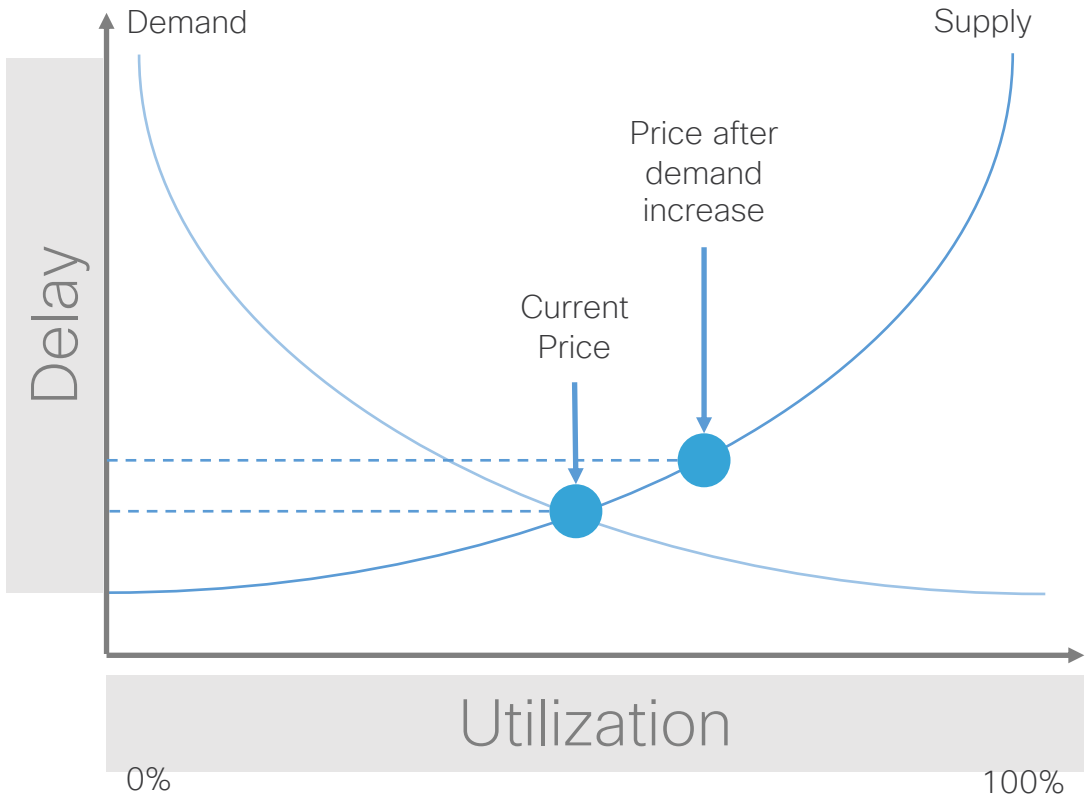



3 Within 1 hr. you see these relationships and metrics in Turbonomic

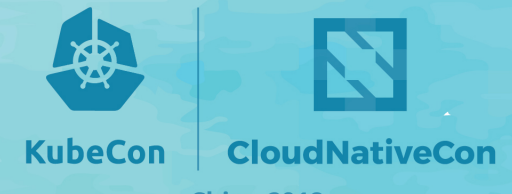
# **Analysis:** Economic Supply, Demand, and Price






- 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

KubeCon | CloudNativeCon

China 2018

Thank you!!

谢谢 ☺