AWS re:Invent

SRV409

A Serverless Journey: Under the Hood of AWS Lambda

Marc Brooker
Senior Principal Engineer
Serverless

Holly Mesrobian
Director of Engineering
AWS Lambda

AWS re:Invent

aws

SERVERLESS AT SCALE IS THE NEW NORM

THOMSON REUTERS

processes **4,000 requests** per second

mlbam

ingests, analyzes and stores **17+ petabytes of data** per season

FINRA

processes **half a trillion** validations of stock trades daily

Driver & Vehicle Licensing Agency

API traffic to register and license more than **47 million driver records** in Great Britain,

Zillow

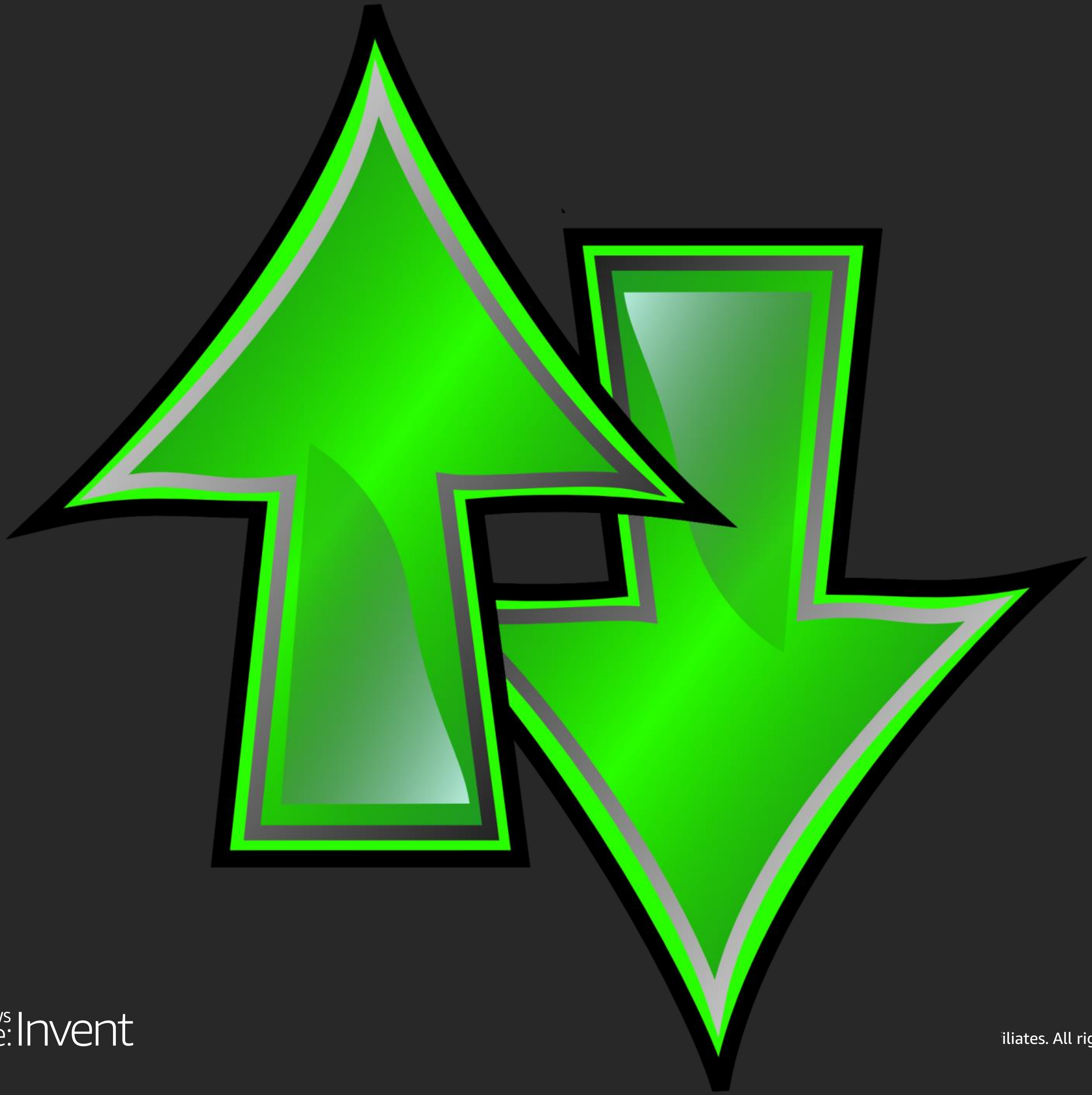
executes **16 million** requests a month

Localytics

processes **tens of billions** of data points monthly

Running Highly Available Large Scale Systems Is a Lot of Work

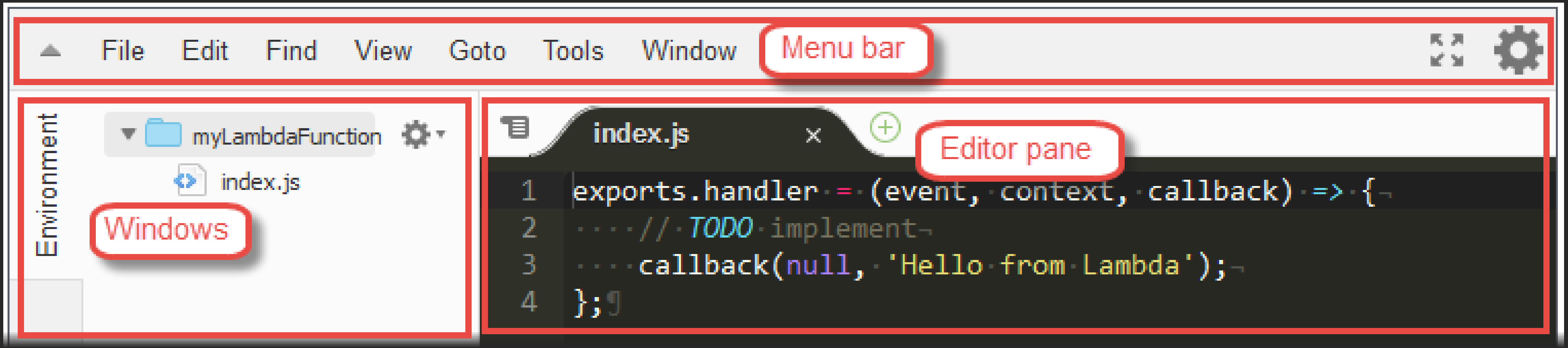
Load Balancing



Scaling Up and Down

Handling Failures

STOP
SITE IS UNDER CONSTRUCTION



```
exports.handler = (event, context, callback) => {
    // TODO implement
    callback(null, 'Hello from Lambda');
};
```

Lambda Handles

Load Balancing

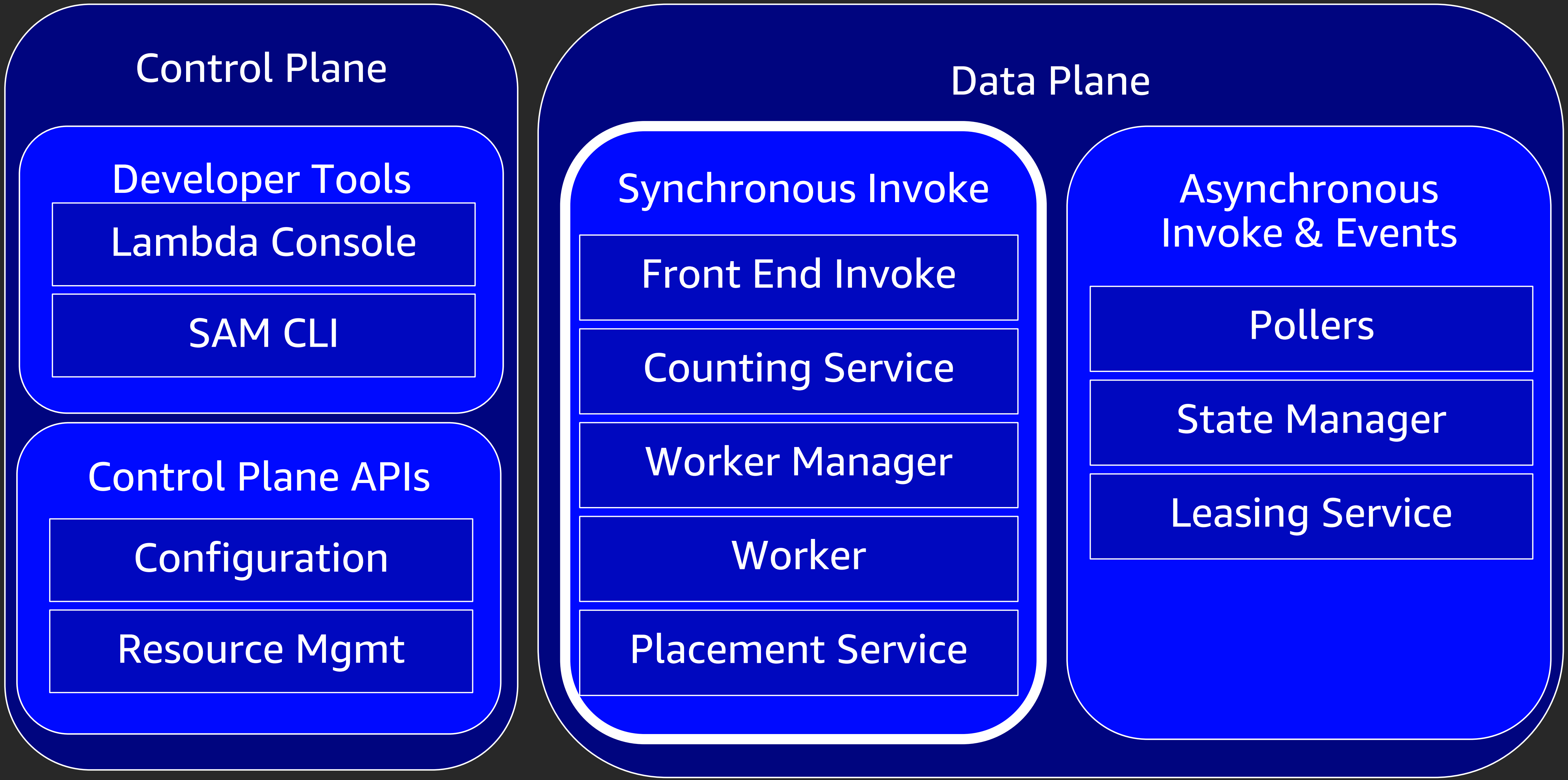
Auto Scaling

Handling Failures

Security Isolation

Managing Utilization

(and many other things) for you


Control Plane
Developer Tools
Lambda Console
SAM CLI
Control Plane APIs
Configuration
Resource Mgmt
Data Plane
Synchronous Invoke
Front End Invoke
Counting Service
Worker Manager
Worker
Placement Service
Asynchronous Invoke & Events
Pollers
State Manager
Leasing Service

Front End Invoke

Orchestrate both synchronous and asynchronous Invokes

Counting Service

Provides a region wide view of customer concurrency to help enforce set limits

Worker Manager

Tracks container idle and busy state and schedules incoming invoke requests to available containers

Worker

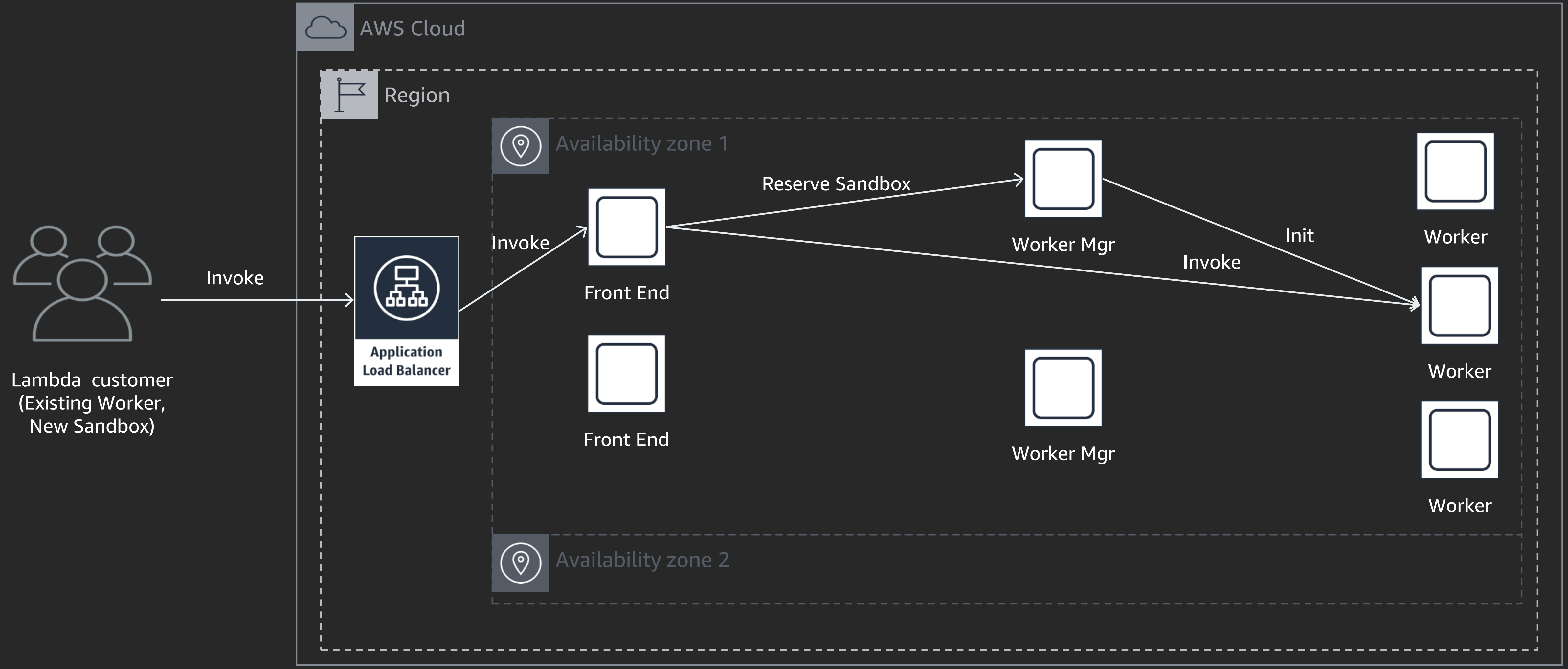
Provisions a secure environment for customer code execution

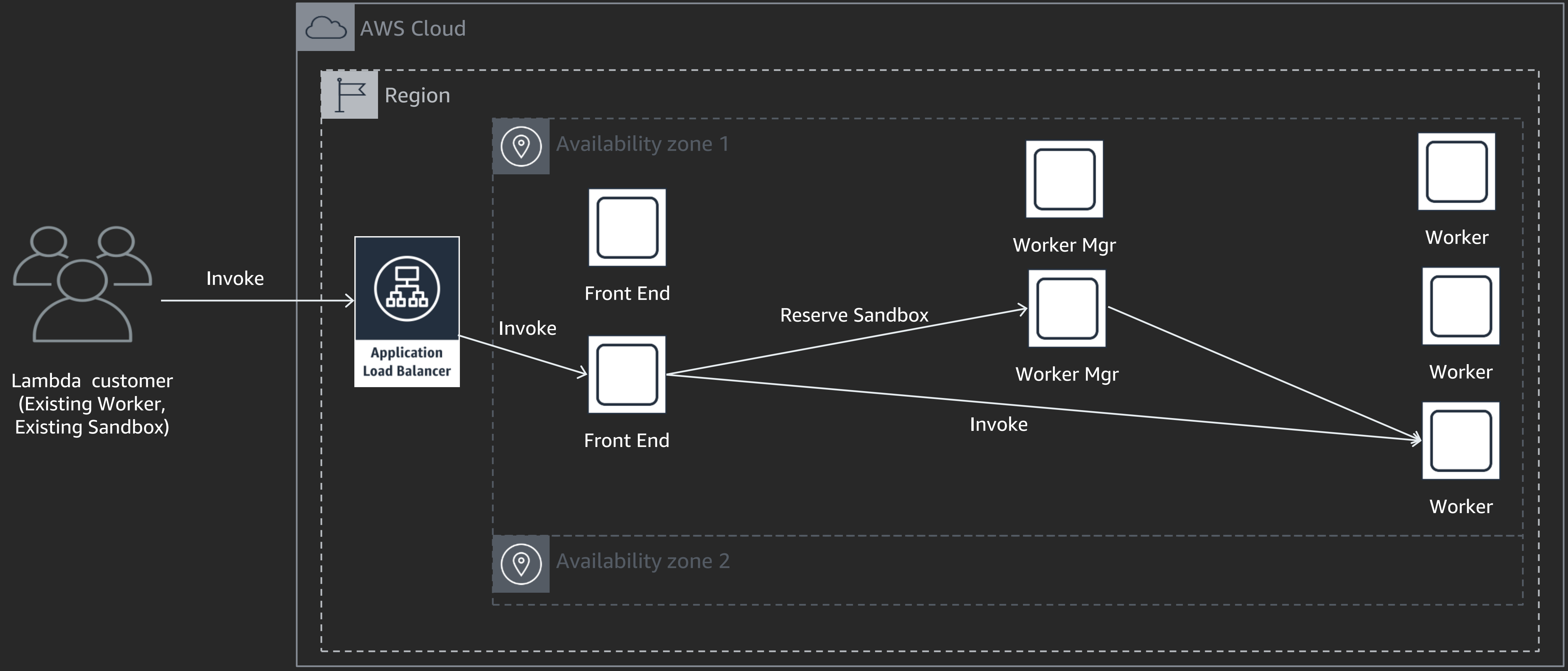
Placement Service

Places sandboxes on workers to maximize packing density without impacting customer experience or cold-path latency

Load Balancing

Routing function traffic across hosts distributed across Availability Zone


AWS Cloud
Region
Availability zone 1
Availability zone 2
Invoke
Lambda customer (Existing Worker, New Sandbox)
Application Load Balancer
Invoke
Front End
Front End
Reserve Sandbox
Worker Mgr
Worker Mgr
Init
Invoke
Worker
Worker
Worker


AWS Cloud
Region
Availability zone 1
Availability zone 2
Lambda customer (Existing Worker, Existing Sandbox)
Invoke
Application Load Balancer
Invoke
Front End
Front End
Reserve Sandbox
Worker Mgr
Worker Mgr
Invoke
Worker
Worker
Worker

SERVERLESS CUSTOMERS

airbnb · mlbam · COMCAST · Experian · Coca-Cola · LexisNexis · New York Public Library · Benchling · Pentair · DBRS

instacart · Zillow · ALT/S · VOGUE · 23andMe · SCHOLASTIC · CITRIX · Mc Graw Hill · The Seattle Times · ADITYA BIRLA GROUP

VidRoll · Localytics · asurion · Expedia · Atlassian · FireEye · vevo · EDIZIONI CONDÉ NAST S.P.A.

BUSTLE · THOMSON REUTERS · ViewLift · SQUARE ENIX. · Nextdoor · AirAsia · ZAPPROVED · FannieMae. · THE HONEST CO. · Vertafore Unleash your potential

FUNNY OR DIE · Genome Institute of Singapore · MORNINGSTAR · Haier · iRobot · ISCS · La Mutuelle Générale · LA NACION · rachio · Discovery COMMUNICATIONS

TTI · FICO · twilio · FINRA · VERACODE · MADSACK MEDIENGRUPPE · BEACHBODY ON DEMAND · AUTODESK · WARBY PARKER

doppler labs · Cornelius · WHALEROCK INDUSTRIES · cimpress · Nishitetsu · Optimizely · ASTRO DIGITAL · SAISON INSURANCE · Serenova shine bright · nextel TÁ NAS SUAS MÃOS.

INTERCOM · COUNTY OF ALAMEDA CALIFORNIA · NATIONAL GEOGRAPHIC · AdRoll · DBS · edmunds · teralytic · HomeAway · Fender · Ringier

AWS re:Invent

aws

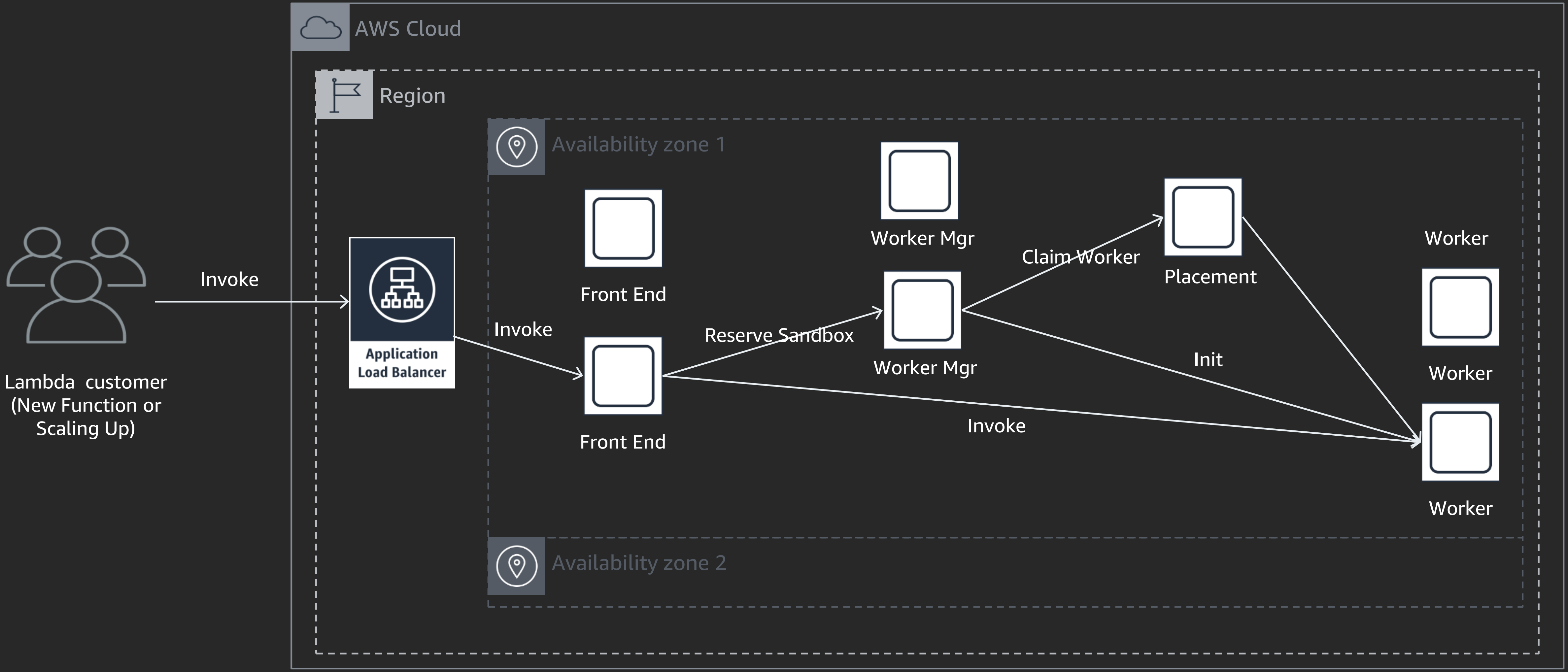
edmunds

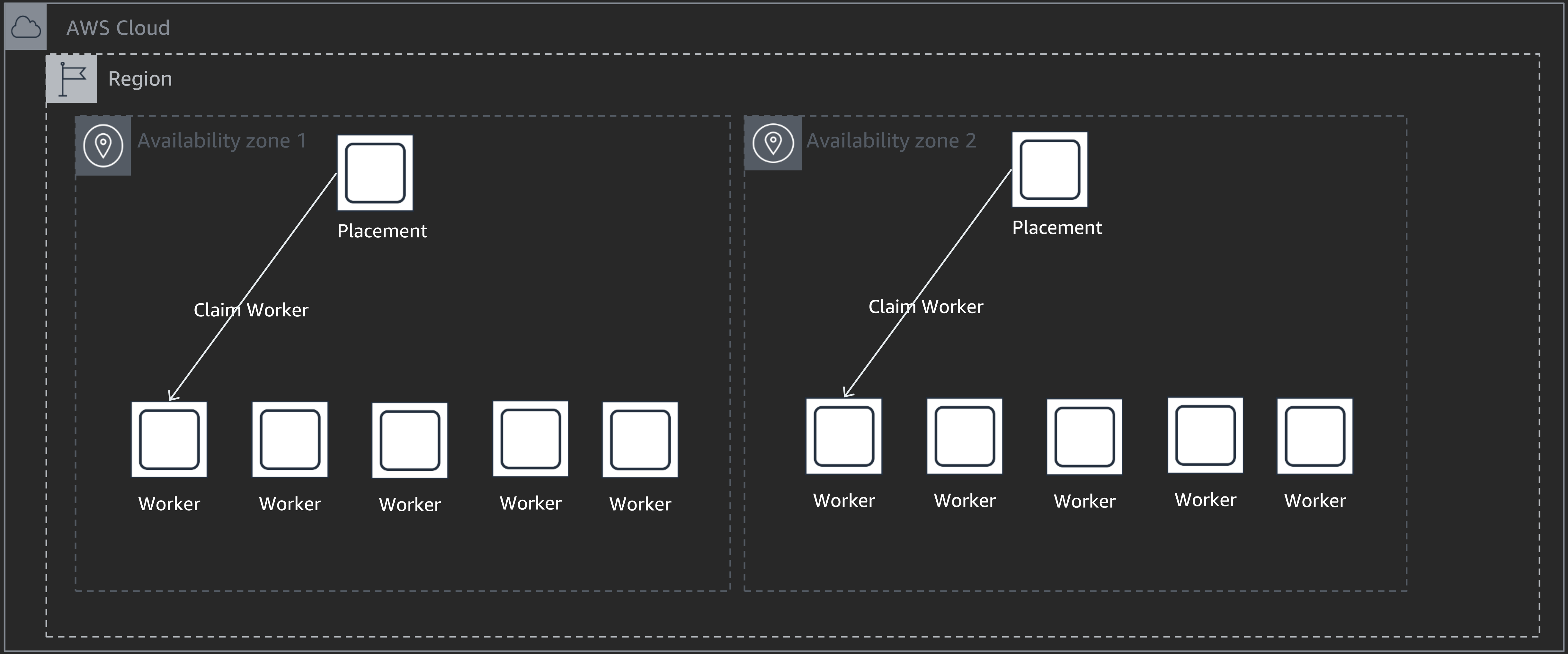
“What took us just a few days to build using a serverless solution based on AWS Lambda would have taken us six months to build from scratch. Our CTO and the rest of the project stakeholders were really happy with how much money and time we saved.”

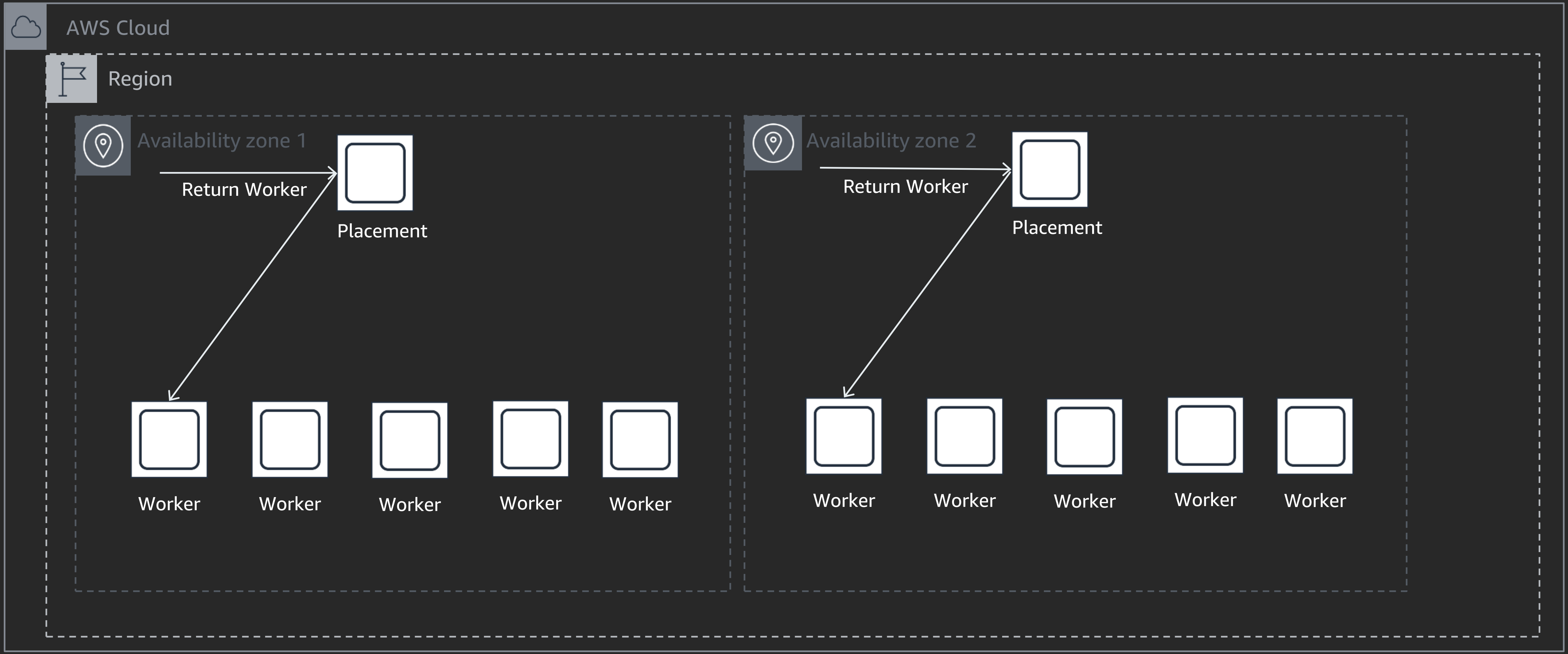
Nitin Mahajan
Executive Director for Service Engineering

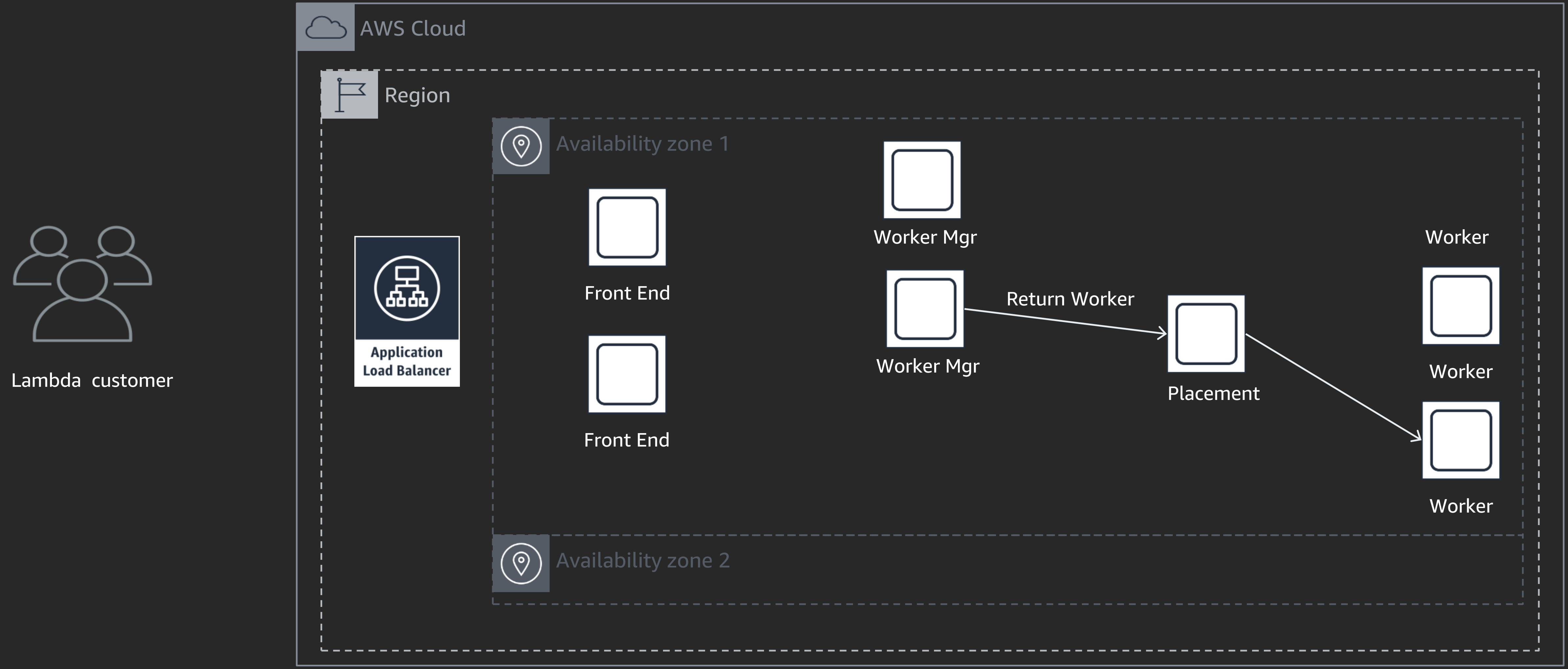
Auto Scaling

Provision function capacity when needed and releasing when not needed


AWS Cloud
Region
Availability zone 1
Availability zone 2
Lambda customer (New Function or Scaling Up)
Invoke
Application Load Balancer
Invoke
Front End
Front End
Reserve Sandbox
Worker Mgr
Worker Mgr
Claim Worker
Placement
Init
Invoke
Worker
Worker
Worker


AWS Cloud
Region
Availability zone 1
Placement
Claim Worker
Worker
Worker
Worker
Worker
Worker
Availability zone 2
Placement
Claim Worker
Worker
Worker
Worker
Worker
Worker


AWS Cloud
Region
Availability zone 1
Return Worker
Placement
Worker
Worker
Worker
Worker
Worker
Availability zone 2
Return Worker
Placement
Worker
Worker
Worker
Worker
Worker


Lambda customer
AWS Cloud
Region
Application Load Balancer
Availability zone 1
Front End
Front End
Worker Mgr
Worker Mgr
Return Worker
Placement
Worker
Worker
Worker
Availability zone 2

SERVERLESS CUSTOMERS

airbnb mlbam COMCAST Experian Coca-Cola LexisNexis New York Public Library Benchling Pentair DBRS

instacart Zillow ALT/S VOGUE 23andMe SCHOLASTIC CITRIX Mc Graw Hill The Seattle Times ADITYA BIRLA GROUP

VidRoll Localytics asurion Expedia Atlassian FireEye vevo EDIZIONI CONDÉ NAST S.P.A.

BUSTLE THOMSON REUTERS ViewLift SQUARE ENIX Nextdoor AirAsia ZAPPROVED FannieMae THE HONEST CO. Vertafore Unleash your potential

FUNNY OR DIE Genome Institute of Singapore MORNINGSTAR Haier iRobot ISCS La Mutuelle Générale LA NACION rachio Discovery COMMUNICATIONS

TTI FICO twilio FINRA VERACODE MADSACK MEDIENGRUPPE BEACHBODY ON DEMAND AUTODESK WARBY PARKER

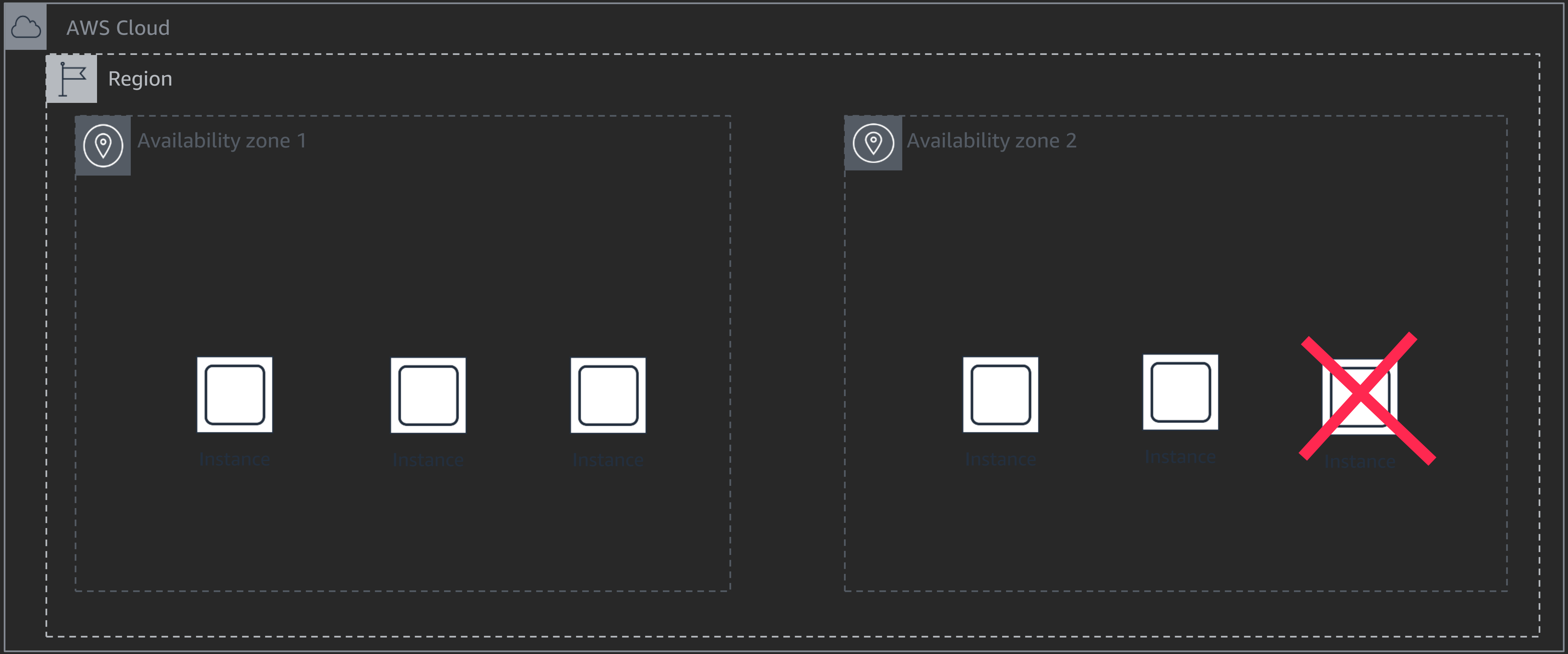
doppler labs Cornelius WHALEROCK INDUSTRIES cimpress Nishitetsu Optimizely ASTRO DIGITAL SAISON INSURANCE Serenova shine bright nextel TÁ NAS SUAS MÃOS.

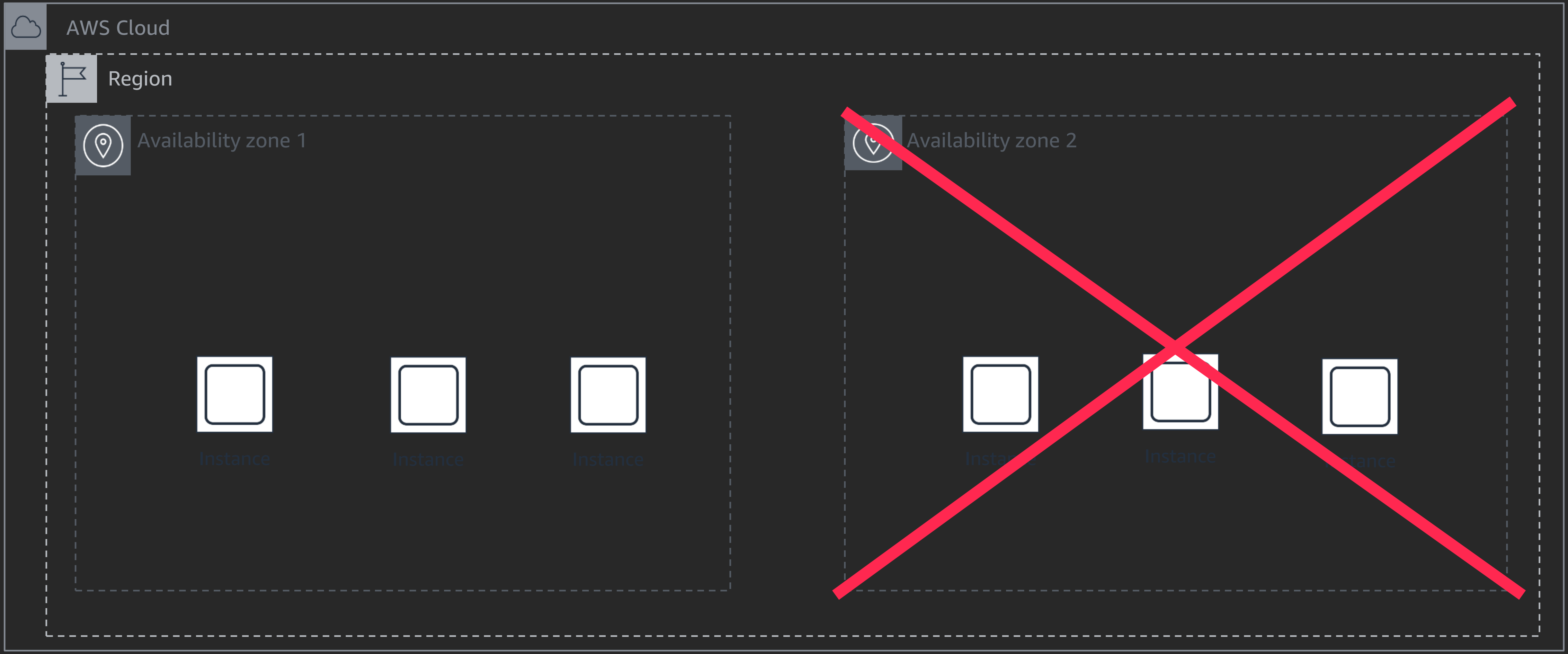
INTERCOM COUNTY OF ALAMEDA CALIFORNIA NATIONAL GEOGRAPHIC AdRoll DBS edmunds teralytic HomeAway Fender Ringier

Handling Failures

Handling Host and Availability Zone failure

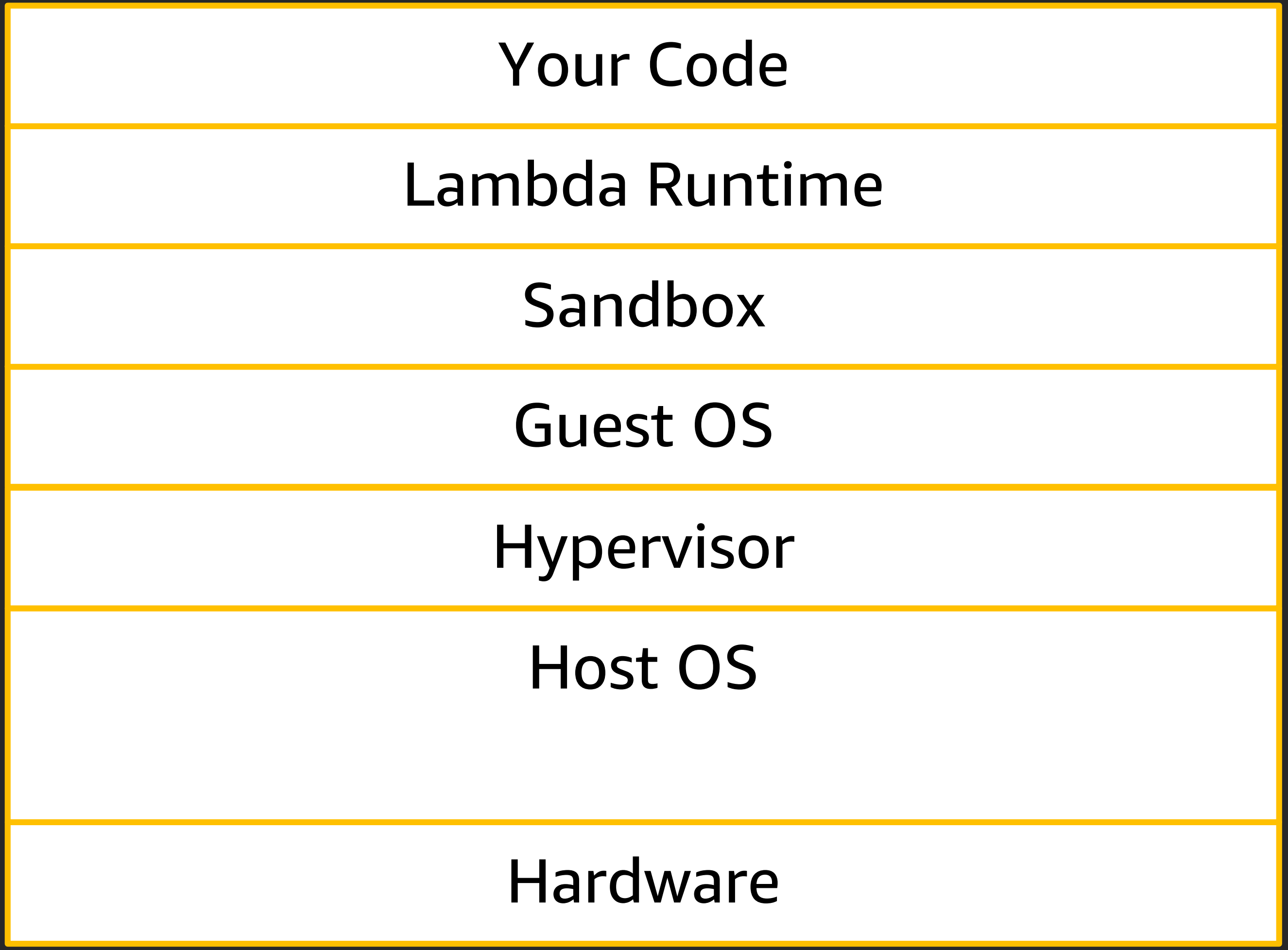
With Lambda:
Always have a healthy host

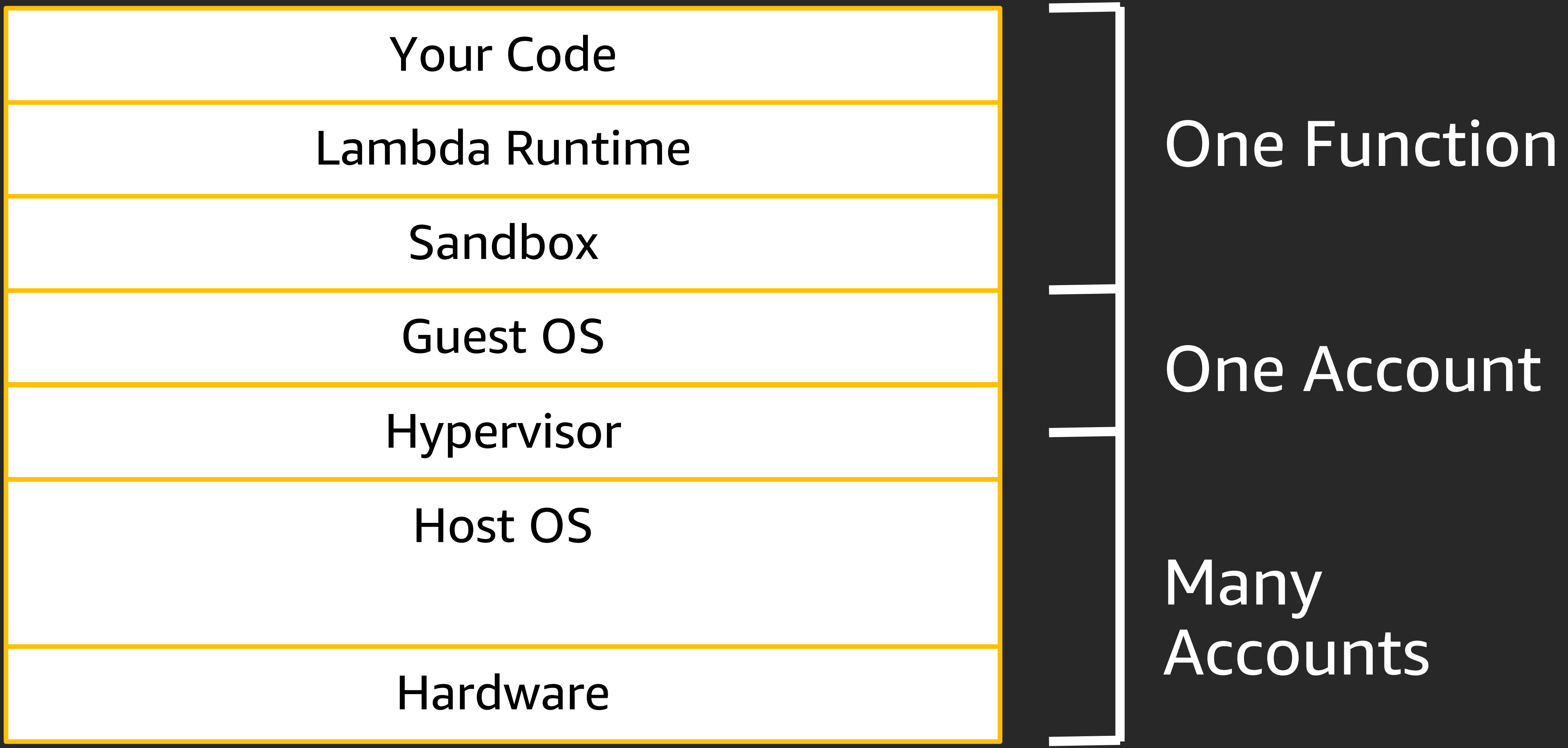

AWS Cloud
Region
Availability zone 1
Availability zone 2
Instance
Instance
Instance
Instance
Instance
Instance

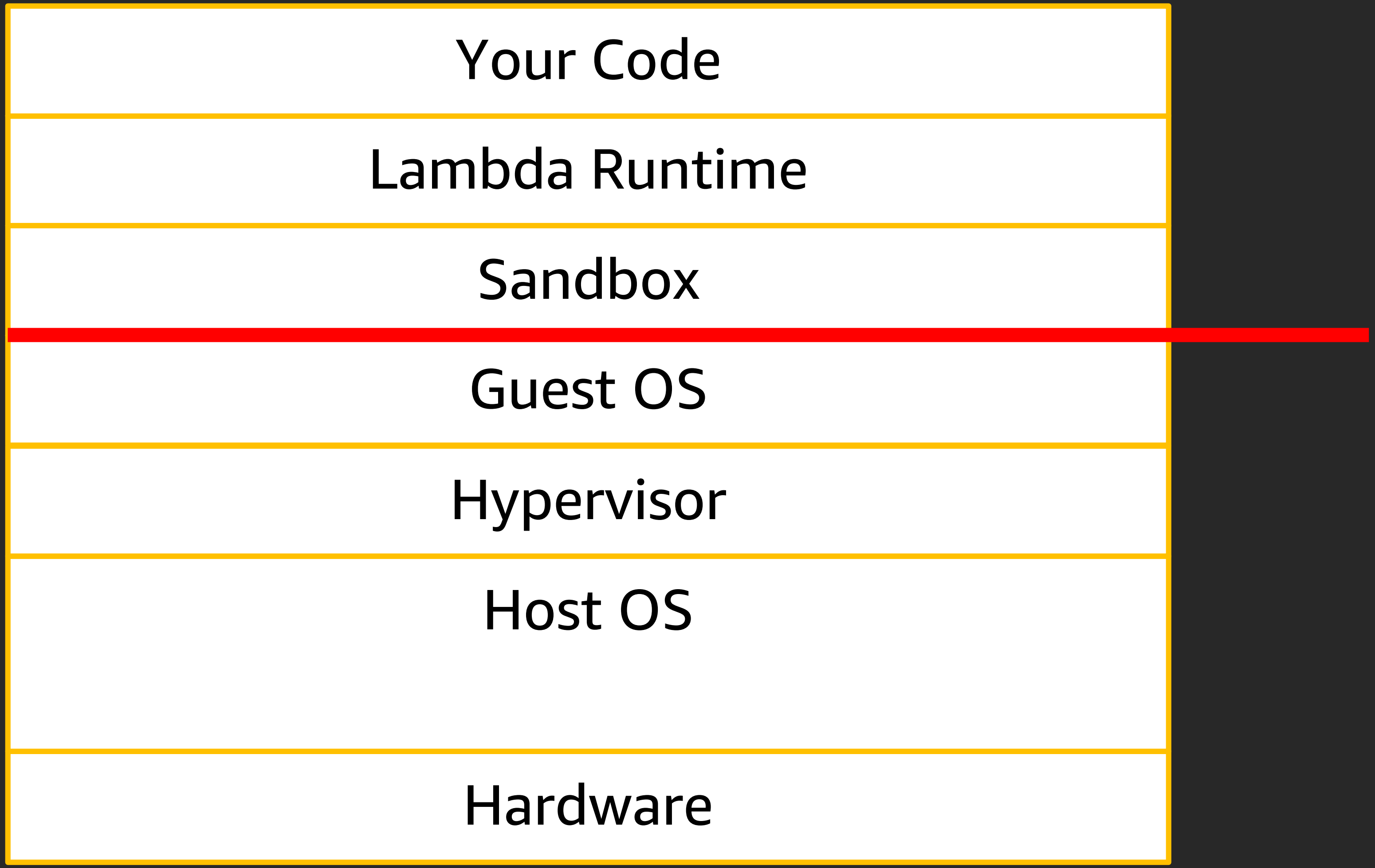

AWS Cloud
Region
Availability zone 1
Instance
Instance
Instance
Availability zone 2
Instance
Instance

Isolation

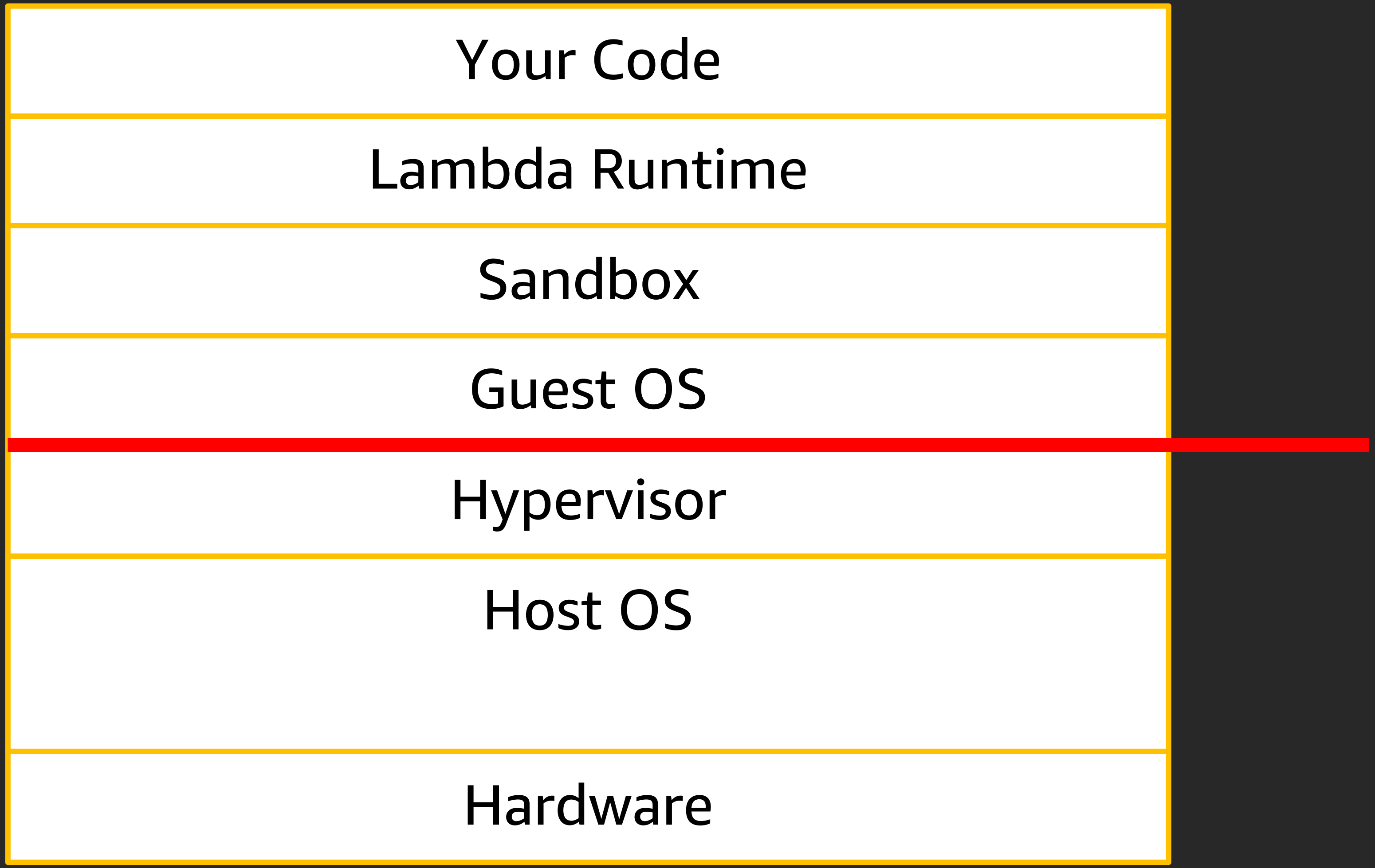
Keeping Workloads Safe and Separate


Your Code
Lambda Runtime
Sandbox
Guest OS
Hypervisor
Host OS
Hardware

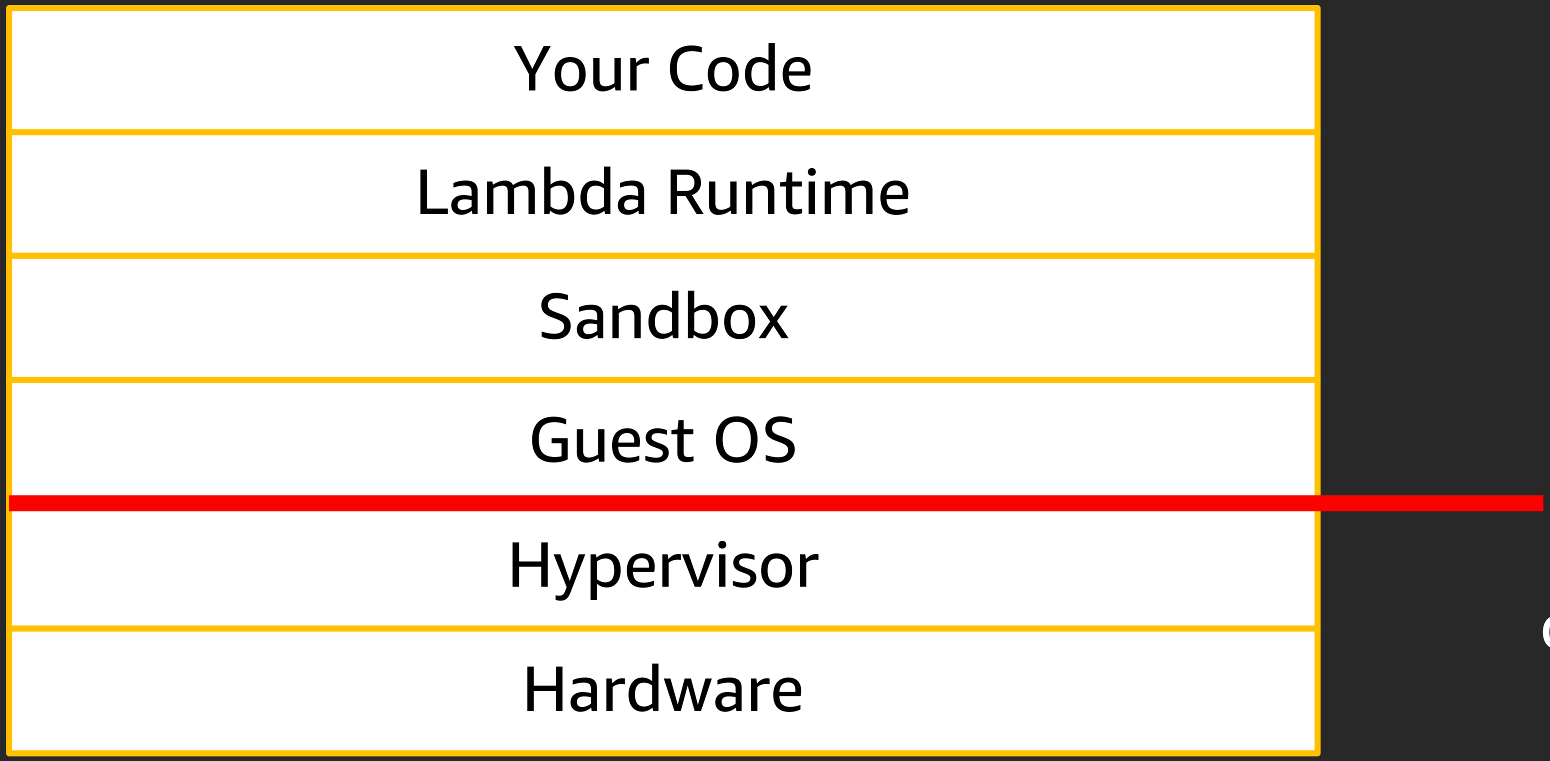

Your Code
Lambda Runtime
Sandbox
Guest OS
Hypervisor
Host OS
Hardware
One Function
One Account
Many Accounts



cgroups,
namespaces,
seccomp,
iptables,
& chroot

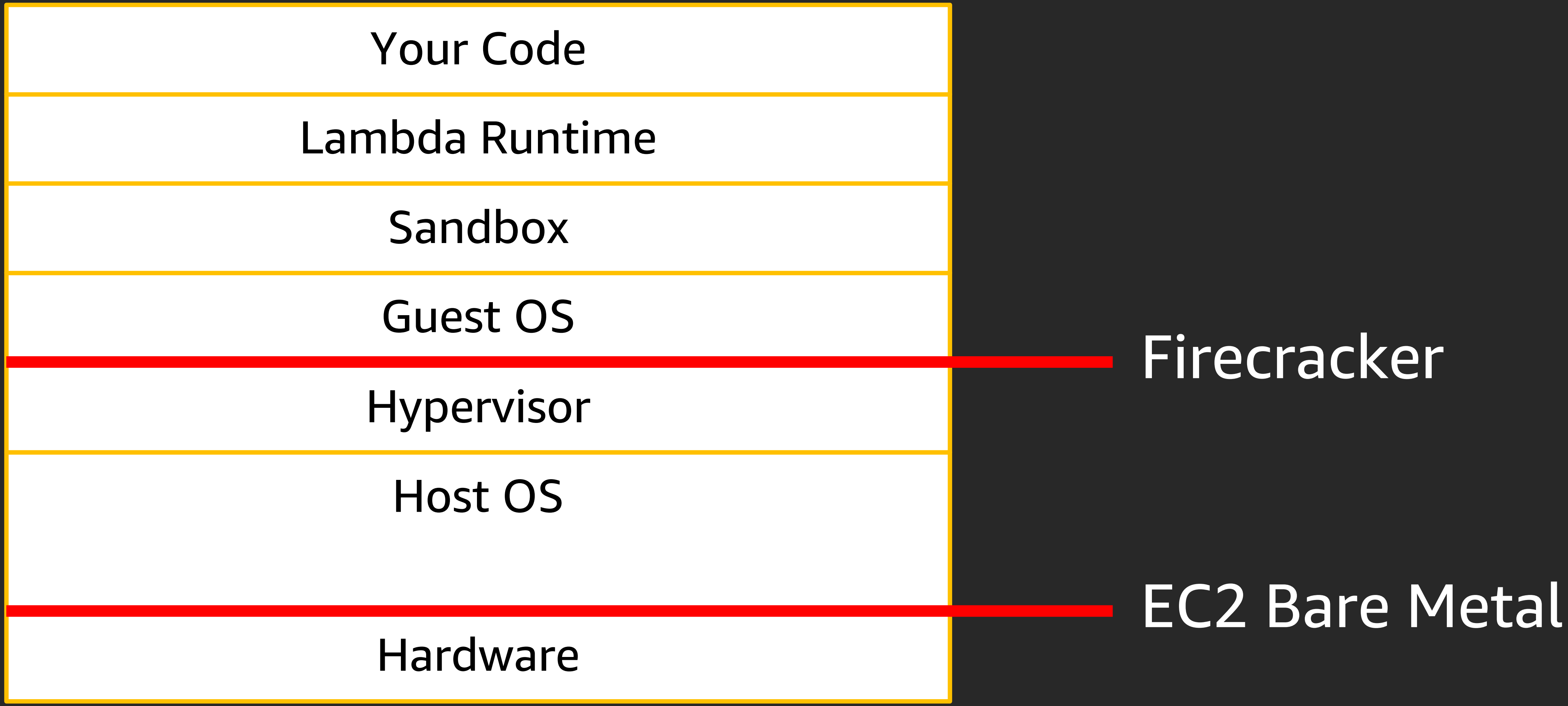


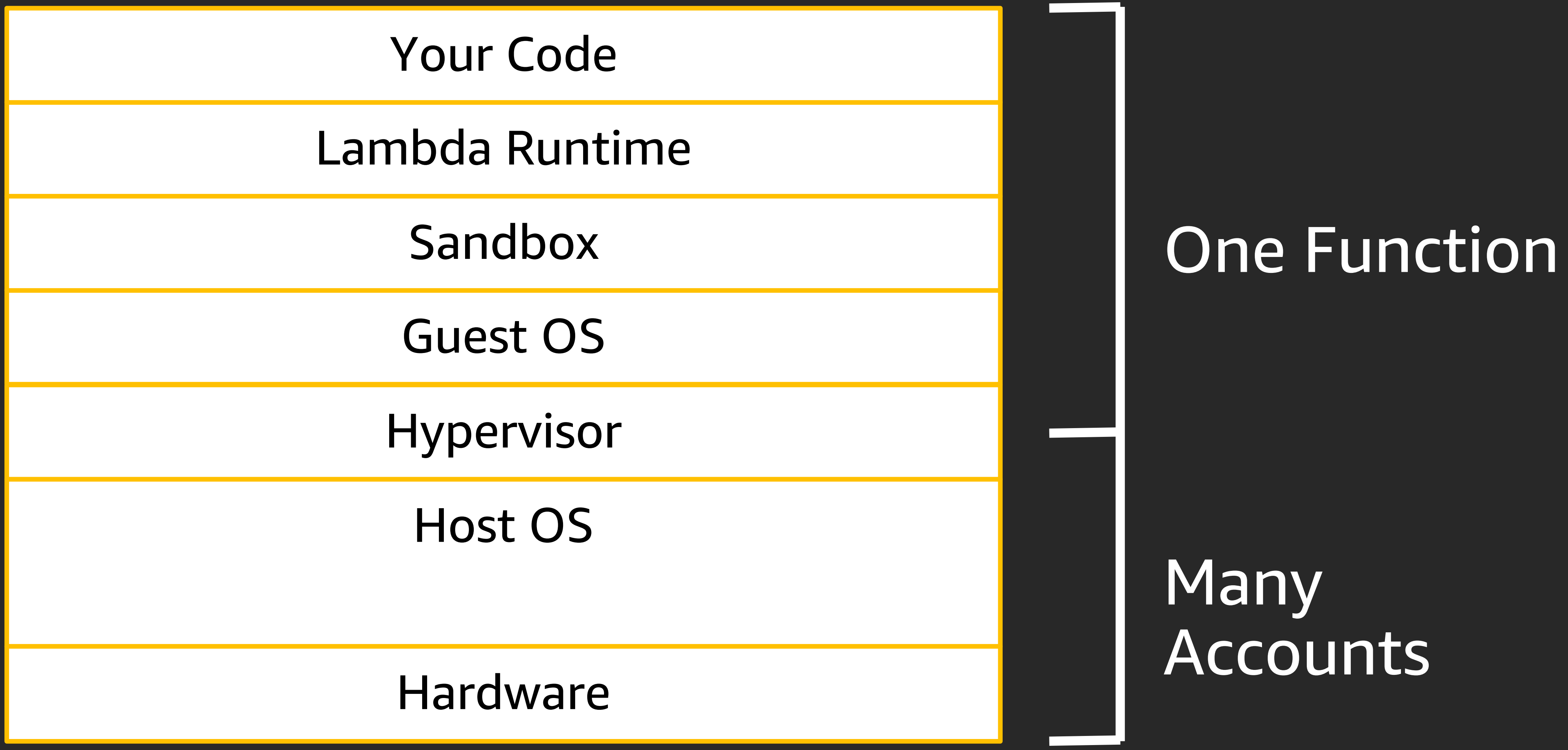
virtualization & device emulation



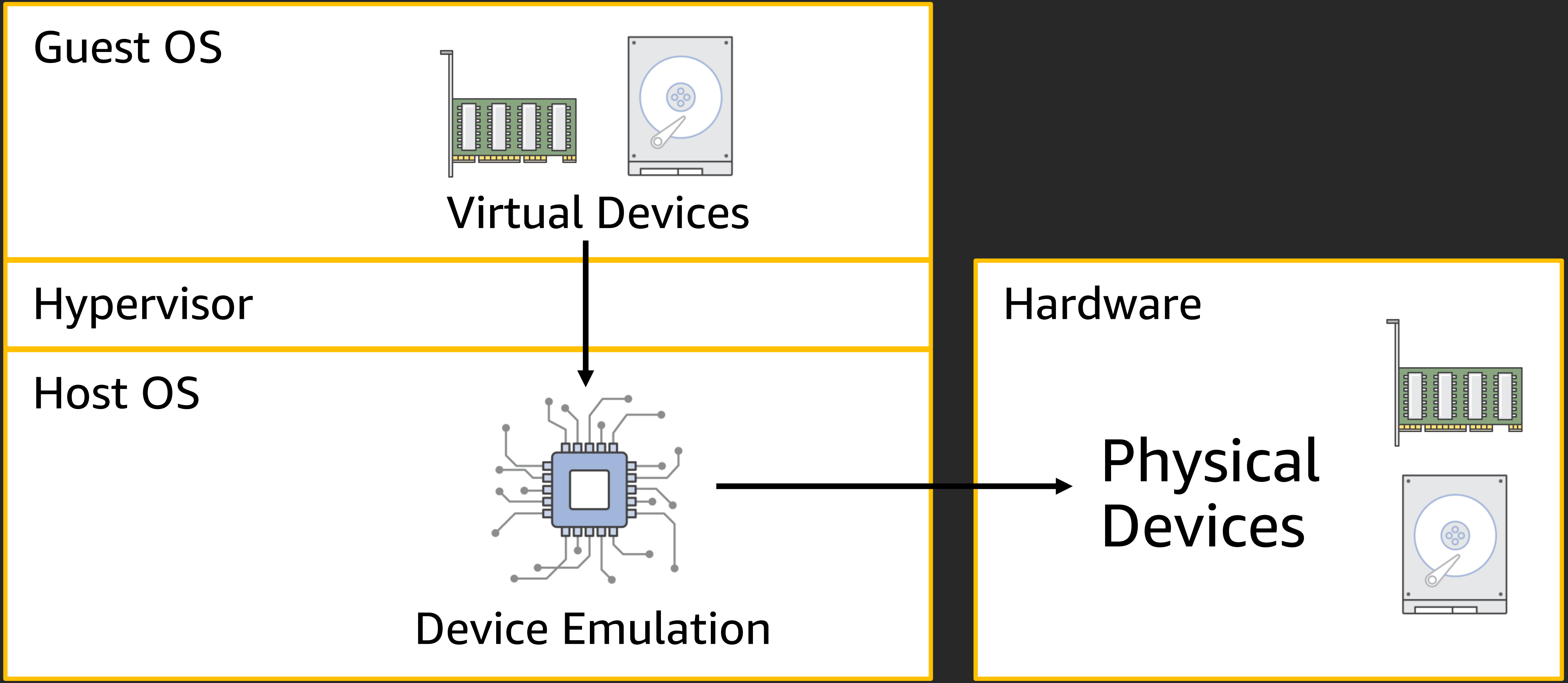
EC2 instances

on EC2 Nitro platform

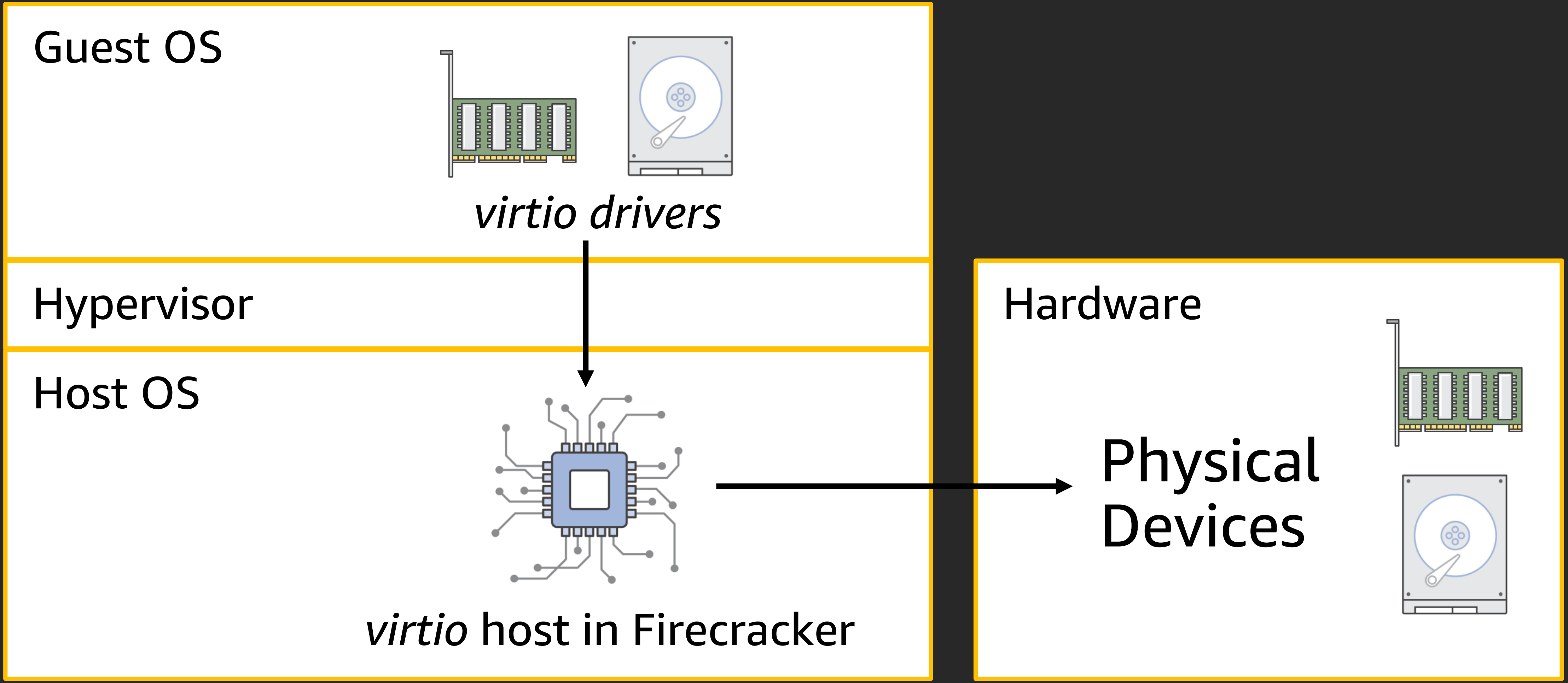

Your Code
Lambda Runtime
Sandbox
Guest OS
Hypervisor
Host OS
Hardware
Firecracker
EC2 Bare Metal


Your Code
Lambda Runtime
Sandbox
Guest OS
Hypervisor
Host OS
Hardware
One Function
Many Accounts

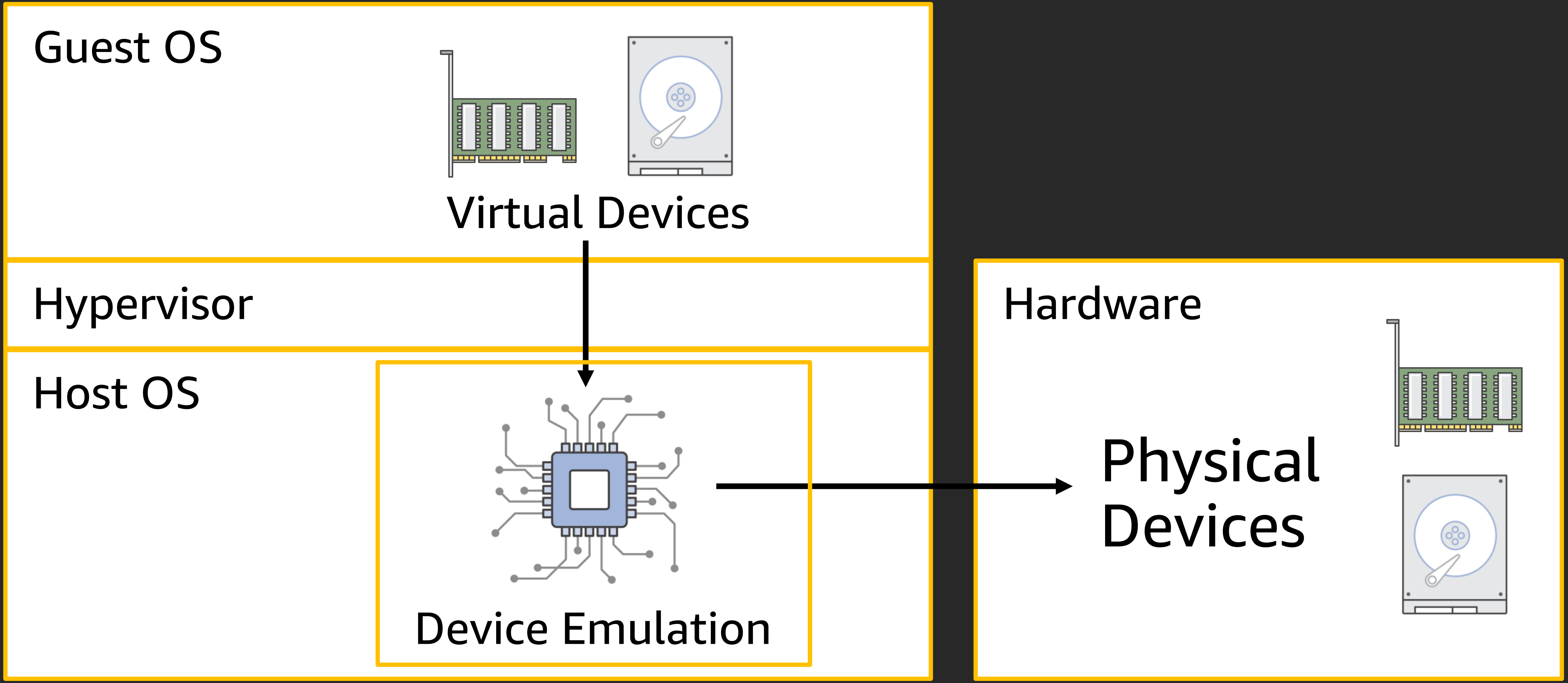
Innovating to Improve Isolation



Innovating to Improve Isolation



Innovating to Improve Isolation



Utilization

Keeping Servers Busy

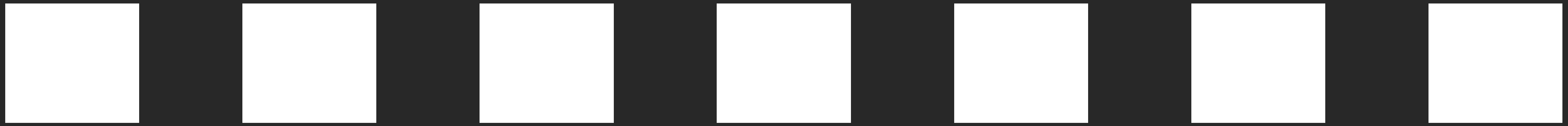
% of Resources Doing Useful Work

(vs. idle or waste)

With Lambda:
Pay only for useful work.

Inside Lambda: Optimize To Keep Servers Busy

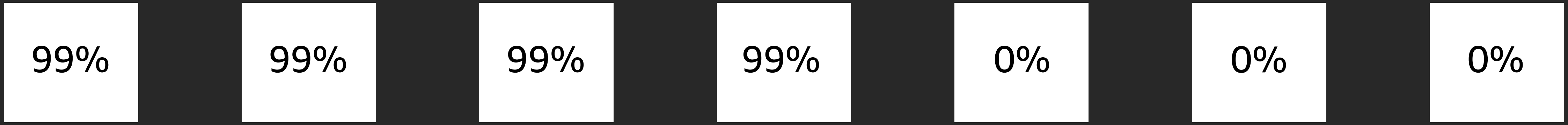
Available Sandboxes For a Function



Bad: Balance The Load

60% 60% 60% 60% 60% 60% 60%

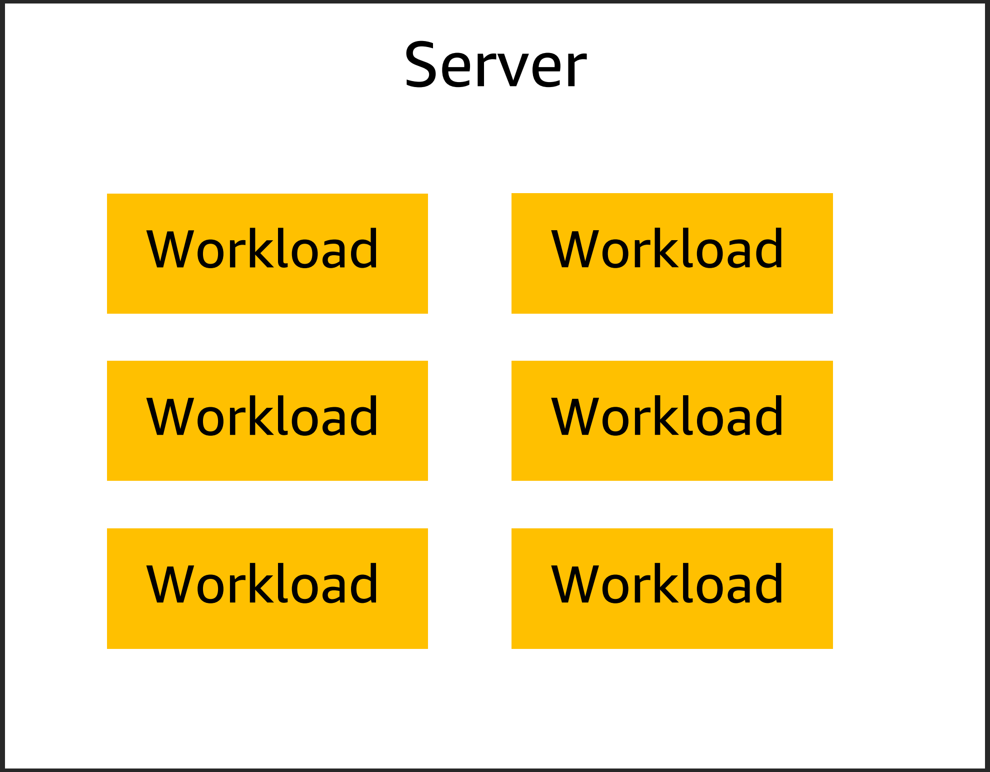
Good: Concentrate The Load



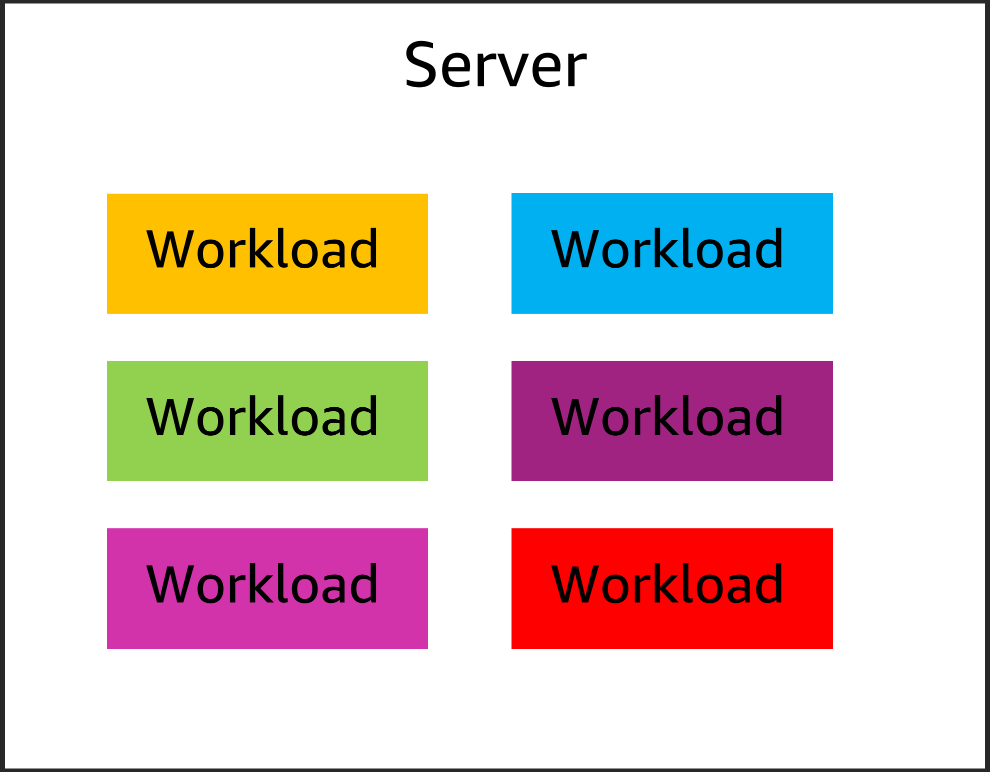
Cache Locality

Ability to Autoscale

Bad: Pack Server With One Workload

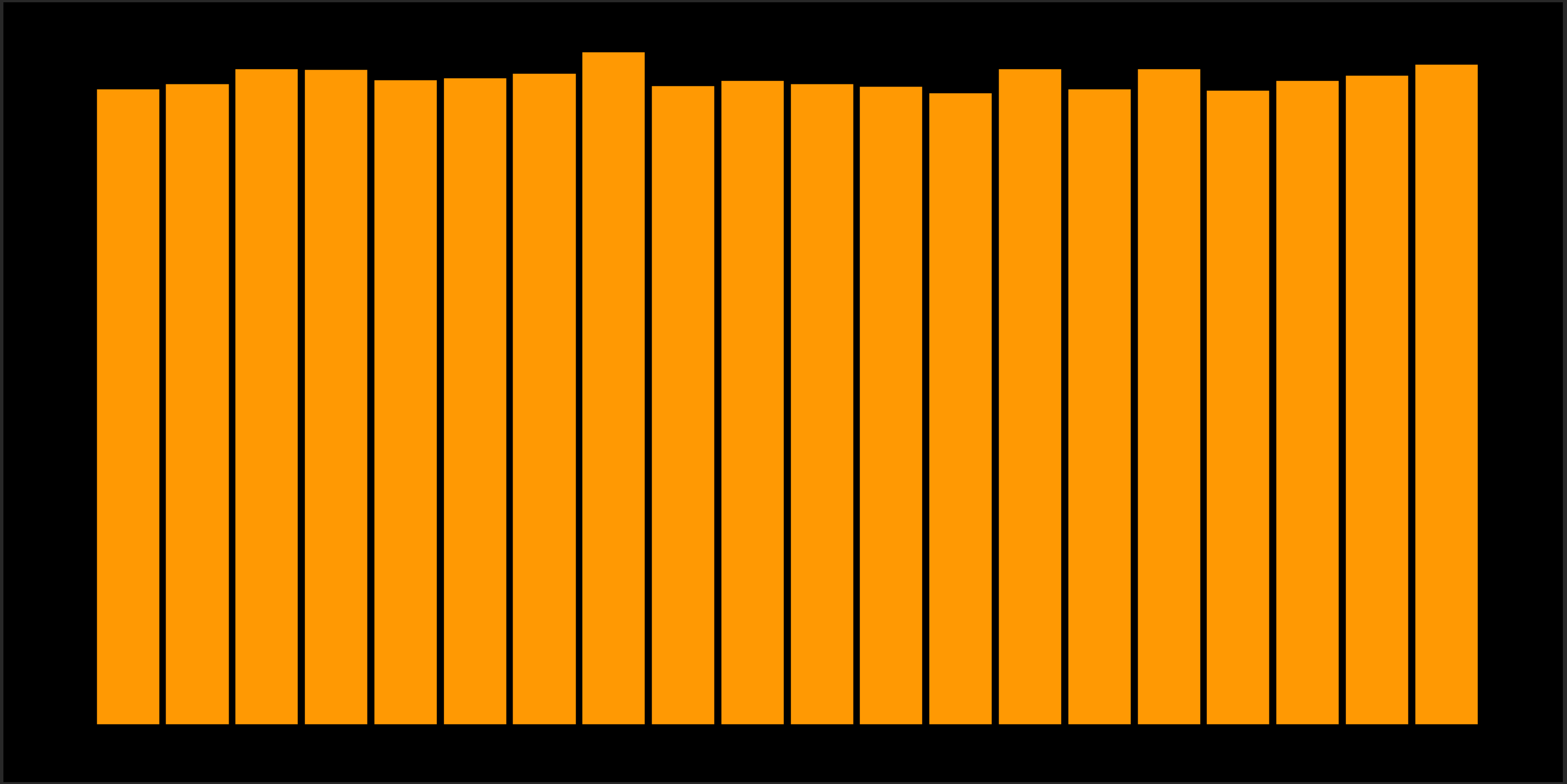


Better: Pack With Many Loads

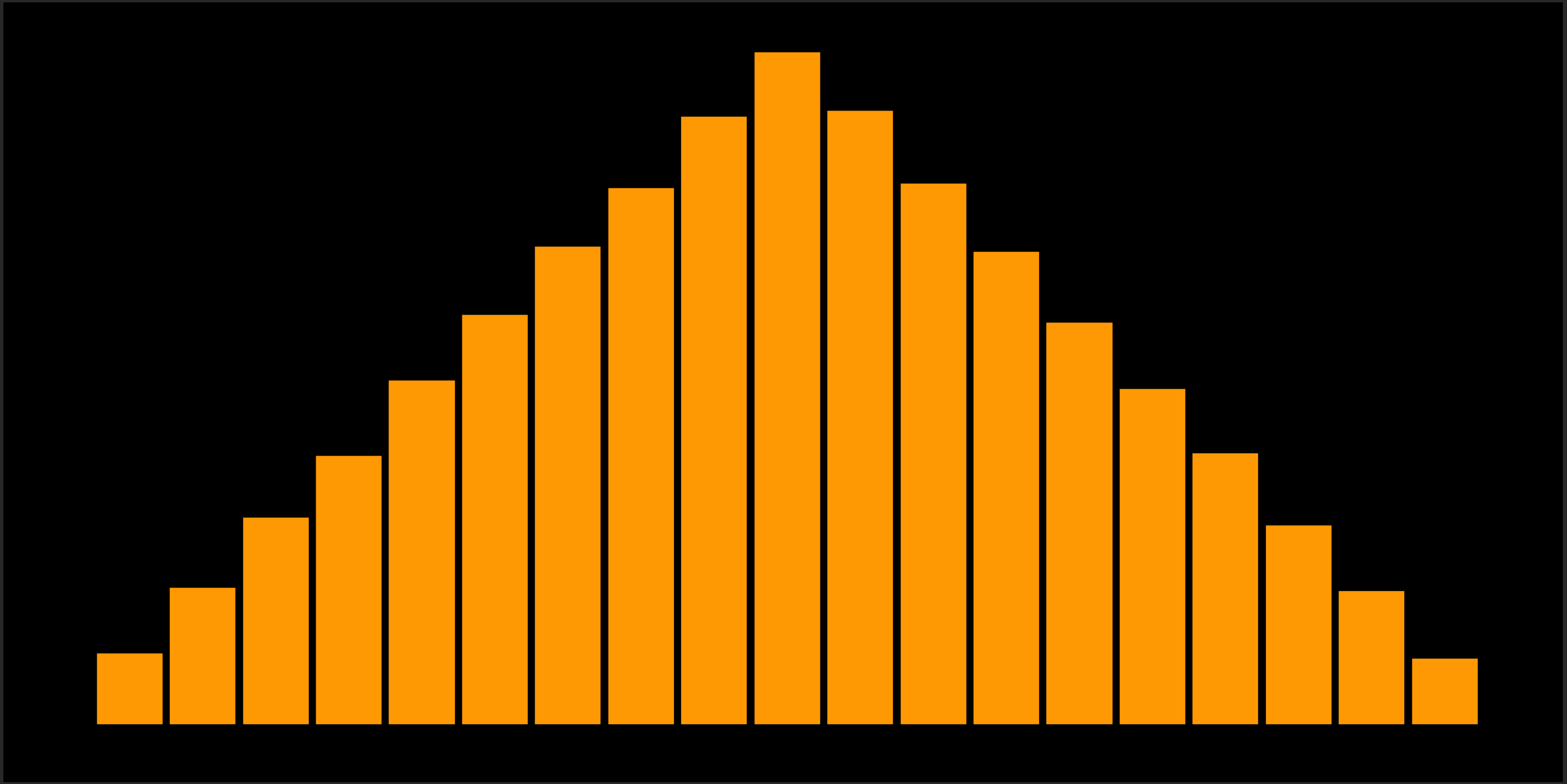


Take advantage of *Statistical Multiplexing*

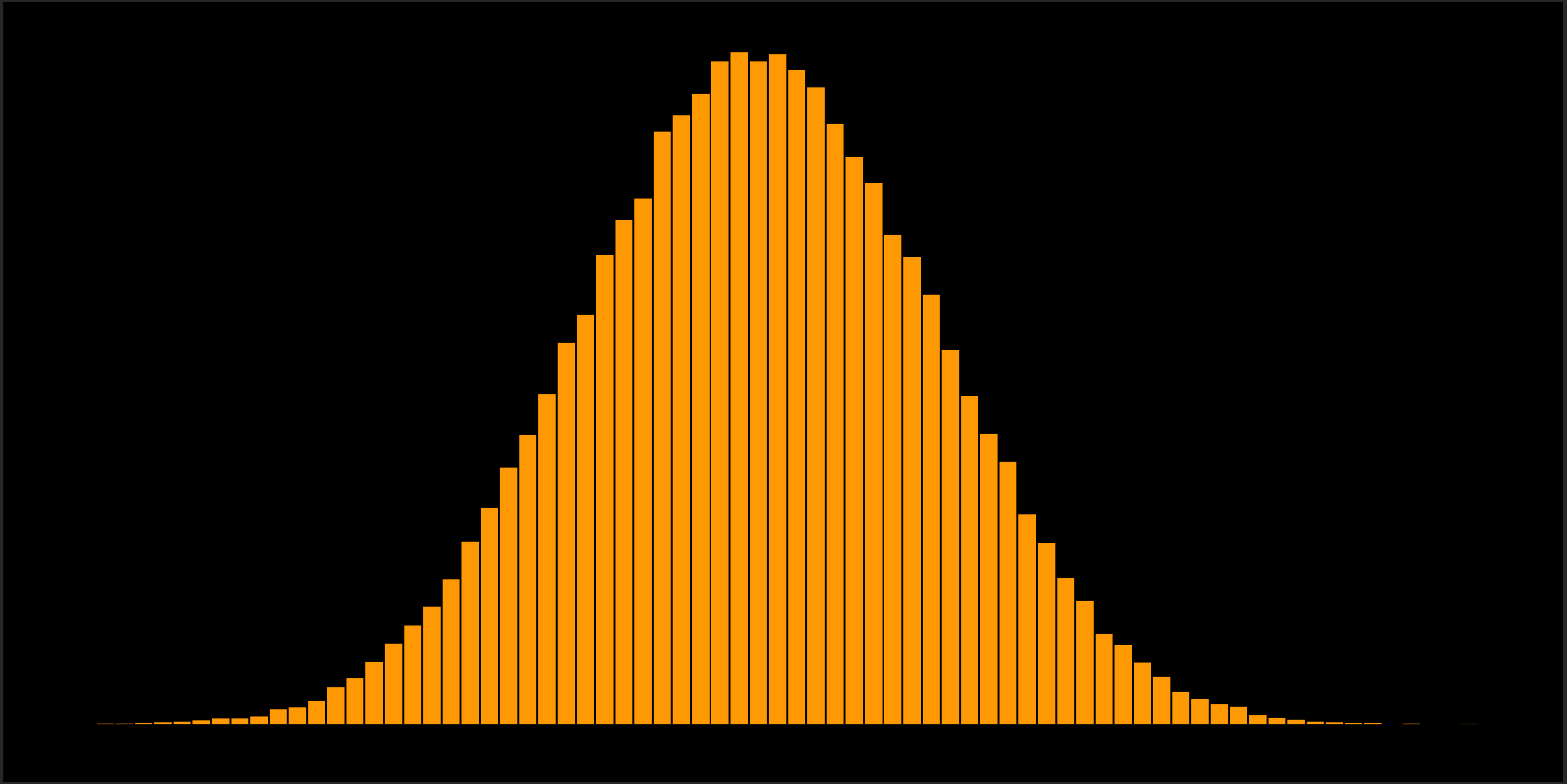
Throwing a d20



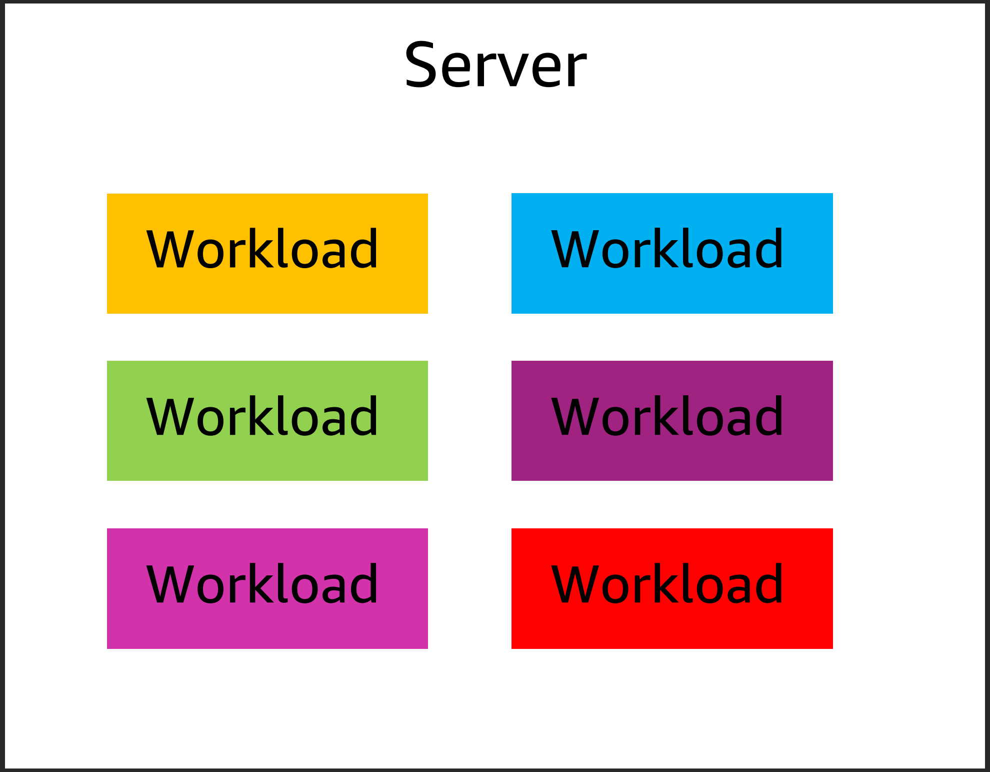
Throwing a d10, twice



Throwing a d10, ten times



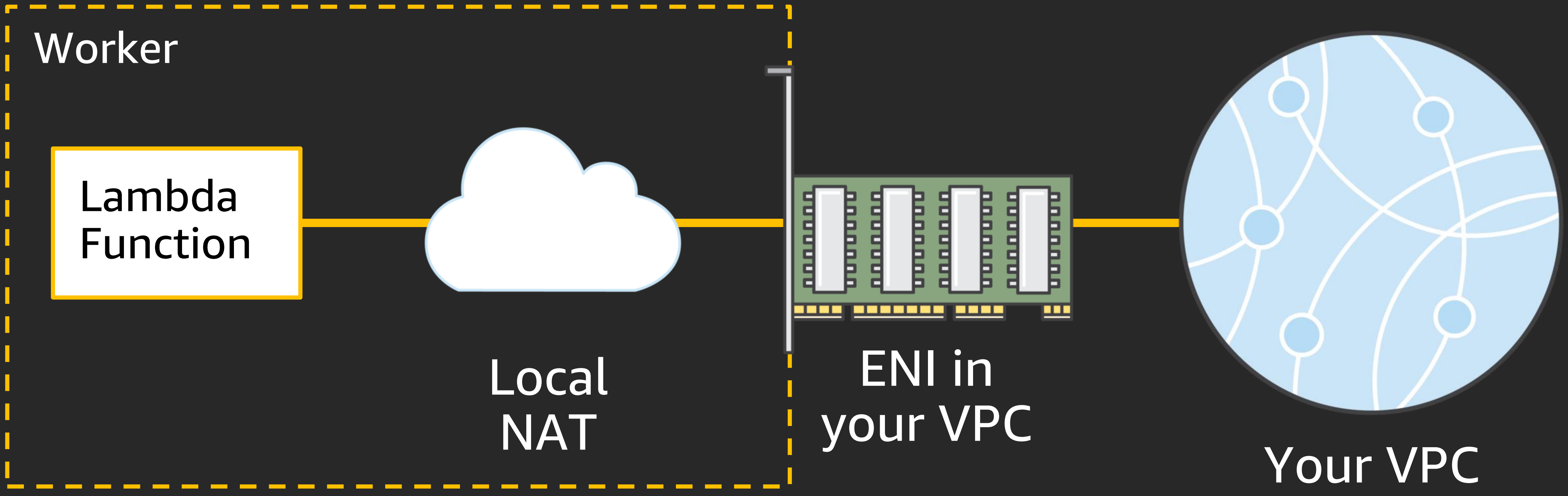
Best: Placement Optimization



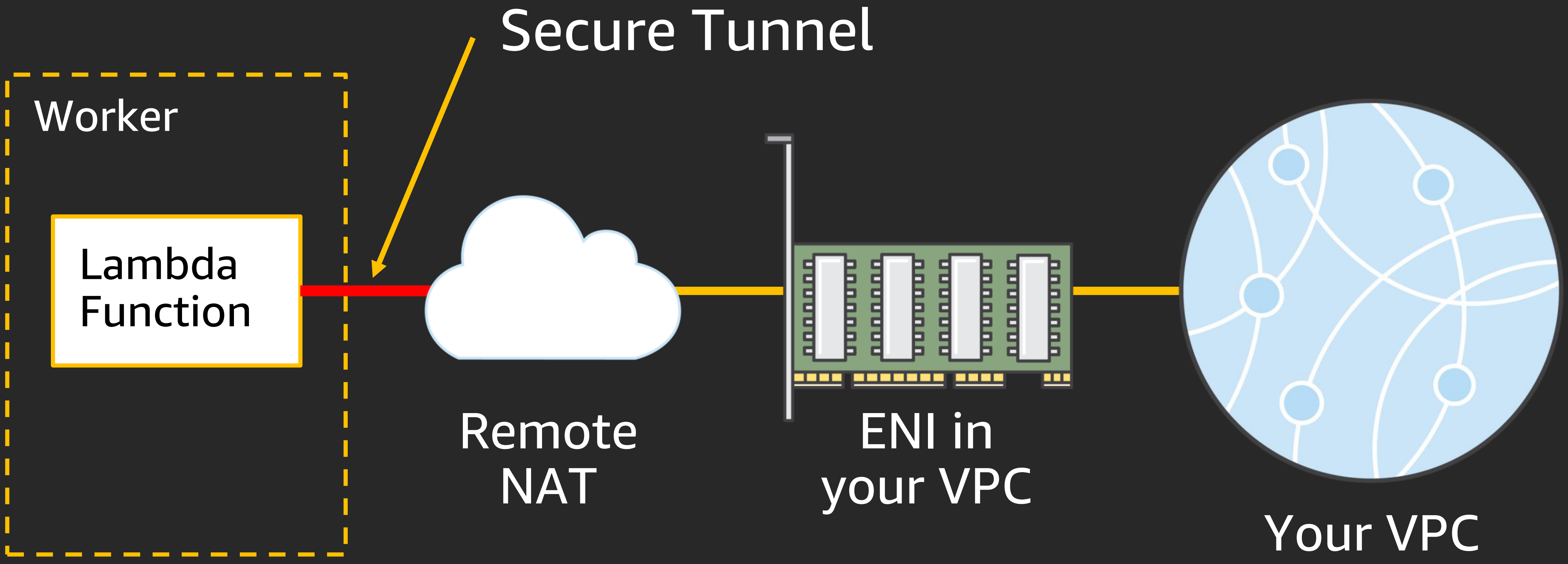
Pick workloads that pack together well.

Minimize contention.

Improving VPC start-up and scaling: now



Improving VPC start-up and scaling: 2019



Improving VPC start-up and scaling: benefits

Faster Scaling

Lower Latency

Easier To Use

Firecracker Hypervisor vs. Others

↓ Startup time

↓ Memory overhead

= Performance

↑ Flexibility

Firecracker Unlocks Higher Utilization and Scale

In Conclusion

Thank you!

Marc Brooker – Senior Principal Engineer, Amazon Serverless
Holly Mesrobian – Director of Engineering, Amazon Lambda

Please complete the session survey in the mobile app.