

The background features a dark blue gradient with abstract, overlapping shapes in shades of purple and magenta. Two thin, light blue lines intersect to form a large 'A' shape. The text is positioned on the left side of the image.

AWS re:Invent

DECEMBER 2 - 6, 2024 | LAS VEGAS, NV

KUB201

The future of Kubernetes on AWS

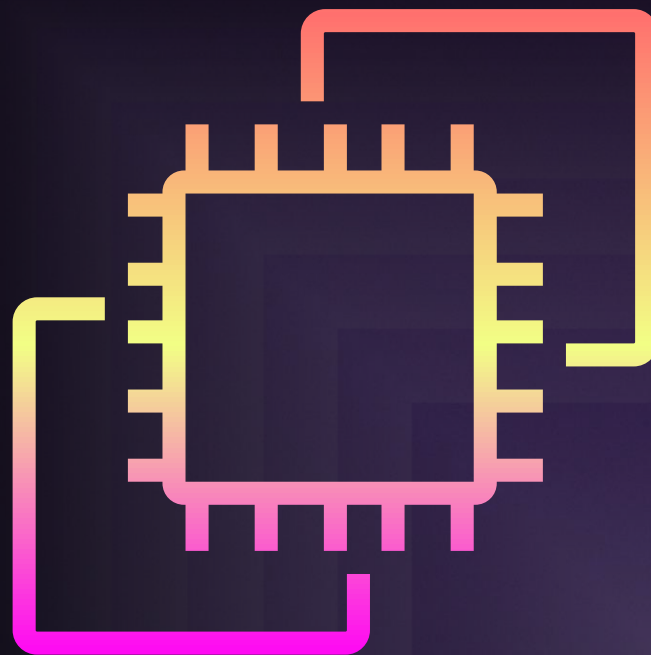
Nathan Taber

Head of Product
AWS

Hyungtae Kim

Principal Engineer
Snowflake







AWS Services launched



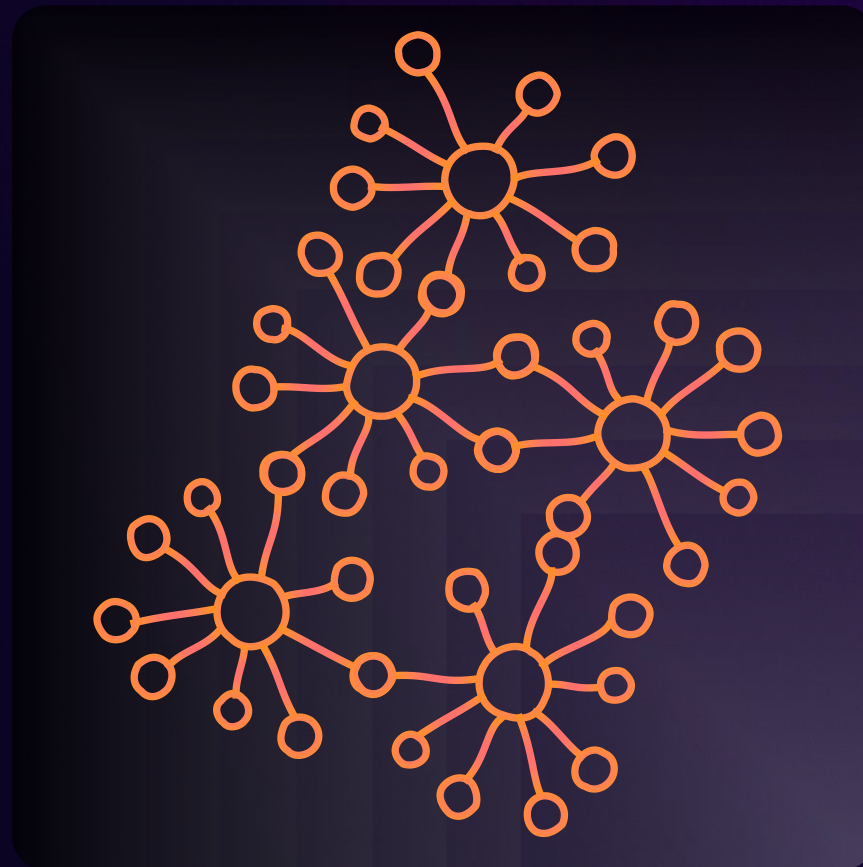




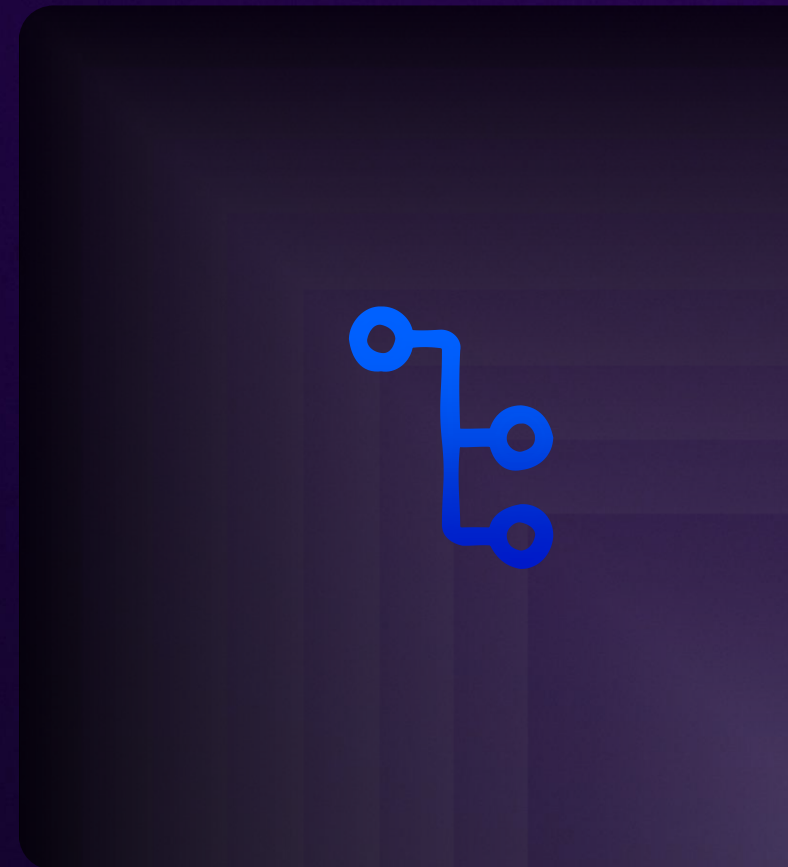
64% Using in production

25% Piloting / evaluating

Simplicity



AWS SDK
10,000 methods



Kubernetes Core
1,500 methods



Simplicity

Consistency

Extensibility

Simplicity

Consistency

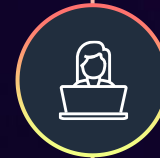
Extensibility



195 CNCF projects



100s of compatible tools



Unlimited customization

7 years of Managed Kubernetes on AWS

2018

2019

2020

2021

2022

2023

2024

EKS Generally Available

Managed Cluster Version Updates

GPU Support

HIPAA eligible

ISO, PCI, and SOC Compliance

Expansion to 15 AWS regions

CSI drivers for EBS, EFS, FSx for Lustre

Pod security policies

Managed Node Groups

EKS Fargate

EKS on AWS Outposts

Price reduced to \$0.10 per hour

Secrets Encryption

SLA raised to 99.95%

EKS CIS Benchmark

ACK Project

Load balancer controller

EKS on AWS Local zones

K8s Resources in EKS console

EKS Add-ons

OIDC access authentication

Karpenter project

Cluster creation reduced by 40%

Control plane scaling

EKS Anywhere GA

EKS Connector

FedRamp High Compliance

EFA, P4d instance support

IPv6 clusters available

PrivateLink support

Local clusters on Outposts

Trainium instances
Add-ons from AWS marketplace

Nitro enclaves

Launch time reductions

Expansion to 32 AWS regions

Kubernetes Network Policy enforcement

Upgrade Insights

Extended version support

Version launch acceleration

Pod Identity

S3 Mountpoints CSI Driver

Amazon Linux 2023

Automatic Version Upgrades

Zonal Shift

Full IPv6 support

Metrics Dashboard

Auto Mode

Hybrid Nodes

Node health & auto-repair





AMAZON EKS

Runs tens
of millions
of clusters every year

snowflake



AWS Enables Modular Platform Construction

Application-ready, production platform components in the cloud and data center

What this means for you:

Focus on your expertise

Reliable systems at scale

A lower risk profile

Kubernetes in context

Applications, Code, Data

Container packaging

Registry

Developer Tooling

Management Tooling

Kubernetes Control Plane

Infrastructure

Kubernetes in context

Applications, Code, Data

Container packaging

Registry

Developer Tooling

IDP, Jobs, and ML workflows

Management Tooling

Deployment, Observability, Governance,
Traffic, Security

Kubernetes Control Plane

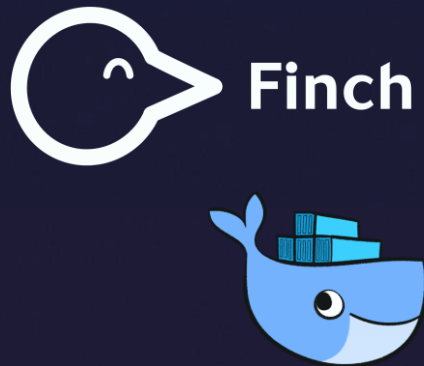
Scale, Availability, Integrations & Extensions

Infrastructure

Compute, Networking, Storage

Everything starts with the container

Registry



Finch

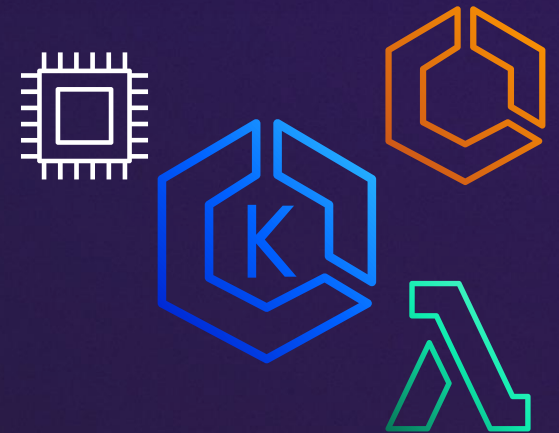
Build

*with any client including
Finch and Docker*



Store

securely with Amazon ECR



Deploy

on EKS, ECS, Lambda or anywhere else

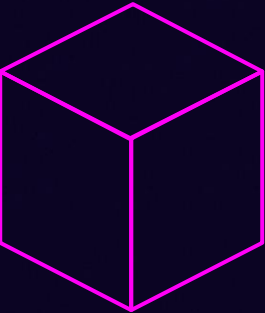
Amazon ECR Basic and **Enhanced** Image Scanning

New!

- Powered by Amazon Inspector
- Improved accuracy
- Results cover over 50 vulnerability databases and 12+ operating systems

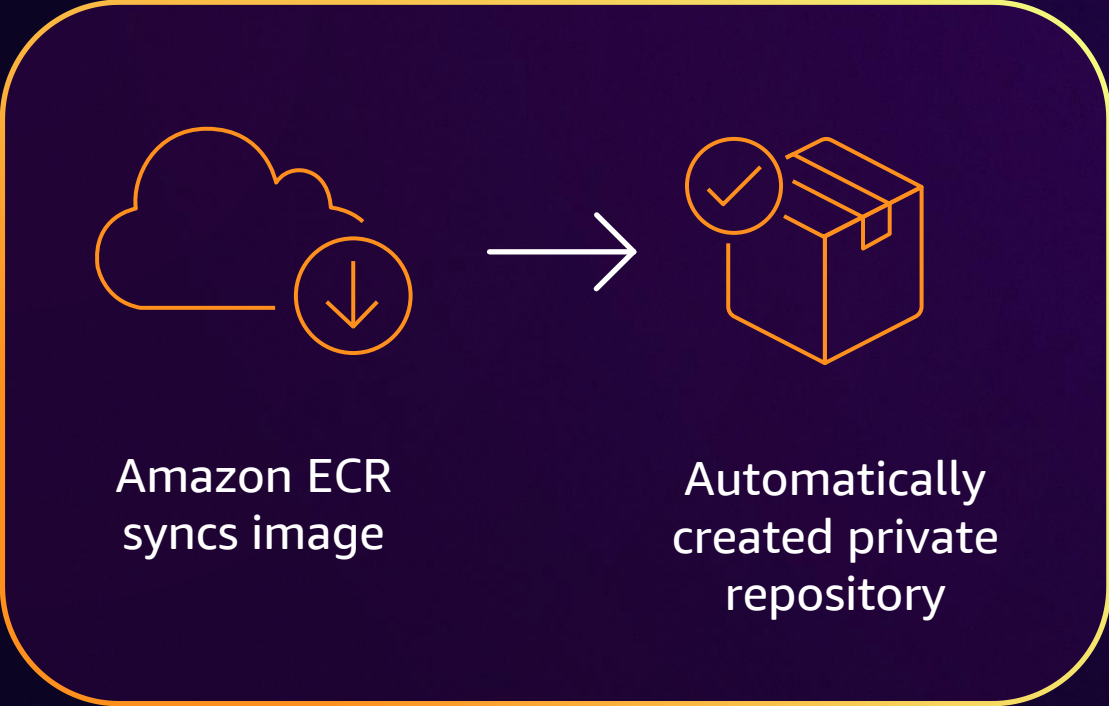
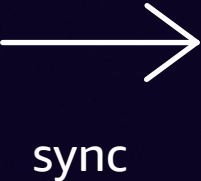


Authenticated pull through cache



Upstream registry

Docker Hub
GitHub Container Registry
...and more!



Amazon ECR syncs image

Automatically created private repository

pull through cache rule



Image pulled securely from ECR



Available Now!



AMAZON ECR

over 2 billion
image pulls

served every day

Kubernetes in context

Applications, Code, Data

Container packaging

Registry

Developer Tooling

Management Tooling

Kubernetes Control Plane

Infrastructure

Kubernetes Control Plane Version launch acceleration



Extended Support for Amazon EKS

- Kubernetes versions supported for an additional 12 months past project support
- Create clusters and upgrade at any time using versions in extended support
- AWS security patching for control plane, default add-ons, and AMIs
- *Upgrade Policy* keeps control planes automatically upgraded on standard support versions

New!



Now available for all EKS versions

Upgrade insights (5) [Info](#)

The table below lists the insight checks performed by EKS against this cluster, along with their associated statuses. EKS automatically refreshes the status of each Insight, which can be seen in the last refresh time column.

🔍 *Filter insights by name, version or status*

< 1 >

Name ▾	Insight status ▾	Version ▾	Last refresh time (UTC-08:00) ▾	Last transition time (UTC-08:00) ▾	Description ▾
Deprecated APIs removed in Kubernetes v1.26	✔ Passing	1.26	10 hours ago	April 16, 2024, 00:38	Checks for usage of deprecated APIs that are scheduled for removal in Kubernetes v1.26. Upgrading your cluster before migrating to the updated APIs supported by v1.26 could cause application impact.
Deprecated APIs removed in Kubernetes v1.32	✔ Passing	1.32	10 hours ago	February 16, 2024, 07:38	Checks for usage of deprecated APIs that are scheduled for removal in Kubernetes v1.32. Upgrading your cluster before migrating to the updated APIs supported by v1.32 could cause application impact.
Deprecated APIs removed in Kubernetes v1.29	✔ Passing	1.29	10 hours ago	November 18, 2023, 15:38	Checks for usage of deprecated APIs that are scheduled for removal in Kubernetes v1.29. Upgrading your cluster before migrating to the updated APIs supported by v1.29 could cause application impact.
Deprecated APIs removed in Kubernetes v1.25	✘ Error	1.25	10 hours ago	November 18, 2023, 15:38	Checks for usage of deprecated APIs that are scheduled for removal in Kubernetes v1.25. Upgrading your cluster before migrating to the updated APIs supported by v1.25 could cause application impact.
Deprecated APIs removed in Kubernetes v1.27	✔ Passing	1.27	10 hours ago	November 18, 2023, 15:38	Checks for usage of deprecated APIs that are scheduled for removal in Kubernetes v1.27. Upgrading your cluster before migrating to the updated APIs supported by v1.27 could cause application impact.

Upgrade insights (5) [Info](#)

The table below lists the insight checks performed by EKS against this cluster, along with their associated statuses. EKS automatically refreshes the status of each Insight, which can be seen in the last refresh time column.

🔍 Filter insights by name, version or status

Name	Insight status	Version
Deprecated APIs removed in Kubernetes v1.26	🟢 Passing	1.26
Deprecated APIs removed in Kubernetes v1.32	🟢 Passing	1.32
Deprecated APIs removed in Kubernetes v1.29	🟢 Passing	1.29
Deprecated APIs removed in Kubernetes v1.25	🔴 Error	1.25
Deprecated APIs removed in Kubernetes v1.27	🟢 Passing	1.27

Deprecation details [Info](#)

Status

🔴 Error

Usage

/apis/policy/v1beta1/poddisruptionbudgets

Stop serving version

1.25

Replaced with

/apis/policy/v1/poddisruptionbudgets

Start serving replacement version

1.21

Client stats (2)

🔍 Search

< 1 >

User agent

Number of requests in the last 30 days

Last request time

kube-state-metrics

5705

11 hours ago

app

11378

10 hours ago



Enhanced control plane observability

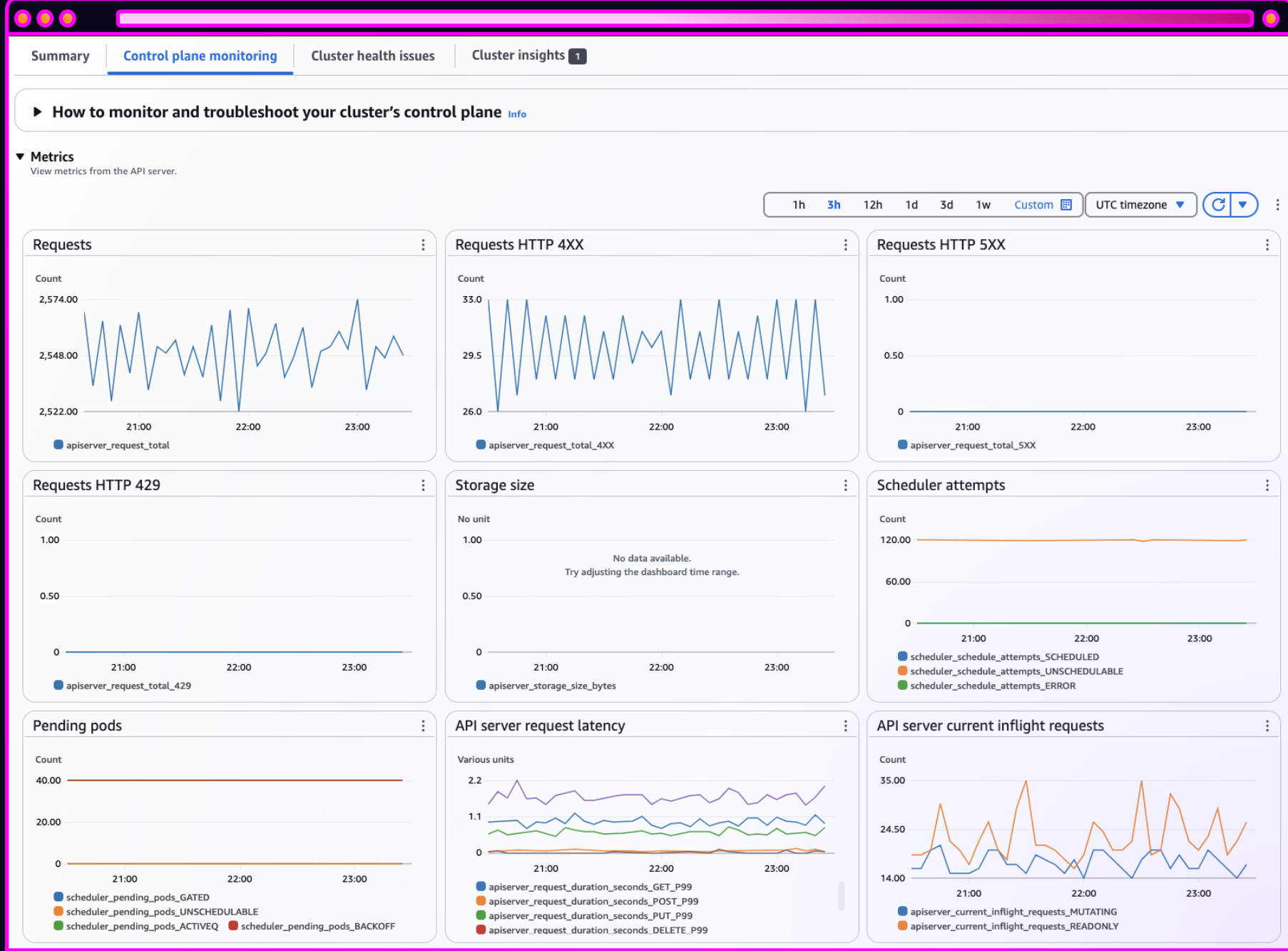
Additional metrics

- Kube-controller-manager
- kube-scheduler

Scape via new Prometheus endpoints!

Pre-configured console dashboards

- Key metrics
- CloudWatch Log Insights queries



Enhanced control plane observability

Additional metrics

- Kube-controller-manager
- kube-scheduler

Scape via new Prometheus endpoints!

Pre-configured console dashboards

- Key metrics
- CloudWatch Log Insights queries

CloudWatch Logs Insights
Each of these lists offers the option to run queries directly. To customize query parameters or change the query content, choose the applicable "View in CloudWatch" button.

[Run all queries](#) 5m 30m 1h 3h 12h Custom Local timezone

Top talkers (1000) Query last run: November 29, 2024, 10:01 (UTC-08:00) [Run query](#) [View in CloudWatch](#)

Show the most frequent callers to the API server, sorted by the number of requests made.

Search

Request URI	Verb	Response status code	User agent	Count
/apis/extensions/v1beta1/Ingresses?limit=500&resourceVersion=0	list	404	kube-state-metrics/v1.9.8 (linux/amd64) kube-state-metrics/e7231551	1701
/apis/certificates.k8s.io/v1beta1/certificatesigningrequests?limit=500&resourceVersion=0	list	404	kube-state-metrics/v1.9.8 (linux/amd64) kube-state-metrics/e7231551	1700
/apis/coordination.k8s.io/v1/namespaces/opentelemetry-operator-system/leases/9f7554c3.opentelemetry.io	get	200	manager/v0.0.0 (linux/amd64) kubernetes/\$Format/leader-election	1682
/apis/coordination.k8s.io/v1/namespaces/kube-system/leases/kube-scheduler?timeout=5s	get	200	kube-scheduler/v1.24.17 (linux/amd64) kubernetes/005dc71/leader-election	1388
/apis/coordination.k8s.io/v1/namespaces/kube-system/leases/cloud-controller-manager?timeout=5s	get	200	aws-cloud-controller-manager/v0.0.0 (linux/amd64) kubernetes/\$Format/leader-election	1388
/apis/coordination.k8s.io/v1/namespaces/kube-system/leases/fargate-scheduler?timeout=5s	get	200	eks-fargate-scheduler/v0.0.0 (linux/amd64) kubernetes/\$Format/leader-election	1384
/apis/coordination.k8s.io/v1/namespaces/kube-system/leases/kube-controller-manager?timeout=5s	get	200	kube-controller-manager/v1.24.17 (linux/amd64) kubernetes/005dc71/leader-election	1378
/api/v1/namespaces/amazon-cloudwatch/configmaps/cwagent-clusterleader	get	200	amazon-cloudwatch-agent/v0.0.0 (linux/amd64) kubernetes/\$Format	1210
/healthz?exclude=kms-provider-0&exclude=kms-provider-1&exclude=kms-providers	get	200	ELB-HealthChecker/2.0	1031
/apis/coordination.k8s.io/v1/namespaces/kube-system/leases/cloud-controller-manager?timeout=5s	update	200	aws-cloud-controller-manager/v0.0.0 (linux/amd64) kubernetes/\$Format/leader-election	856



Available Now!

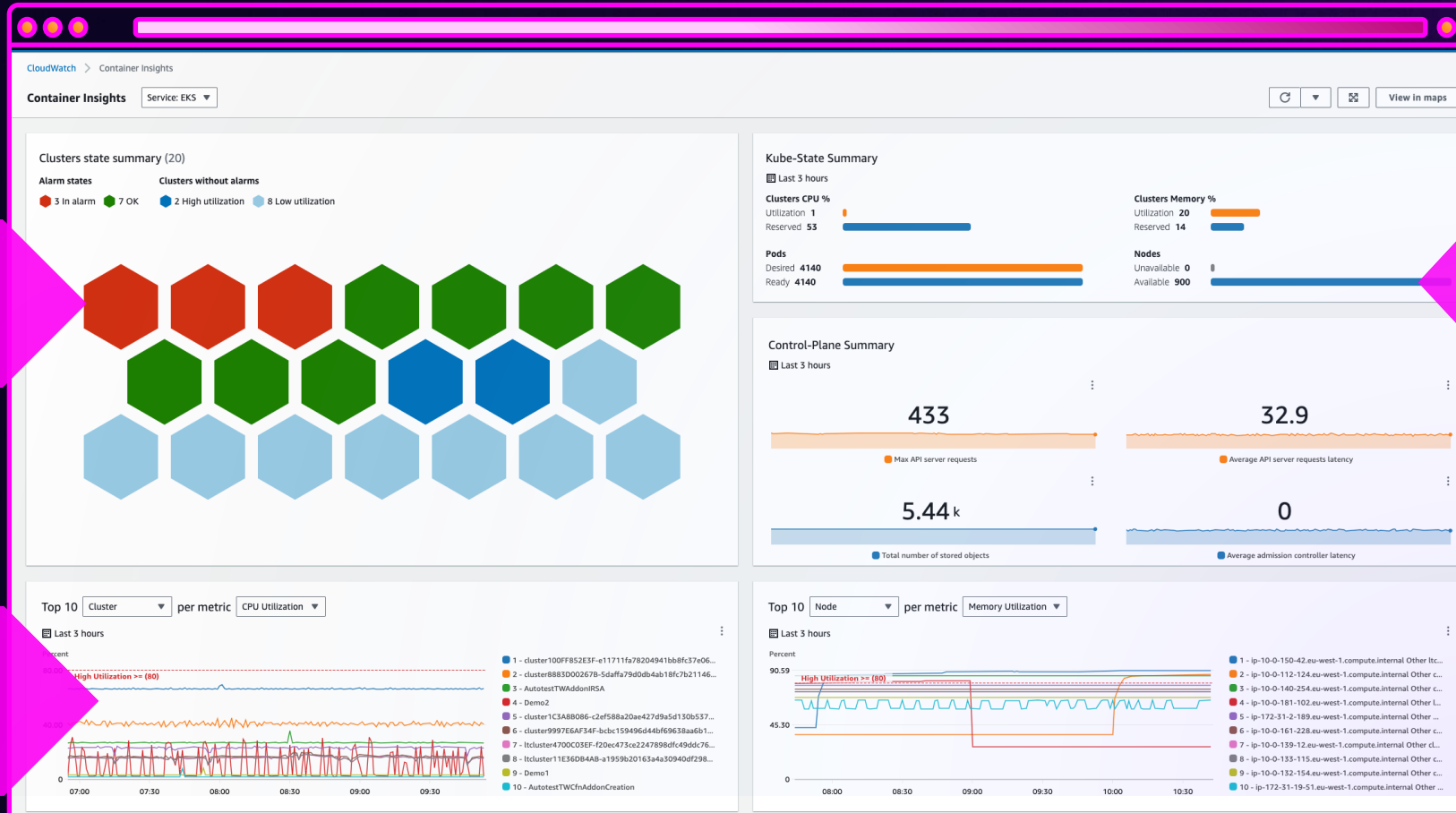
CloudWatch Container Insights with Enhanced Observability for Amazon EKS

ENHANCED CONTAINER OBSERVABILITY WITH EASY GETTING STARTED AND FASTER TROUBLESHOOTING

Performance overview with alarms and utilization status

Take proactive action from top resource consumers

Metrics give visibility into cluster health



New: GPU, Neuron, and Windows support

Available Now!



EKS support for CloudWatch Network Flow Monitor

New!

The AWS Network can be a Blackbox

- Customers lack visibility and transparency into the performance of their networks on AWS.
- Customers cannot identify AWS Network driven impairments to their workloads' performance.
- Customers experience prolonged MTTD and MTTR as they work with support and disparate data.

Answer the question – is it the AWS Network?

- Reduce MTTD for network performance driven workload impairments with loss and latency metrics.
- Kubernetes workload metadata annotations included by default.
- Accelerate MTTR via re-configuration as required using metrics and AWS Network Infrastructure Health Indicators.
- Expedite Root Cause Analyses using shared performance data when working with AWS support.

MTTD: Mean Time to Detection
MTTR: Mean Time to Resolution



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Available Now!



Split Cost Allocation Data (SCAD) for EKS

LineItem/ResourceId	LineItem/LineItemType	LineItem/UsageType	LineItem/UnblendedCost	SplitLineItem/ParentResourceId	SplitLineItem/SplitUsage	SplitLineItem/SplitCost	SplitLineItem/UnusedCost
i-12345	Usage	BoxUsage:m7g.2xlarge	0.33				
EC2-Pod1	Usage	EKS-EC2-vCPU-Hours		i-12345	2	0.051	0
EC2-Pod1	Usage	EKS-EC2-GB-Hours		i-12345	6	0.019	0.001
EC2-Pod2	Usage	EKS-EC2-vCPU-Hours		i-12345	3	0.076	0
EC2-Pod2	Usage	EKS-EC2-GB-Hours		i-12345	10	0.032	0.002
EC2-Pod3	Usage	EKS-EC2-vCPU-Hours		i-12345	2	0.051	0
EC2-Pod3	Usage	EKS-EC2-GB-Hours		i-12345	6	0.019	0.001
EC2-Pod4	Usage	EKS-EC2-vCPU-Hours		i-12345	2	0.051	0
EC2-Pod4	Usage	EKS-EC2-GB-Hours		i-12345	8	0.025	0.002

- Allocate EC2 costs across pods, namespaces, and clusters
- Automatically ingests true EC2 costs for EKS clusters
- Native to AWS Cost and Usage Reporting



Amazon EKS add-ons

Clusters with batteries-included

Configure add-ons before launch

Launch clusters without core networking add-ons **New!**

Subscribe to marketplace add-ons directly from EKS

Expanded Catalog **New!**

CloudWatch Container Insights

CSI Snapshot Controller

Pod Identity Agent

Node Monitoring Agent

Over 40 Marketplace add-ons including:

Kubecost, Datadog, Upbound UXP, Kubearmor, Gloo,

Akuity, New Relic, Splunk, Datree, Dynatrace, Rafay,

Stormforge, Kong, and more!



Add-ons integration with EKS Pod Identity

Simplify secure cluster setup

Create and associate pod identity roles with add-ons at cluster creation or add-on install.

Simplifies setup of critical cluster operational software that needs to interact with AWS services outside the cluster.

Faster application ready clusters

Expands the selection of Pod Identity compatible EKS add-ons from AWS and Marketplace available for installation through the EKS console during cluster creation.



Default encryption with KMS v2

Improved security posture by default

Default envelope encryption with an AWS owned key. defense-in-depth for your Kubernetes applications.

Encryption for all cluster objects

EKS now encrypts all Kubernetes API data in addition to secrets.

Optionally, continue to use your own customer managed key (CMK) in AWS KMS to envelope encrypt all cluster objects.

Improved performance

With KMS v2, a new DEK is only generated on API server startup and when the KMS plugin informs the API server that a KEK rotation has occurred.



Cedar Access Controls for Kubernetes

New!

Consolidated policy authoring experience

Cedar is an open source policy language created by AWS. Author policies for both authorization and admission in Kubernetes using expressive permissions.

Enhanced authorization features

Support for features not available in K8s RBAC today like denials, conditions, and attribute and label-based access controls.

Part of our vision for secure clusters with truly separate tenants.



Now available in alpha

github.com/awslabs/cedar-access-control-for-k8s



IPv6

Scale EKS with full support for IPv6

- EKS management APIs
- Cluster endpoints **New!**
- Single stack pods, dual stack nodes
- Assign security groups to IPv6 pods

Available Now!

Amazon Application Recovery Controller (ARC) ^{New!}

Improved resilience

Works in tandem with Kubernetes native built-in protections to improve application environment fault tolerance.

Deeply integrated

Enable ARC on additional services like ELB to provide AWS service wide resilience for applications.

Flexible implementation

Enable manually or via ARC zonal autoshift.



Available Now!

AWS Controllers for Kubernetes (ACK)

Cloud-Native control

Define the AWS resources your applications need directly within the cluster.

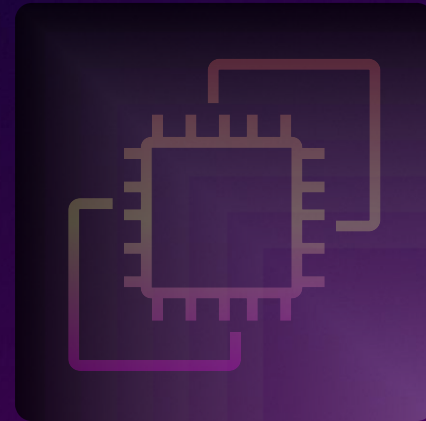
Always up to date

Controllers refreshed automatically from AWS SDKs

Harness the full cloud

41 AWS services supported in GA, plus 12 in preview.

20 new services in GA this year!



Now available

github.com/aws-controllers-k8s



Kube Resource Orchestrator (KRO) **New!**

Simplify platform building

Build high-level abstractions with complex Kubernetes resource configurations.

Develop using common expression language

Publish abstractions as APIs

Unify cloud and Kubernetes resource management

Automate the dynamic creation of custom Kubernetes resources in the cluster.

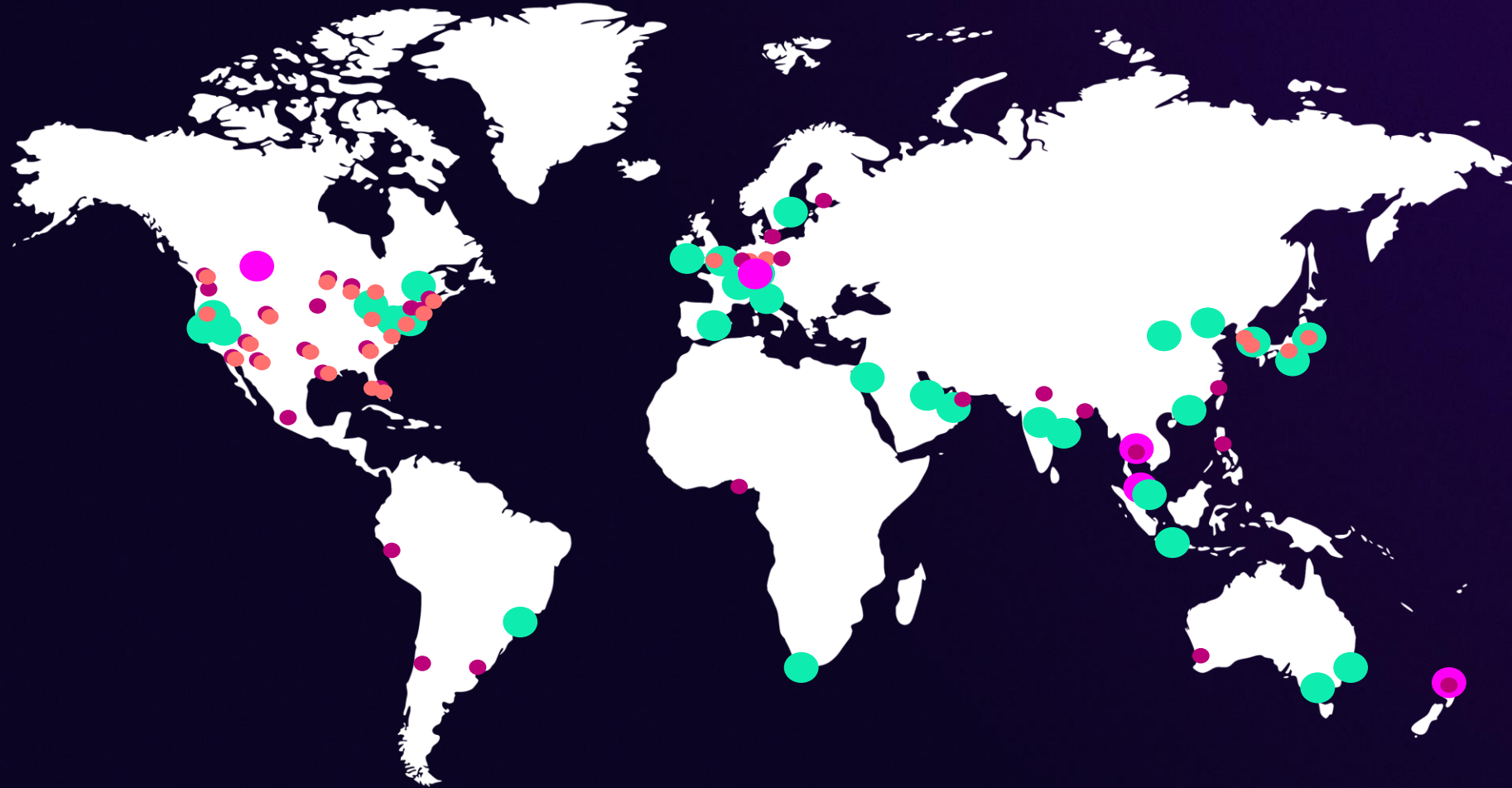
Support for any Kubernetes objects including native objects like services and jobs along with custom resources like ACK.



Now available in Alpha
github.com/awslabs/kro










Global reach



Run EKS in

- 34 Geographic Regions
- 108 Availability Zones
- 41 Local Zones
- 29 Wavelength Zones

Run Kubernetes everywhere

	 EKS Distro	 EKS Anywhere	 Hybrid Nodes	 EKS on Outposts	 EKS on Wavelength	 EKS on Local Zones	 Amazon EKS
Hardware	Any	Customer	Customer	aws	aws	aws	aws
Location	Any	On-prem	On-Prem	On-prem	Wavelength	Local Zone	aws
K8s control plane	Any	On-prem	aws	aws	aws	aws	aws
K8s Nodes	Any	On-prem	On-prem	Outpost	Wavelength	Local Zone	aws
Support	Community	aws	aws	aws	aws	aws	aws
Region connectivity required	No	No	Yes	Yes	Yes	Yes	Yes

Customer Managed
AWS Managed





AMAZON EKS

Hybrid Nodes

Extend clusters anywhere

Amazon EKS Hybrid Nodes

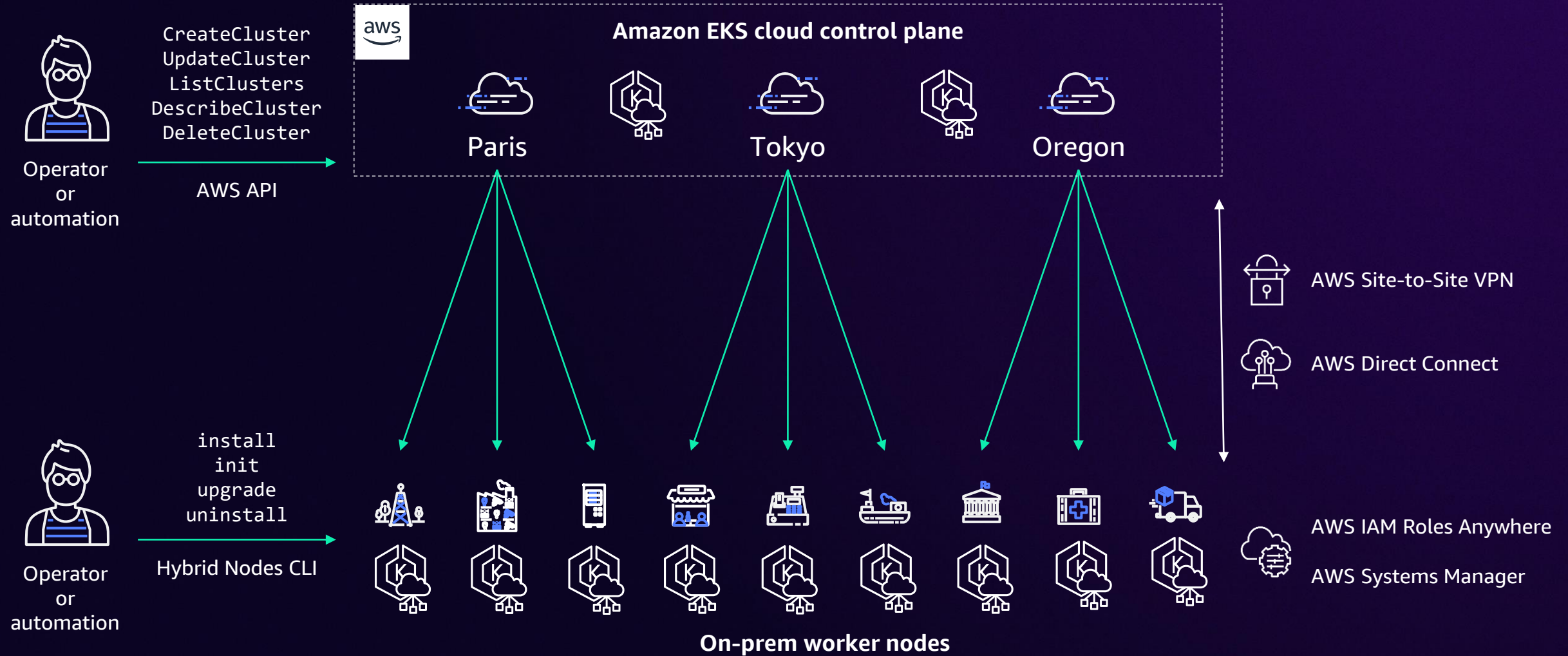
BRING THE POWER OF AMAZON EKS TO YOUR ON-PREMISES APPLICATIONS

Customers can now use existing on-premises and edge infrastructure as nodes in Amazon EKS clusters for unified Kubernetes management across environments

- ✓ Improve operational efficiency by unifying Kubernetes operations across environments
- ✓ Reduce total cost of ownership of managing Kubernetes
- ✓ Get the benefits of AWS Cloud on premises
- ✓ Gain the flexibility to run your workloads anywhere



Amazon EKS Hybrid Nodes architecture



Amazon EKS Hybrid Nodes in action

mi-0089533dd63828ed1 Structured view Raw view

hybrid-demo

Delete cluster Upgrade version

ⓘ End of standard support for Kubernetes version 1.30 is July 28, 2025. On that date, your cluster will enter the extended support period with additional fees. For more information, see the [pricing page](#). Upgrade now

▼ Cluster info Info

Status ✔ Active	Kubernetes version <a>Info 1.30	Support period <a>ⓘ Standard support until July 28, 2025	Provider EKS
---------------------------	---	--	------------------------

Overview | Resources | Compute | Networking | Add-ons **1** | Access | Observability | Update history | Tags

Nodes (1) Info

🔍 *Filter Nodes by property or value* < 1 >

Node name	Instance type	Compute	Managed by	Created	Status
mi-0089533dd63828ed1	-	Hybrid	-	Created 📅 October 9, 2024, 15:41 (UTC-05:00)	✔ Ready

Amazon EKS Hybrid Nodes in action

hybrid-demo

End of standard support for Kubernetes version 1.30 is July 28, 2025. On that date, your cluster will enter the extended support period with additional fees. For more information, see the [Kubernetes version 1.30 end of support page](#).

Cluster info

Status: Active | Kubernetes version: 1.30 | Support period: [Standard support until July 28, 2025](#)

Overview | Resources | **Compute** | Networking | Add-ons 1 | Access | Observability | Update history | Tags

Nodes (1)

Filter Nodes by property or value

Node name	Instance type	Compute	Managed by	Created
mi-0089533dd63828ed1	-	Hybrid	-	October 9, 2024, 15:41 (UTC-05:00)

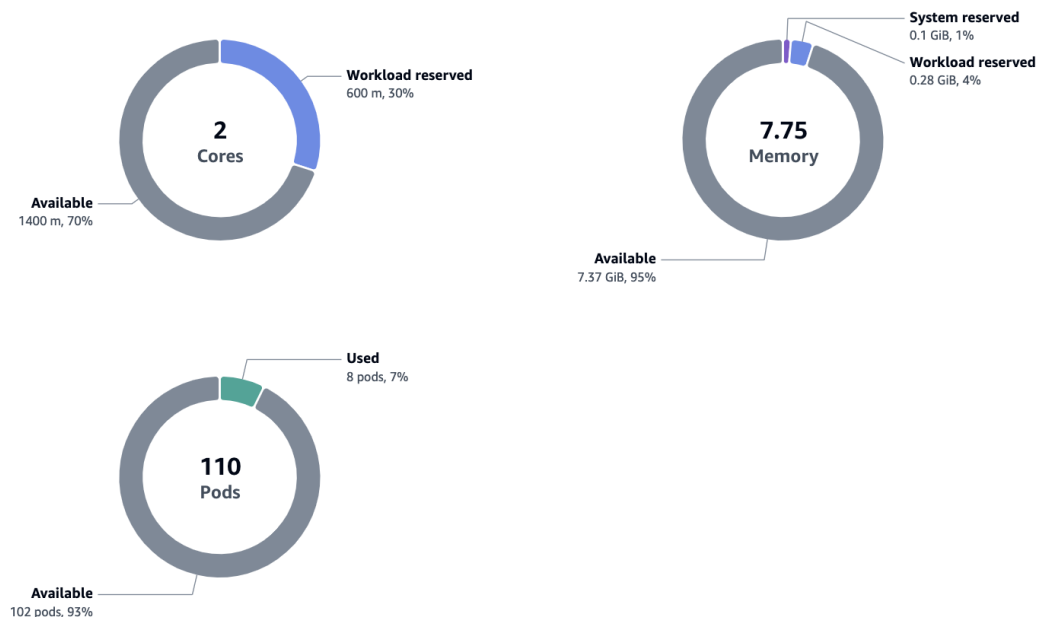
mi-0089533dd63828ed1

Structured view | Raw view

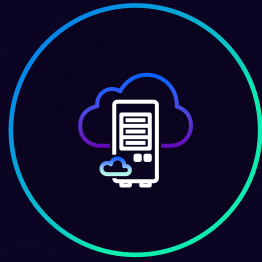
Details

Status: Ready	Kernel version: 5.15.0-122-generic	Created: October 9, 2024, 15:41 (UTC-05:00)
Last transition time: October 15, 2024, 02:46 (UTC-05:00)	Node group: -	Container runtime: containerd://1.7.12
OS (Architecture): linux (amd64)	Instance: mi-0089533dd63828ed1	Kubelet version: v1.30.0-eks-036c24b
OS image: Ubuntu 22.04.4 LTS	Instance type: -	

Capacity allocation



Amazon EKS Hybrid Nodes use cases



Enterprise
modernization



Machine
learning



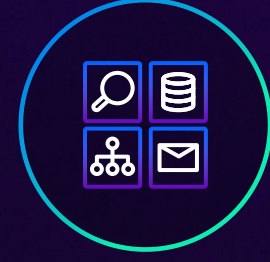
Financial
services



Media
streaming



Manufacturing



Internal IT apps

Categories

General purpose

Burstable

Compute intensive

Memory intensive

Storage (high I/O)

Dense storage

GPU compute

Graphic intensive

Capabilities

Choice of processor
(AWS, Intel, AMD, Apple)

High memory footprint
(up to 24 TiB)

Accelerated computing
(GPUs and FPGA)

Instance storage (HDD
and NVMe)

Size
(up to 112x large)

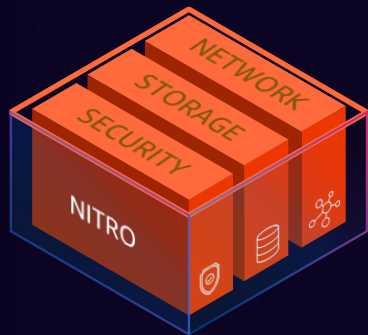
Networking
(up to 800 Gbps)

800+

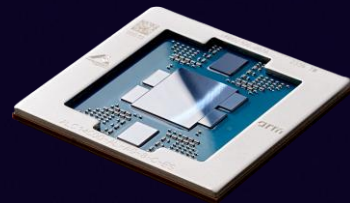
instance types

for virtually
every workload

Silicon innovation



Powered by
AWS Nitro System



AWS Graviton Processors best
price performance for cloud
workloads

Purchase options

ON-
DEMAND
INSTANCES

SAVINGS
PLANS

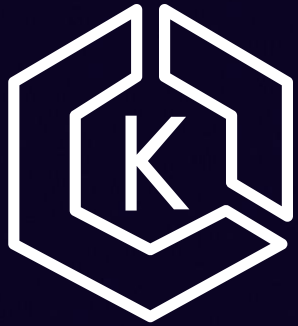
SPOT
INSTANCES



AMAZON EKS

Auto Mode

Automate your Kubernetes
cluster infrastructure



Amazon EKS Auto Mode

**FULLY AUTOMATE YOUR
KUBERNETES CLUSTERS**



Increase agility and accelerate innovation by offloading cluster operations to AWS

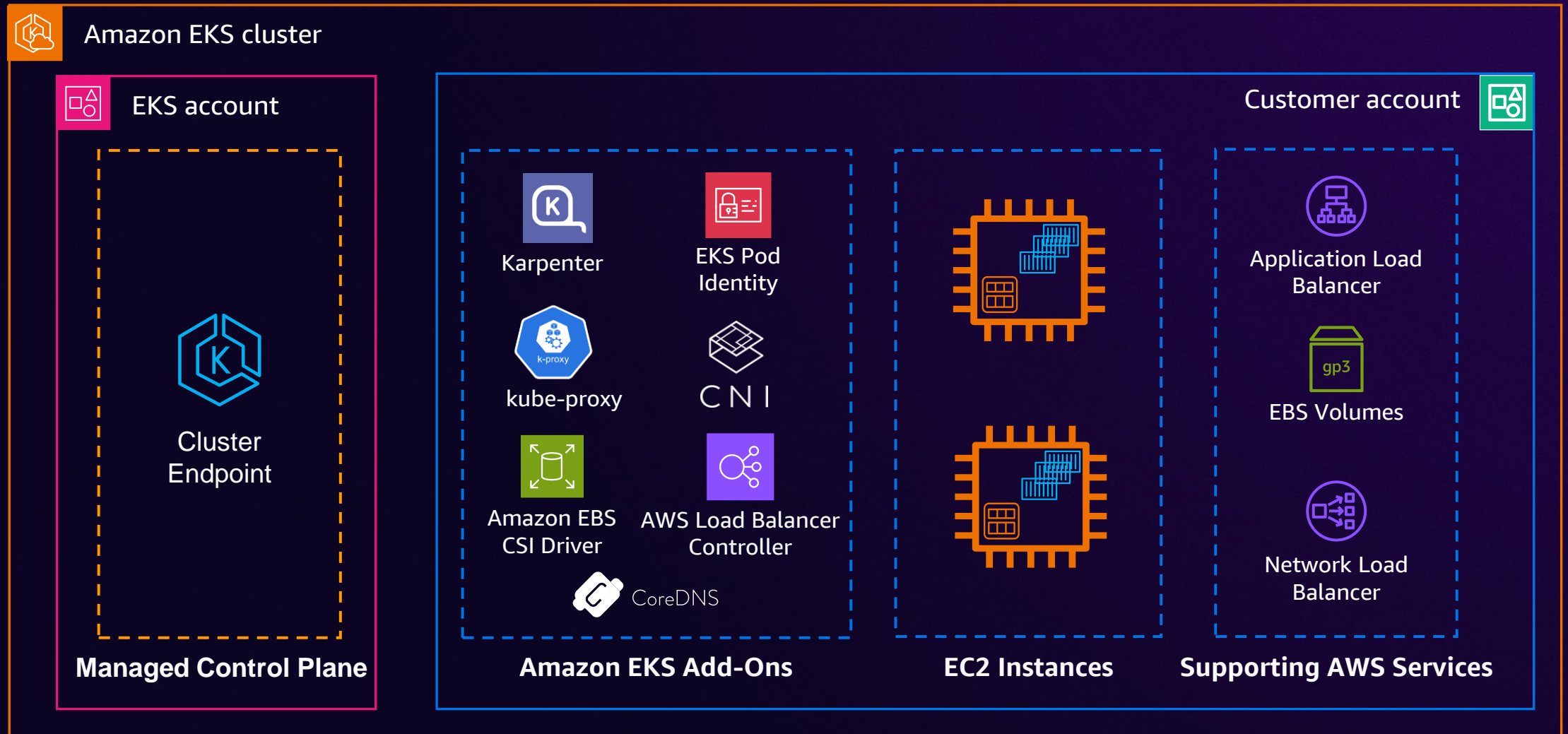


Improve performance, availability, and security of your applications with AWS operational excellence

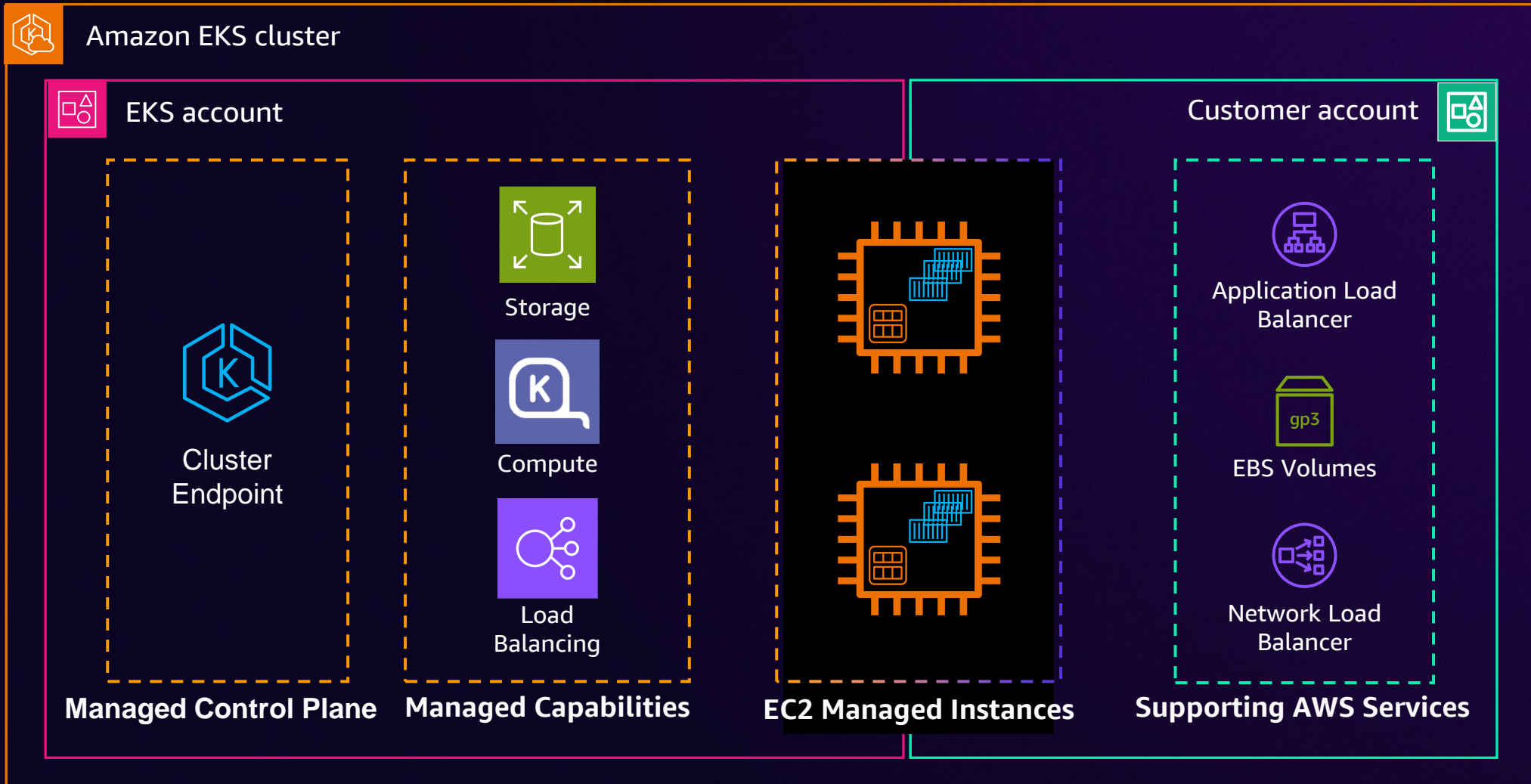


Optimize compute costs with automatic capacity planning and dynamic scaling

Previous EKS cluster architecture



EKS cluster architecture with Auto Mode



Easier and faster to get started



Get started with one click



Available for any new or existing cluster



Easily switch mode on or off



Built-in best practices

Configure cluster

Configuration options [Info](#)

Choose how you would like to configure the cluster.

Quick configuration (new with EKS Auto mode)

Quickly create a cluster with production-grade default settings. The configuration uses EKS Auto mode to automate infrastructure tasks like creating nodes and provisioning storage.

Custom configuration

To change default settings prior to creation, choose this option. This configuration gives the option to use EKS Auto mode and customize the cluster's configuration.

Cluster configuration

Name

Use the auto-generated name or enter a unique name for this cluster. This property cannot be changed after the cluster is created.

Autogenerated_cluster_name

The cluster name should begin with letter or digit and can have any of the following characters: the set of Unicode letters, digits, hyphens and underscores. Maximum length of 100.

Version

Select Kubernetes version for this cluster.

1.28 (Latest version)

Cluster IAM role [Info](#)

Select the IAM role to allow the Kubernetes control plane to manage AWS resources on your behalf. This cannot be changed after the cluster is created.

MyEKSRole



Create recommended role [↗](#)

Node IAM role [Info](#)

Nodes need an EC2 Instance IAM Role to launch and register with a cluster. This cannot be changed after the cluster is created.

Choose a role



Create recommended role [↗](#)

VPC [Info](#)

Select a VPC to use for your EKS cluster resources.

VPC-0jhf7465uhf83huh | Default



Create VPC [↗](#)

Subnets [Info](#)

Choose the subnets in your VPC where the control plane may place elastic network interfaces (ENIs) to facilitate communication with your cluster.

Select subnets

subnet-0ihfg83478h
us-west-2a 172.31.0.0/20

subnet-8hugh47ytd
us-west-2b 172.31.12.0/20

subnet-38ghu4uh
us-west-2c 172.31.0.18/20

▶ View quick configuration defaults

Cancel

Create

Amazon EKS Node Health and Auto-Repair ^{New!}

Comprehensive monitoring and remediation

Continuously monitors the health of node components including kubelet, CNI, disk, and CPU.

Can automatically replace nodes when issues arise.

Kubernetes native

Awareness of existing Kubernetes disruption controls such as Pod Disruption Budgets. Enable with one click in EKS.

GPU optimized

Detects accelerated instance driver failures and automatically reschedules workloads away from impaired GPUs.



Available Now with auto mode!

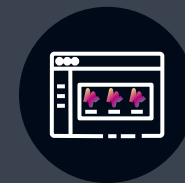
EKS Auto Mode key capabilities



Easier and faster
to get started



Access to all EC2
instances types



Fully managed
core cluster
capabilities



Secure
by default



Automatically
upgraded



Run any
Kubernetes
workload

Accelerate AIML with Amazon EKS



Autonomous Vehicles



Generative AI Models



Robotics



Inference at Scale

EKS

Open Source ML at Cloud Scale



Node Health and Auto-Repair

Accelerated AMIs

Capacity Block Reservations

Container Insights for Accelerated Instances



EFA K8s Device Plugin

EC2 UltraServers

S3 Mountpoint CSI Driver

Native OSS Frameworks



Control Plane Parameter Tuning

Optimized Compute



Node Health and Auto-Repair



Accelerated AMIs

EC2 UltraServers

Capacity Block Reservations



Optimized Compute



Node Health and Auto-Repair



Accelerated AMIs

EC2 UltraServers

Capacity Block Reservations



Optimized Compute



Node Health and Auto-Repair



Accelerated AMIs

EC2 UltraServers

Capacity Block Reservations



Optimized Compute



Node Health and Auto-Repair



Accelerated AMIs

EC2 UltraServers

Capacity Block Reservations



Accelerate Data Management



S3 Mountpoint CSI Driver

EFA K8s Device Plugin



Accelerate Data Management



S3 Mountpoint CSI Driver

EFA K8s Device Plugin



Streamline Kubernetes ML



Native OSS Frameworks



Container Insights for Accelerated Instances



Streamline Kubernetes ML



Native OSS Frameworks



Container Insights for Accelerated Instances



snowflake

Hyungtae Kim

principal software engineer



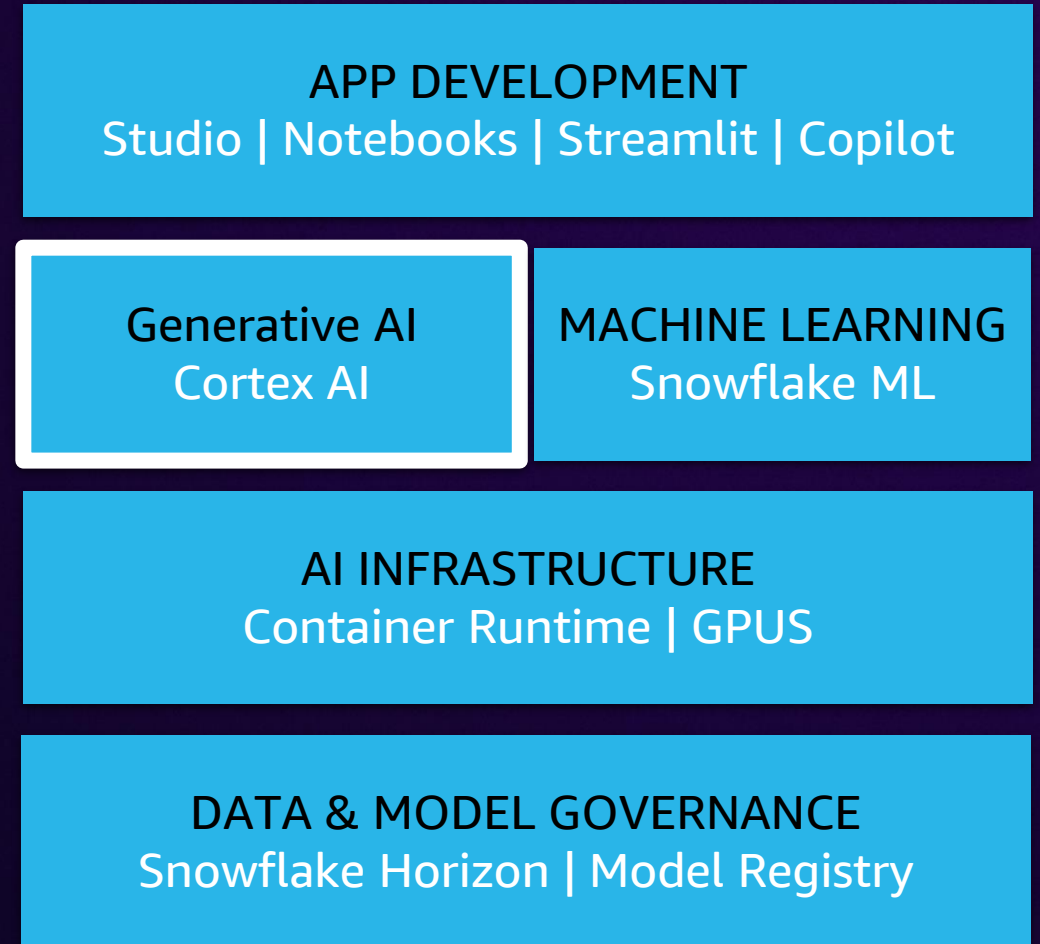
Snowflake: Cortex AI

Inference

- Enterprise-scale batch processing
- Real-time interactive systems

Training

- Arctic models
- Fine-tuning



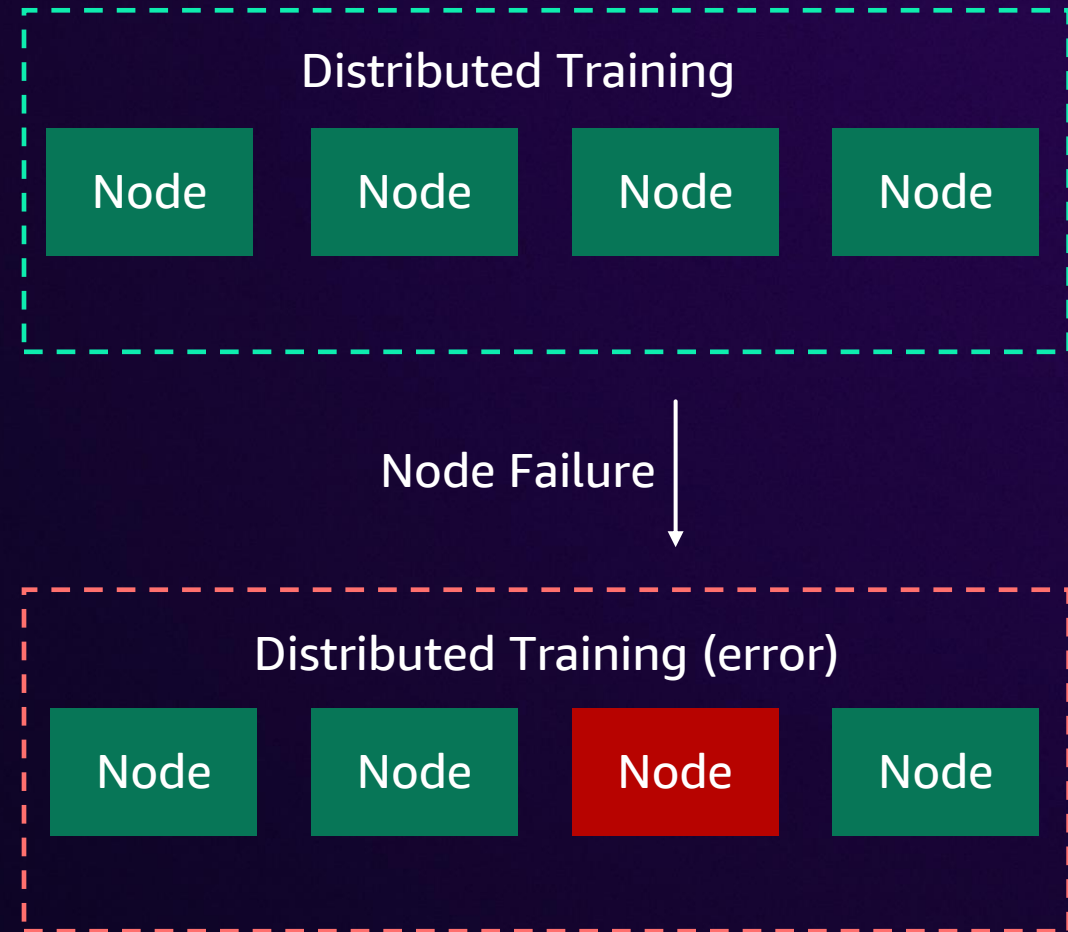
Challenges

GPU capacity

- Scarcity and high costs
- Limited scaling flexibility
- Complex cluster upgrades

System fragility

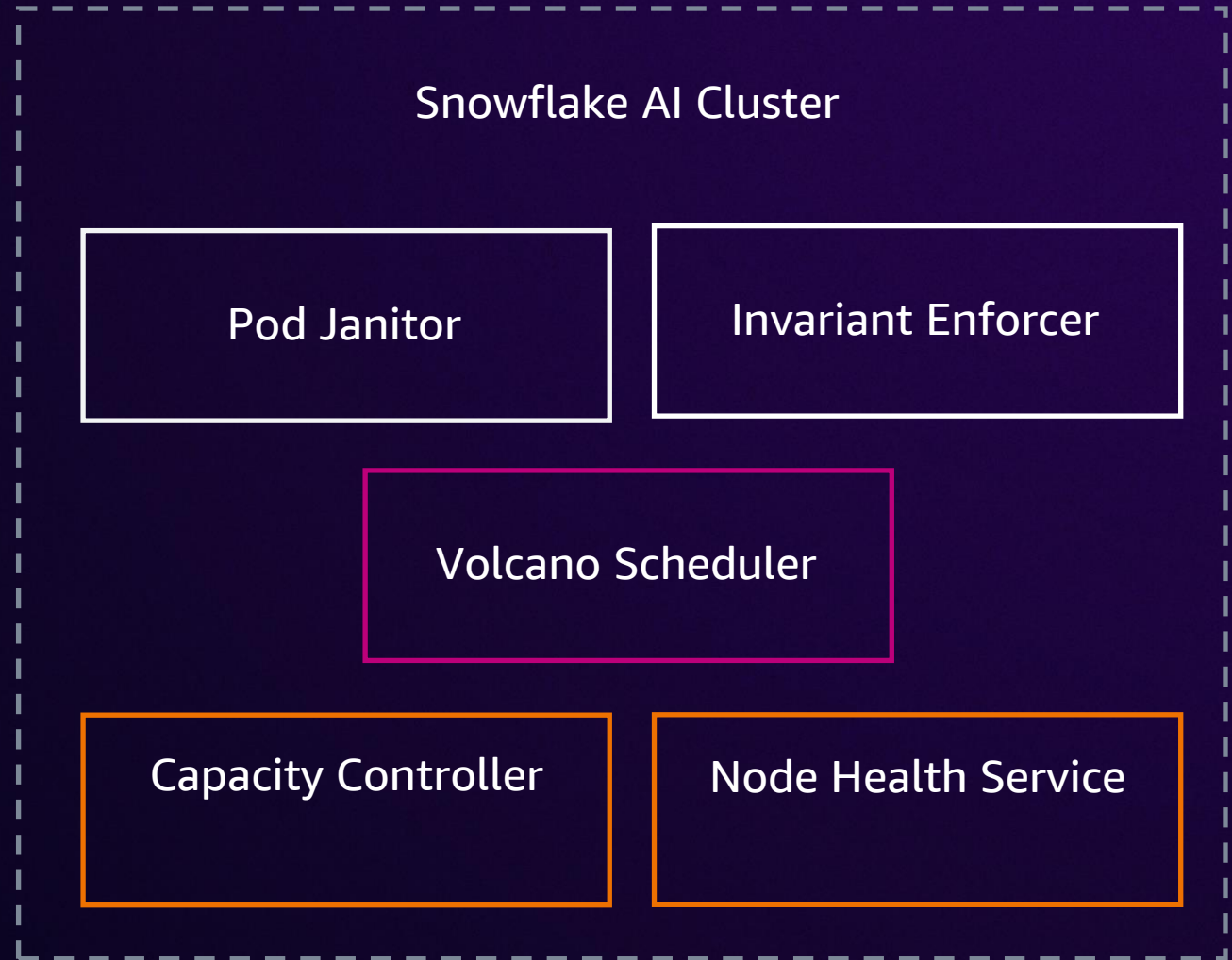
- Higher GPU failure rates
- All-or-nothing requirements



Snowflake AI cluster

Single cluster per region

Powers all AI workloads



Capacity controller

Optimizes GPU allocation across workload types

Capacity Buckets (CRD)

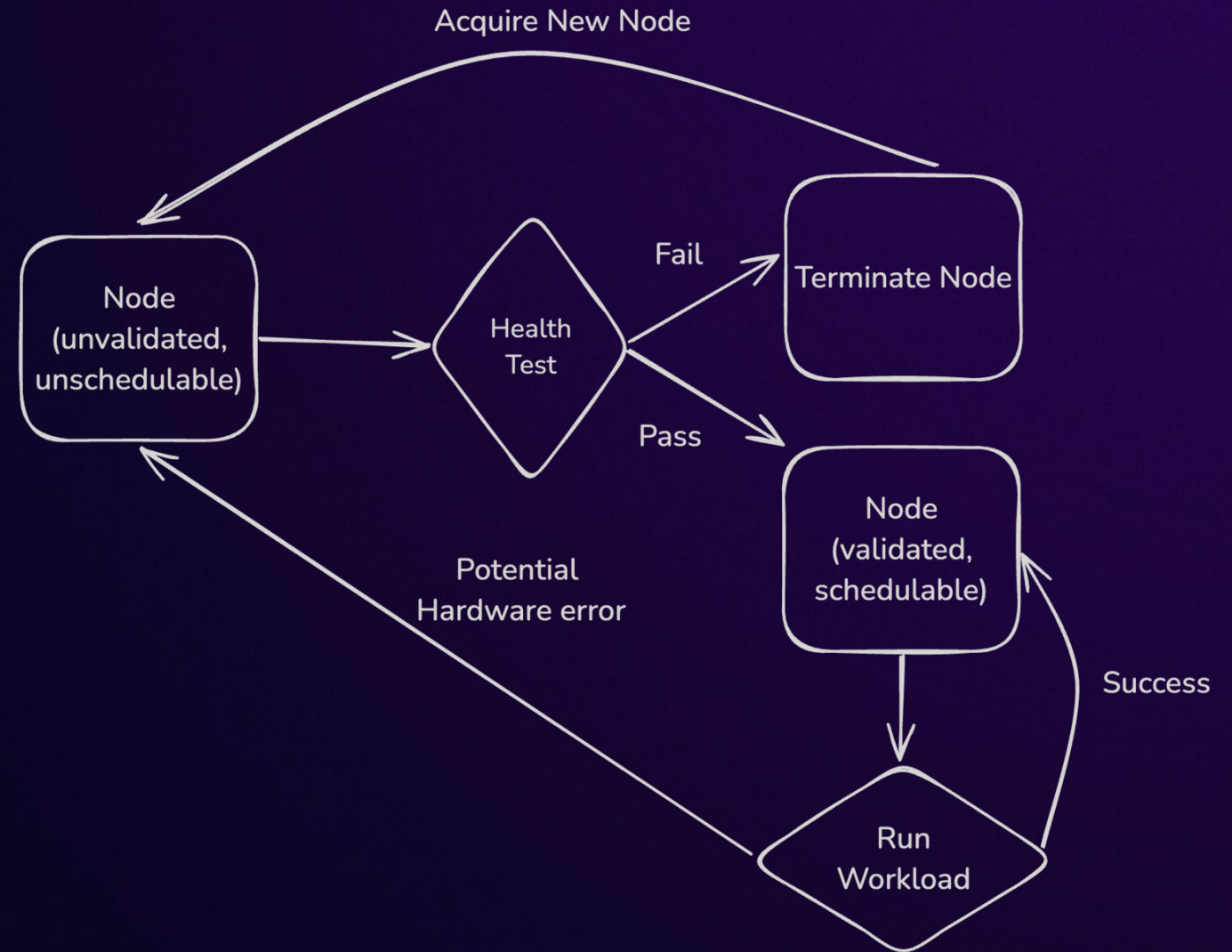
- Defines workload requirements
- Sets resource boundaries
- Controls node allocation

```
apiVersion: mlscheduler.snowflake.com/v1beta1
kind: CapacityBucket
metadata:
  name: cortex-inference
spec:
  capacityPriorityClass: high
  ownersNamespaces:
    - cortex-inference
  request:
    - nodeType: p5.48xlarge
      min: 3
      max: 200
    - nodeType: p4d.24xlarge
      min: 3
      max: 200
```

Node health service

Proactive node validation

- GPU
- Networking
- Performance
- Storage



Benefits of EKS

Performance & Compatibility

EFA Networking

Accelerated AMIs

NVIDIA NGC
Support

Storage Tiers

FSx for Lustre

EBS & EFS

S3

Performance & Compatibility

Node Remediation

Auto-scaling

Simplified
Management

Initial Challenges

Capacity

- ✗ Limited GPU capacity
- ✗ Complex upgrades

Fragility

- ✗ Hardware failures
- ✗ All-or-nothing jobs

Transform



Wins

k8s native AI

- ✓ Dynamic auto-scaling
- ✓ Upgrades
- ✓ Automated resilience

Improve utilization

- ✓ 30% higher utilization
- ✓ Resource sharing
- ✓ Less operational burden

Key lessons learned

- ✓ **Design for resilience**
Embrace impermanence · Optimize for recovery
- ✓ **Hardware validation**
Proactively validate · Reserve good nodes
- ✓ **Automate recovery**
Actively monitor and remediate cluster state
- ✓ **Constantly iterate**
Continuous improvement · Snowflake AI cluster is on its 8th gen

The future of Amazon EKS





App modernization

.NET apps

Legacy homegrown
Linux apps

Monoliths



AI/ML

Autonomous vehicles

Generative AI

Robotics

Modeling, training,
and inference



Data processing

Real time

MapReduce

Batch



Backends

Apps and services

Mobile

IoT



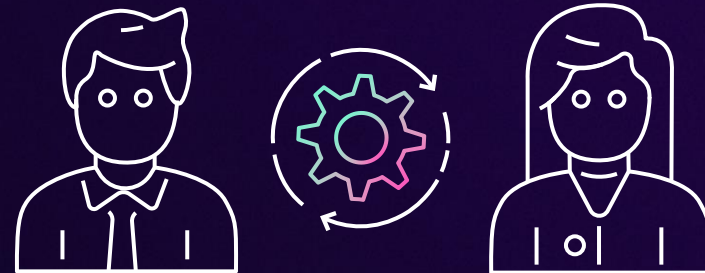
Web applications

Static websites

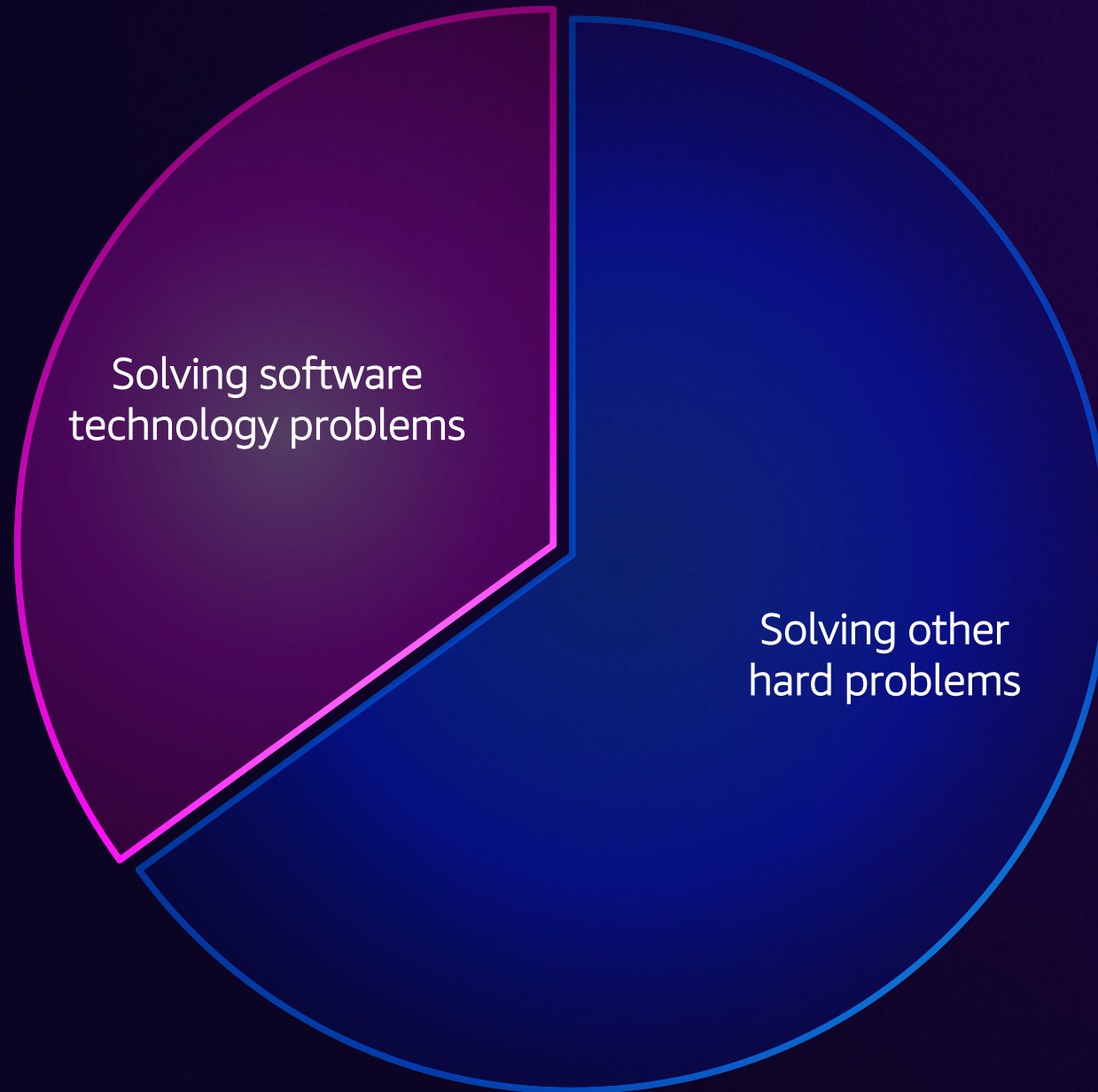
Complex web apps



Solving software
technology problems



Solving other
hard problems

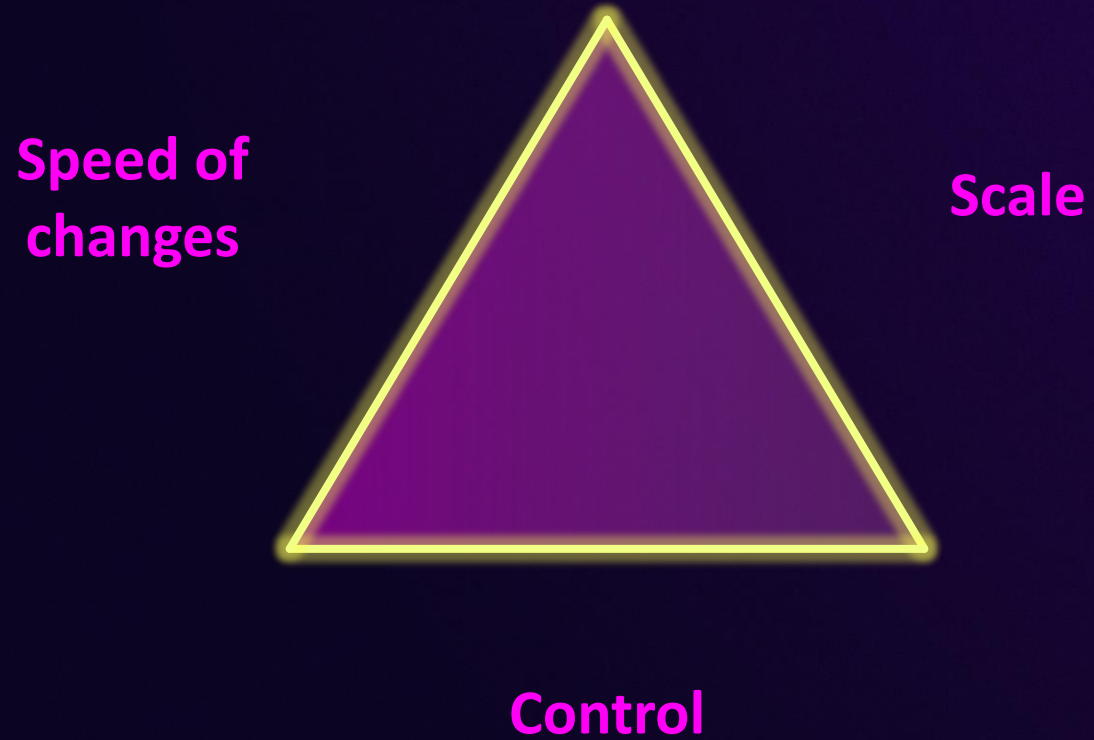


1. How to use technology without becoming a technology company.

1. How to use technology without becoming a technology company.
2. The future is here, it's just not evenly distributed.

1. How to use technology without becoming a technology company.
2. The future is here, it's just not evenly distributed.
3. Open source software can become expensive at scale.

Systems Scaling Trilema



AWS + Open Standards

let us go faster together

Accelerate time to value

Democratize Innovation

Lower the cost of entry

Turn CAPEX into OPEX

Evolution of EKS

2018

2019

2020

2021

2022



Managed Control Plane

Managed Data Plane

Managed Operational Tools



Evolution of EKS

2018 2019 2020 2021 2022 2023 2024 2025 Beyond



Managed Clusters

Integrated Hybrid

Managed Platform Components

Integrated Developer Experience



Investment Priorities for the next 3 years

1. Provide optimized experiences for critical workload patterns at any scale
2. Deepen AWS integrations and tooling for management and efficiency
3. Meet your workloads where they are
4. Simplify platform building
5. Accelerate the flywheel of innovation in the community and seamlessly bring that innovation to customers

Goals for **Customers**

Automate more things in and alongside the cluster

Natively bring you the latest AWS innovations through Kubernetes

Ensure compatibility with and support of community projects which make Kubernetes innovative and powerful



Goals for **Partners**

Make it easier to build on EKS for your products and services

Provide simple paths to enable EKS customers and sell with AWS

Provide ongoing guidance, support, and ideas to improve your product and our partnership






“ You’ve got to start with the customer experience and work backward to the technology.

Steve Jobs





“ We innovate by starting with the customer and working backwards.

Jeff Bezos



Public Roadmap

- Stay up to date with what we're working on.
- Give us feedback and propose ideas.
- Get notified when new features ship.

github.com/aws/containers-roadmap



Resources



docs.aws.amazon.com/eks/latest/best-practices

EKS Best Practices Guide

Deep dive into advanced best practices
Regularly updated and curated by AWS experts.

New! Best practices for machine learning



eksworkshop.com

EKS Workshop

Free and open training for using EKS.
Modules from 200 – 400 level

New! Developer workshop



aws-ia.github.io/terraform-aws-eks-blueprints
aws-quickstart.github.io/cdk-eks-blueprints/

EKS Blueprints

Frameworks and examples for deploying complete clusters.
Available for Terraform and AWS CDK

Thank you!

Nathan Taber

 [linkedin.com/in/natetaber](https://www.linkedin.com/in/natetaber)

Hyungtae Kim

 [linkedin.com/in/hyungtaekim](https://www.linkedin.com/in/hyungtaekim)



Please complete the session survey in the mobile app