

Optimización de costos para cargas de trabajo con Amazon **EKS**

Iago Banov
Sr. Specialist SA - Containers

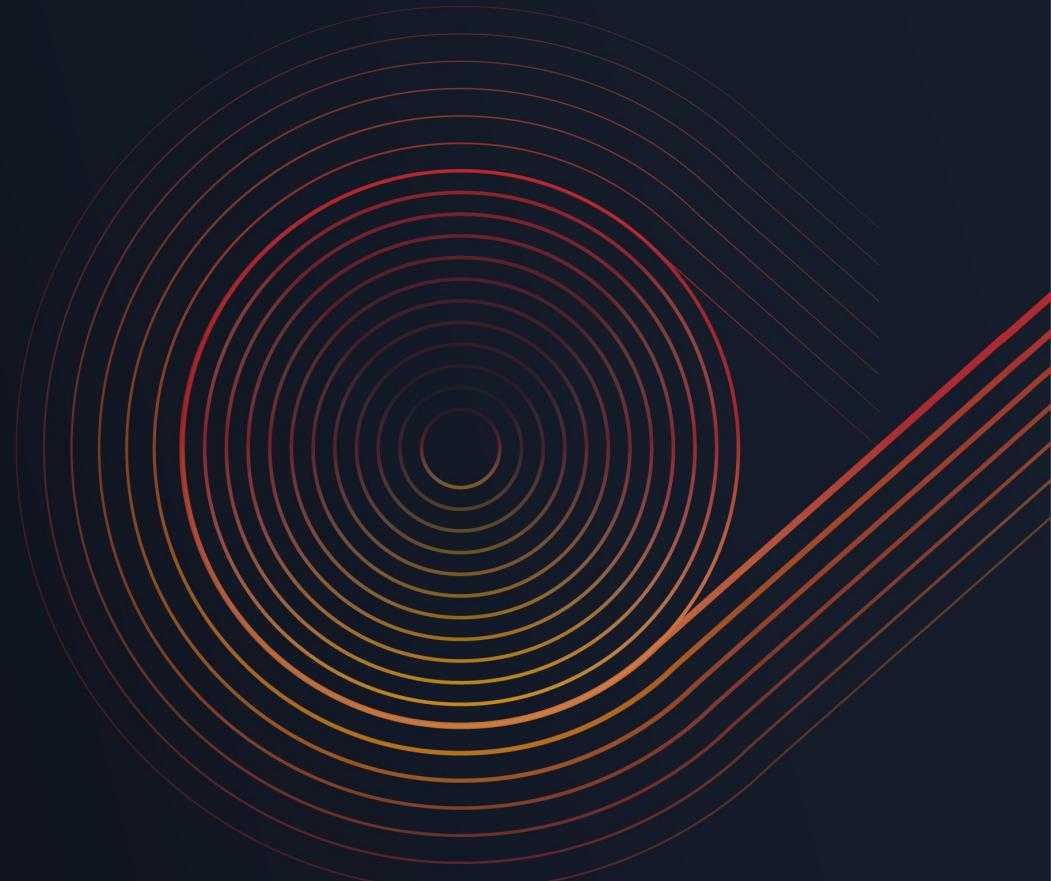
Si no puedes medir, no puedes gestionar.

- Peter Drucker

Dimensiones de costes



Visibilidad de costos: conceptos introductionios

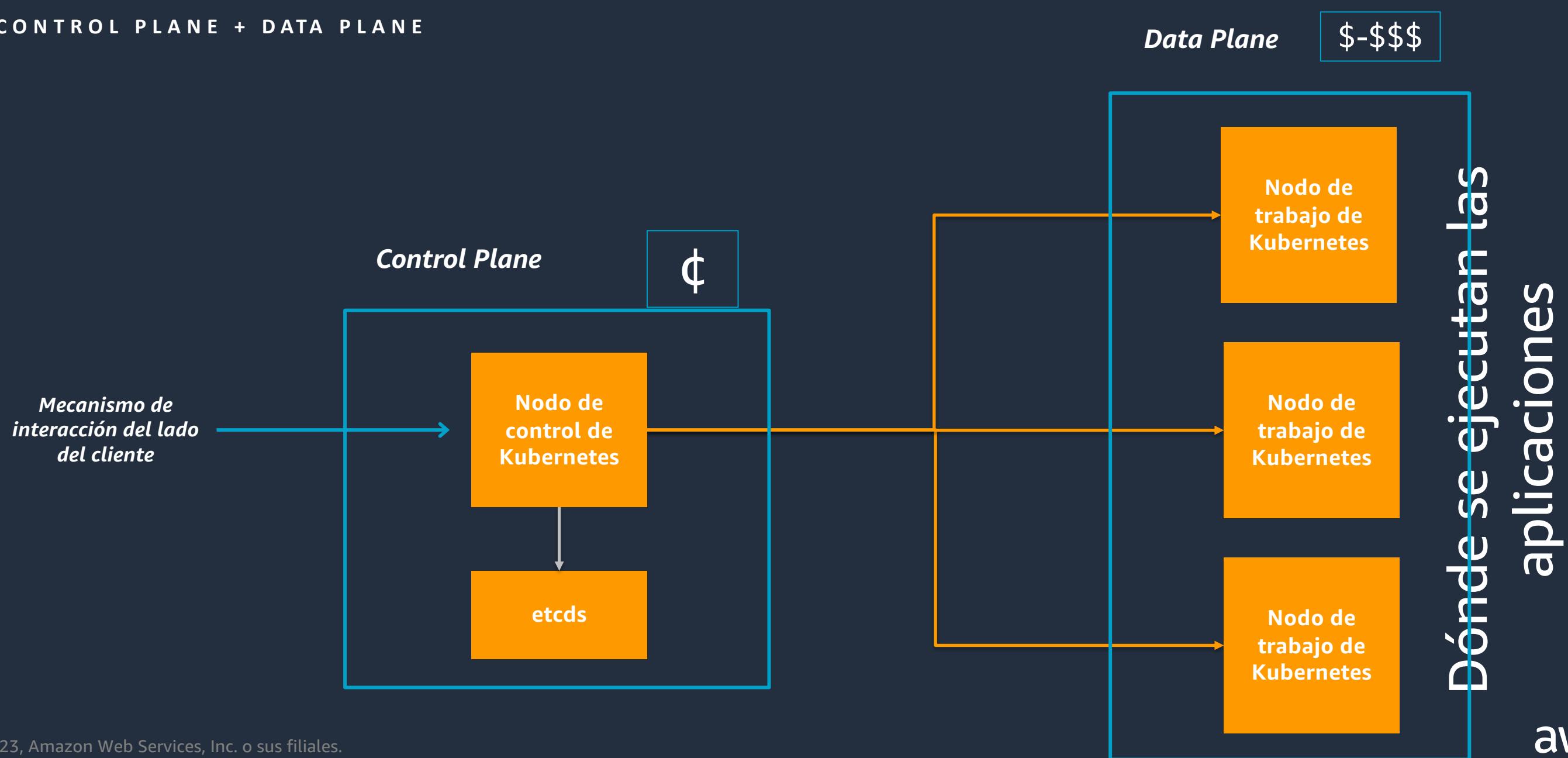


Arquitectura de Kubernetes de alto nivel

CONTROL PLANE + DATA PLANE

Data Plane

\$-\$-\$



Granularidad de los costes

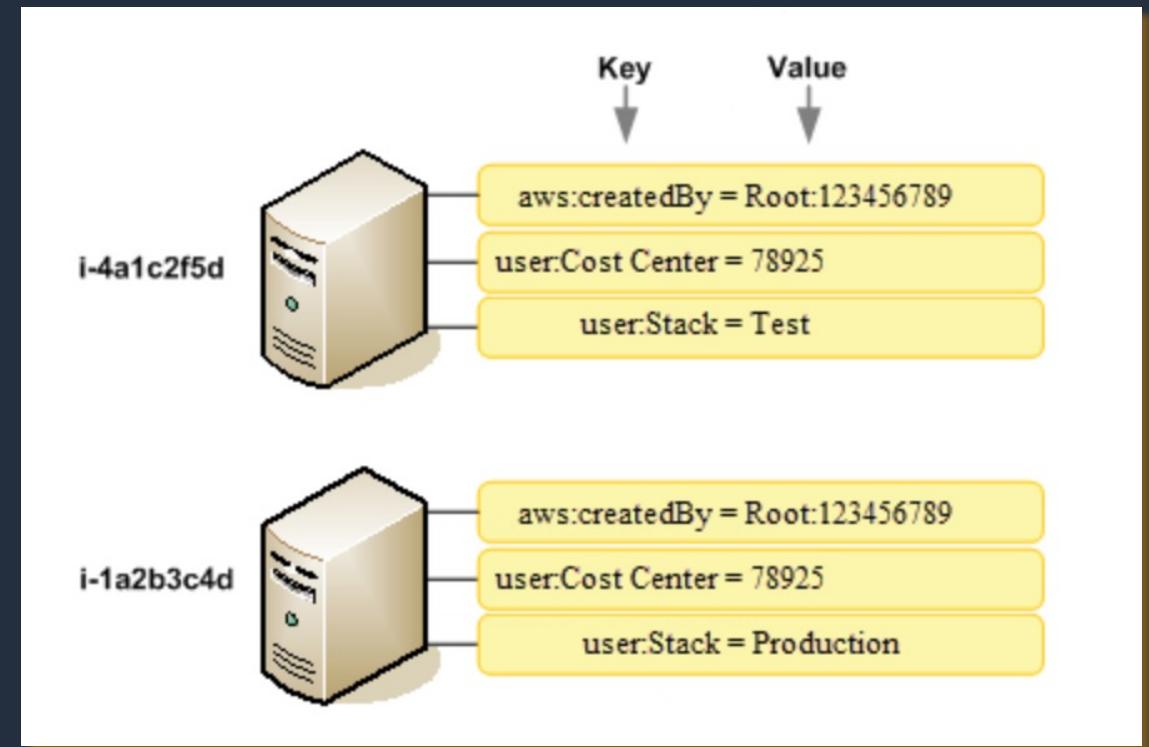


Visibilidad de costos con AWS

- Uso de etiquetas
- Explorador de costos
- Informes de uso y costos de AWS

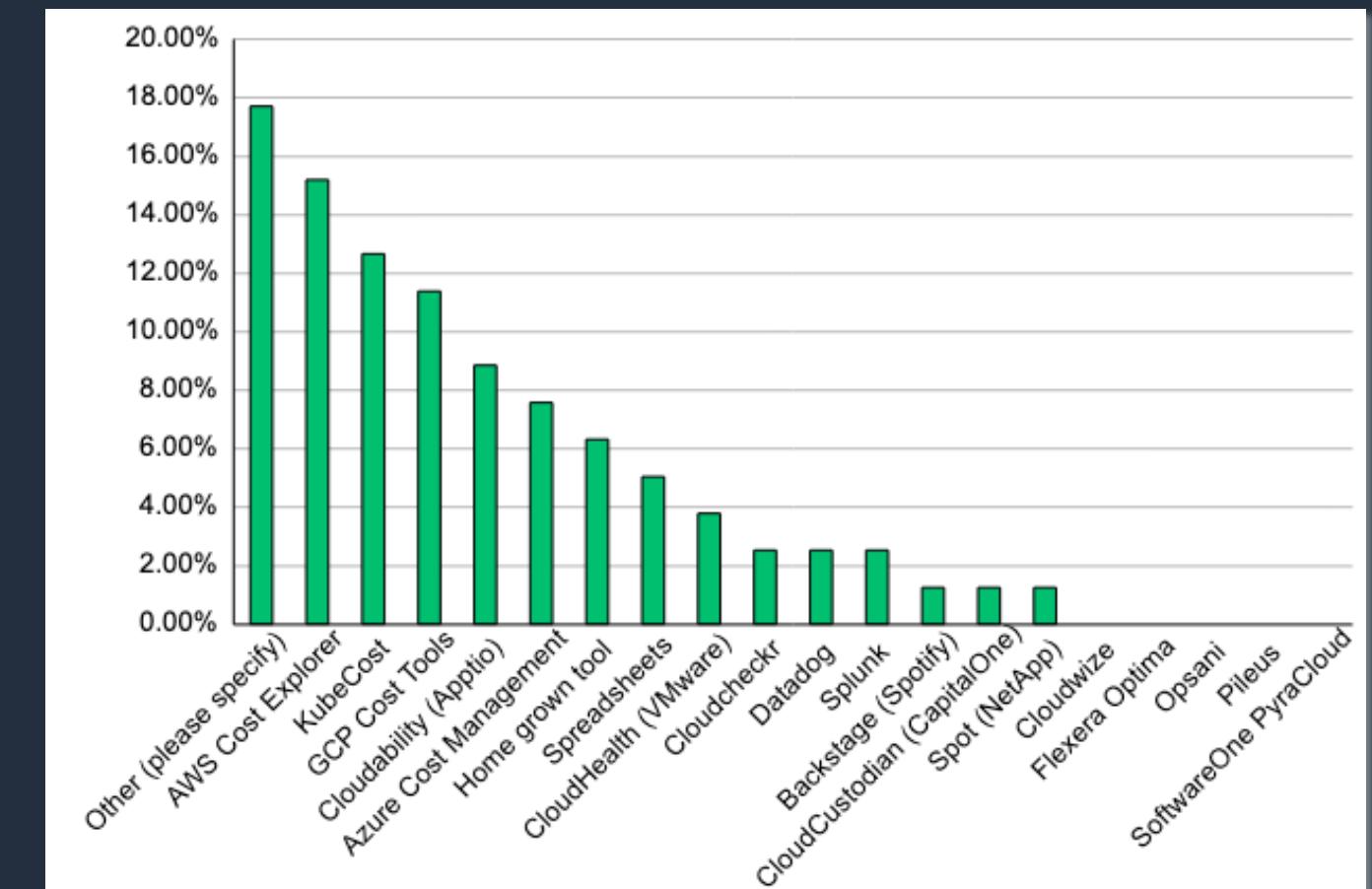
NOVEDAD

La tag generada por AWS, **aws:eks: cluster-name**, se agrega automáticamente a las instancias de Amazon EC2 que participan en un clúster de Amazon EKS.



Gestión de costos: ecosistema

Socios de
AWS



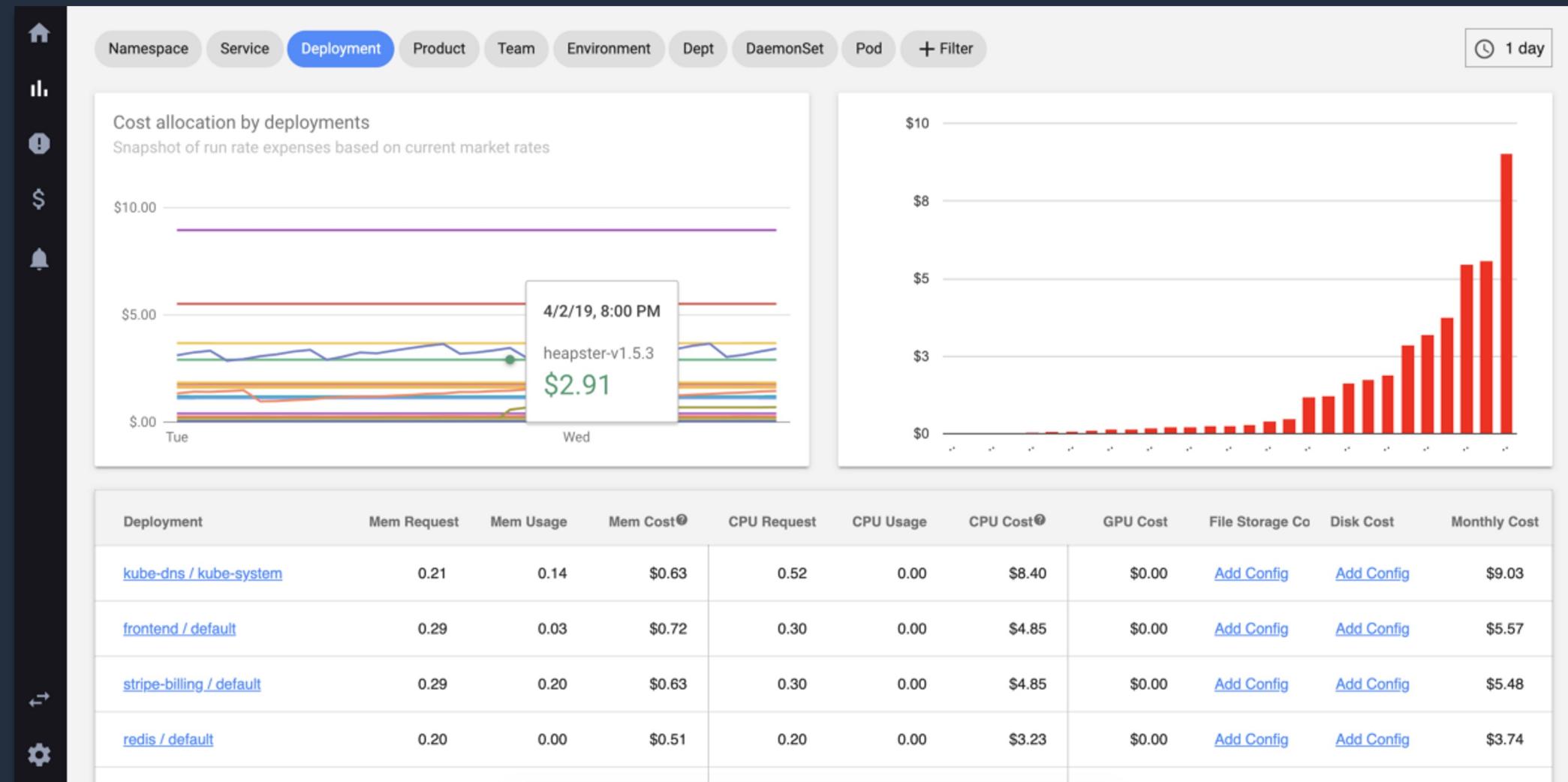
¿Qué utilizan los clientes para supervisar los costos de Kubernetes?
- Encuesta FinOps de CNCF

Kubecost: monitoreo de costos en tiempo real

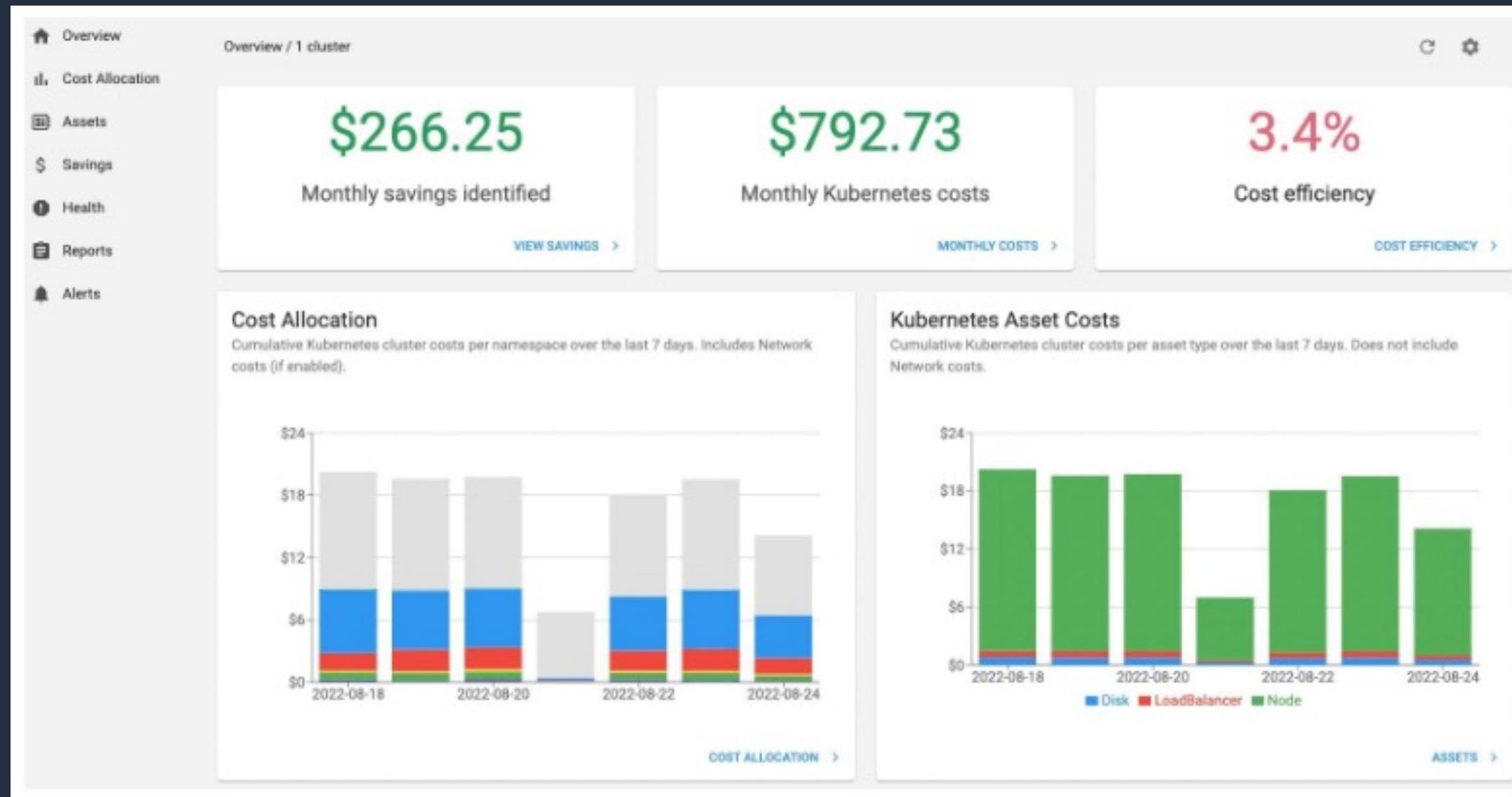


- *Opensource*;
- Evaluación de los costos y el uso de forma granular:
 - Service, Deployment, Namespace, Label, StatefulSets, DaemonSet, Pod y Container.
- Atribución a conceptos organizacionales: equipos o centro de costos;
- Métricas de Prometheus para determinar el uso por parte de las aplicaciones;
- Hace recomendaciones sobre la optimización de los recursos.

Kubecost: visibilidad de costos en tiempo real



Kubecost: *información* para la optimización

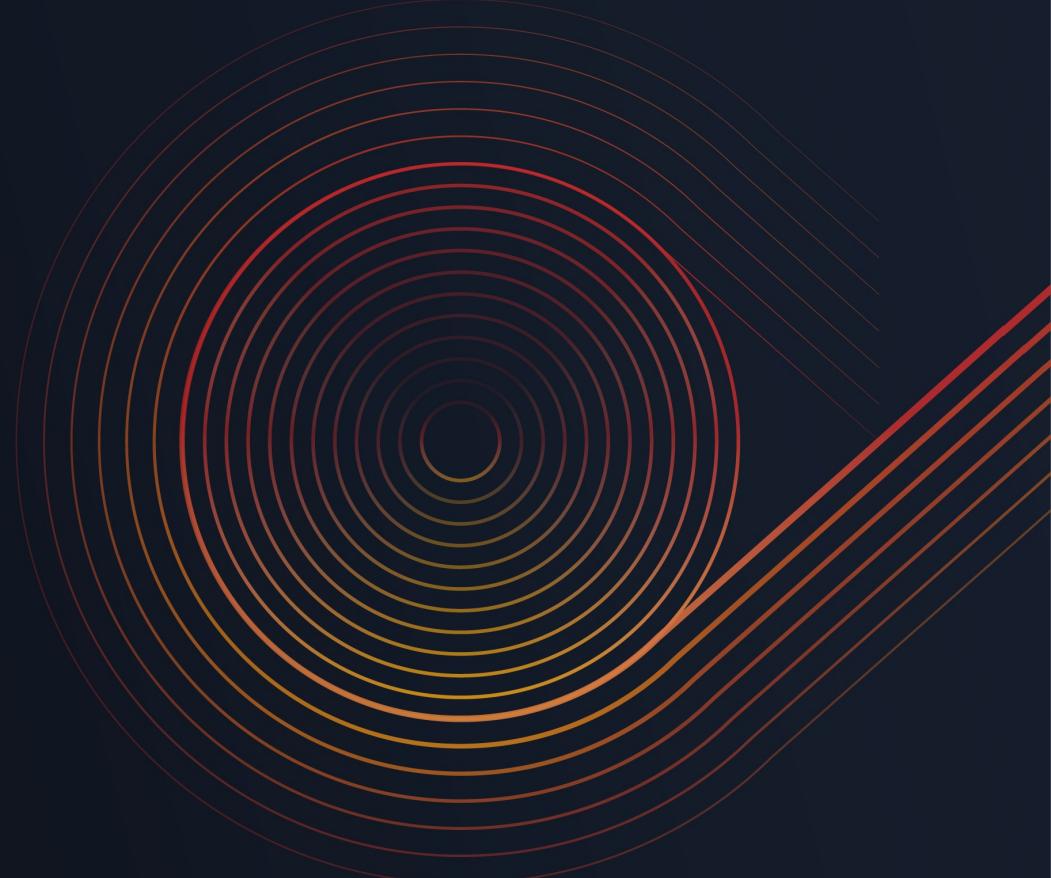


Mejores prácticas para la visibilidad de los costos

- Uso de etiquetas en los recursos de AWS para aumentar la visibilidad en las herramientas nativas;
- Aproveche las herramientas del ecosistema para comprender sus costos con mayor profundidad;
- Monitoreo continuo de costos;
- Implemente el *showback* o *contracargo para los usuarios internos de la plataforma.*



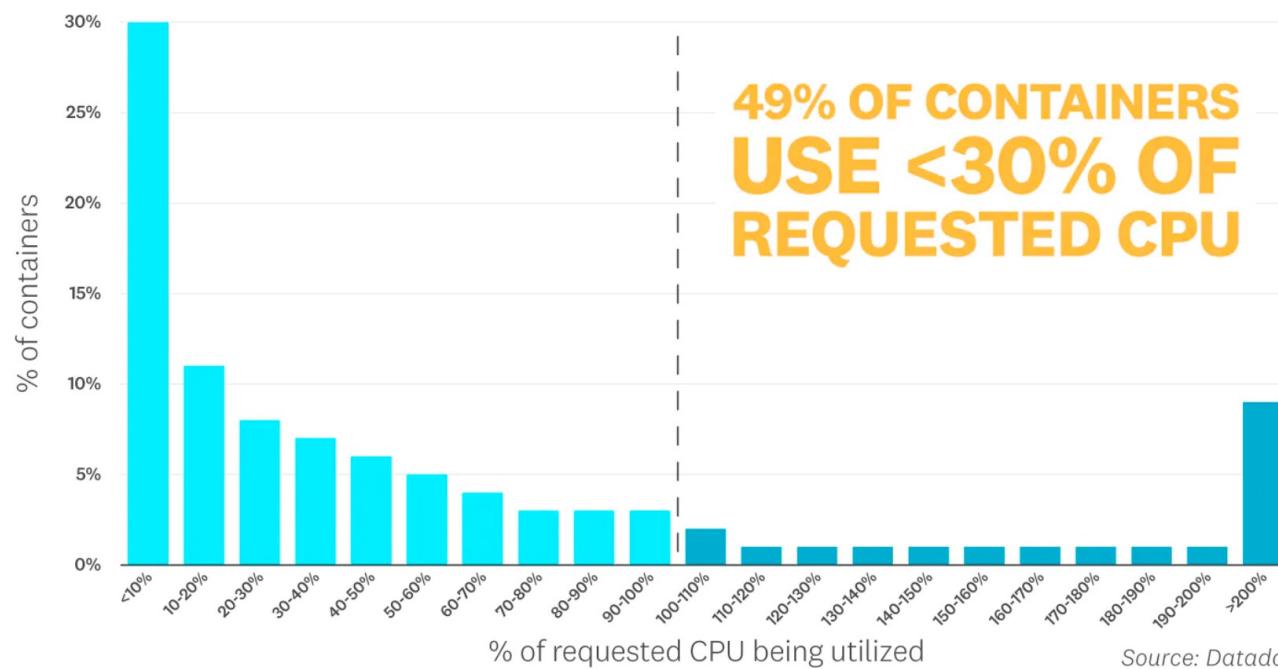
Perfiles de aplicaciones... ¿Hola?



Métricas del uso promedio de contenedores

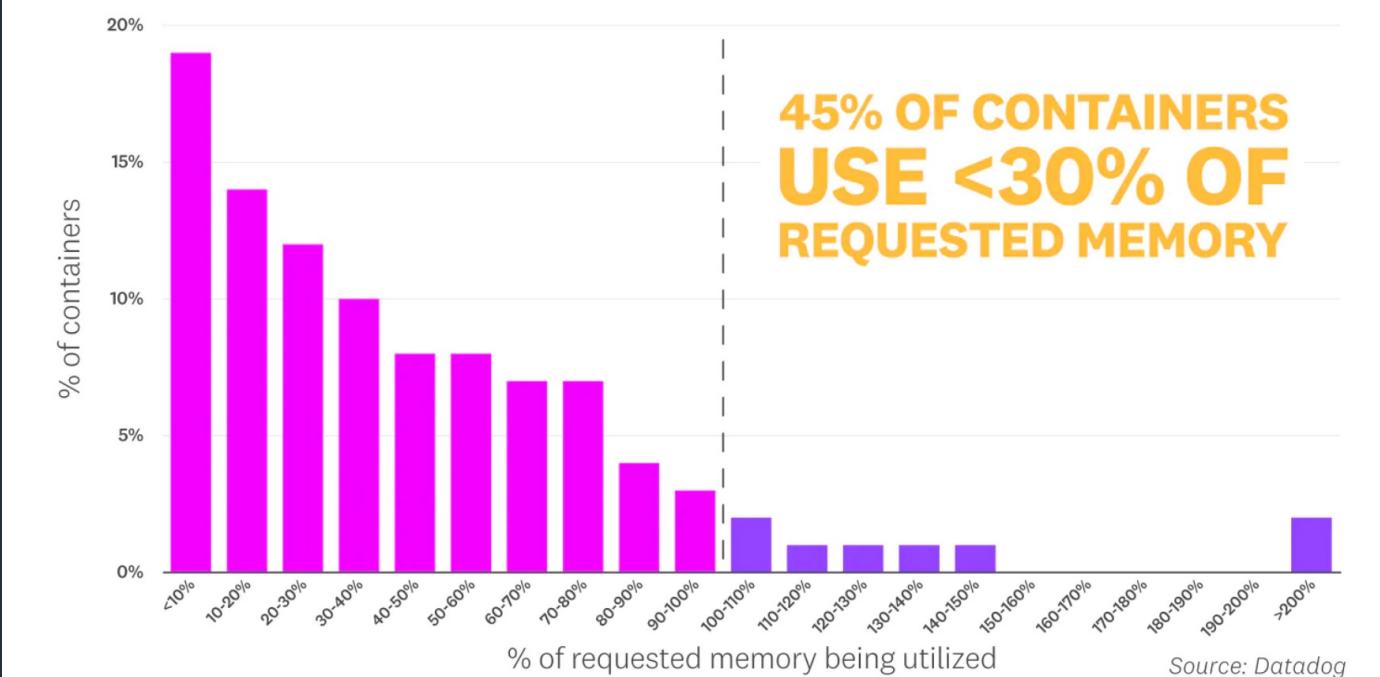
Uso de CPU

Usage of Requested CPU



Uso de memoria

Usage of Requested Memory



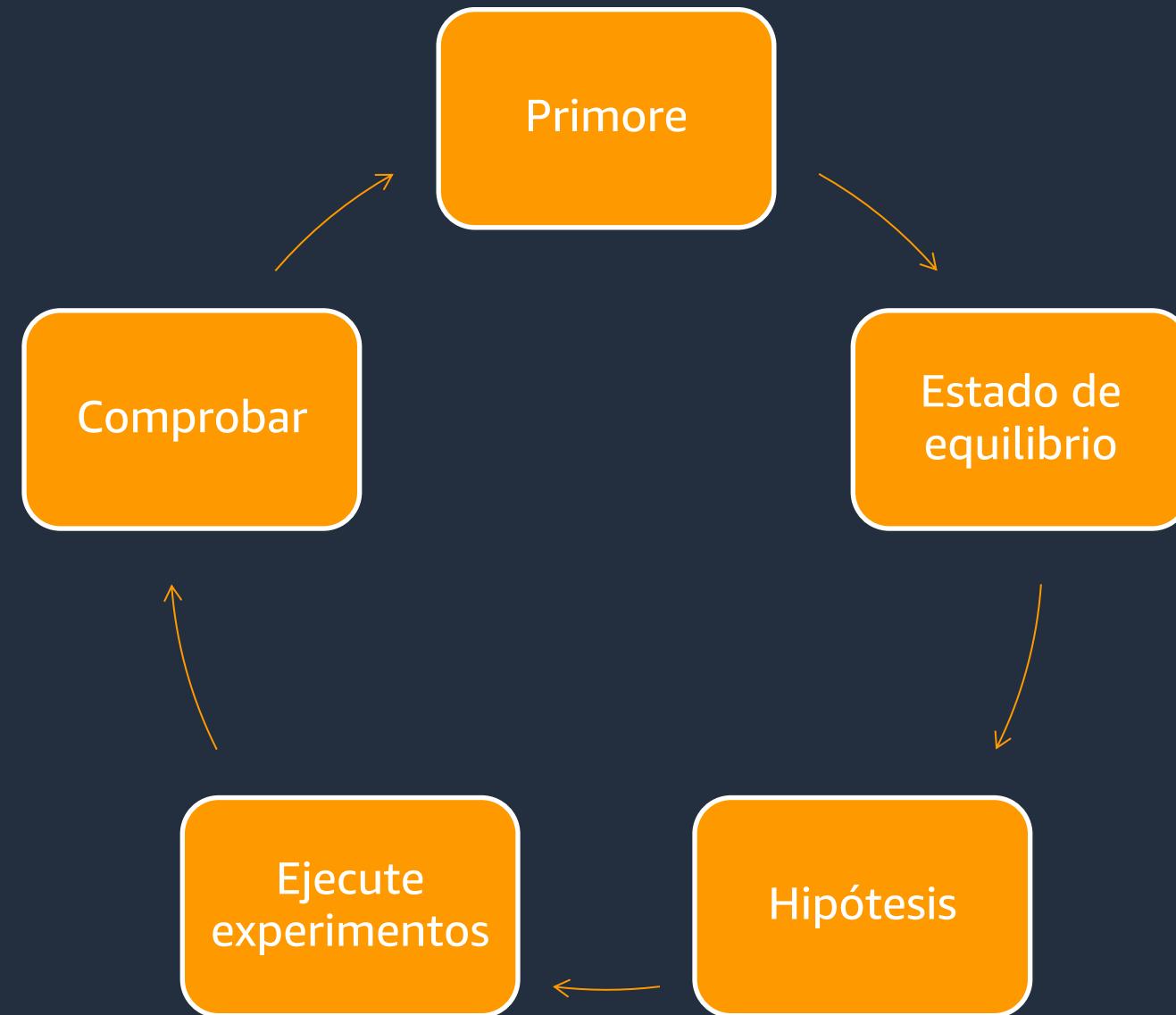
Elaboración de perfiles de aplicaciones

En raras ocasiones, la asignación de CPU y memoria para una *carga de trabajo* será correcta la primera vez;

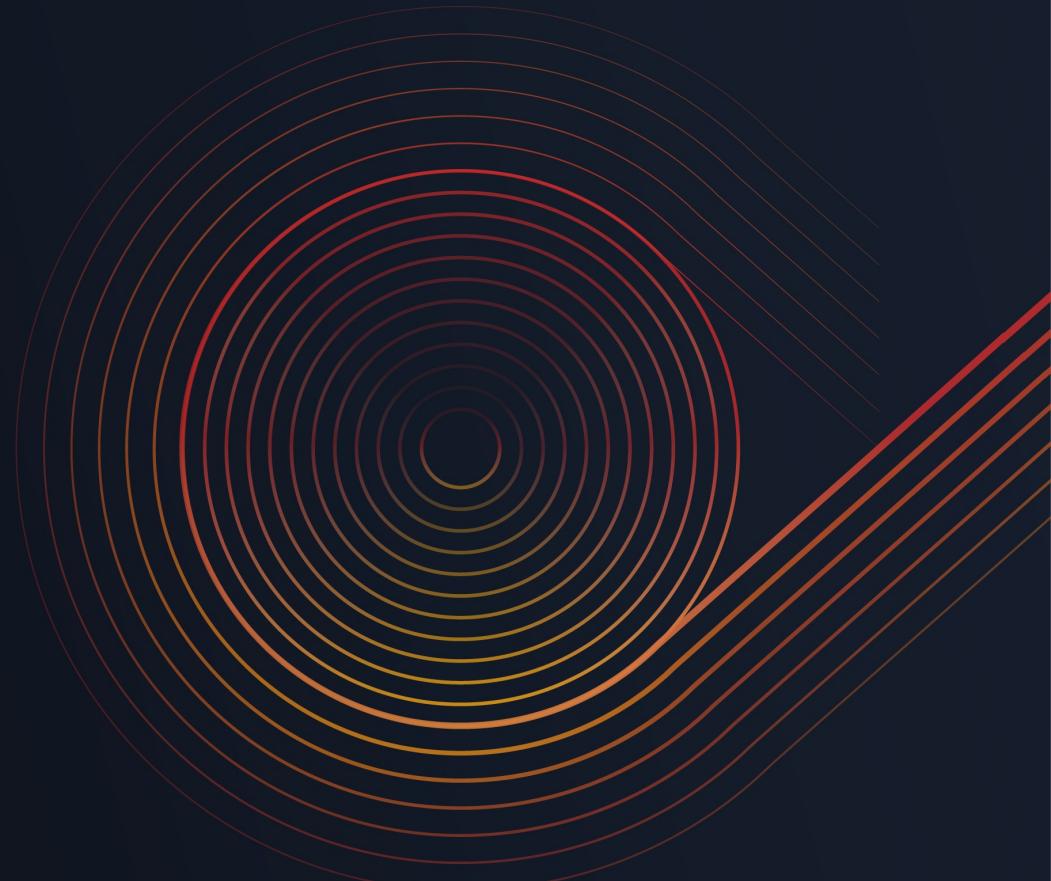
- La *creación de perfiles* del formulario continuo le ayuda a ajustar las asignaciones de recursos para evitar el sobreaprovisionamiento.



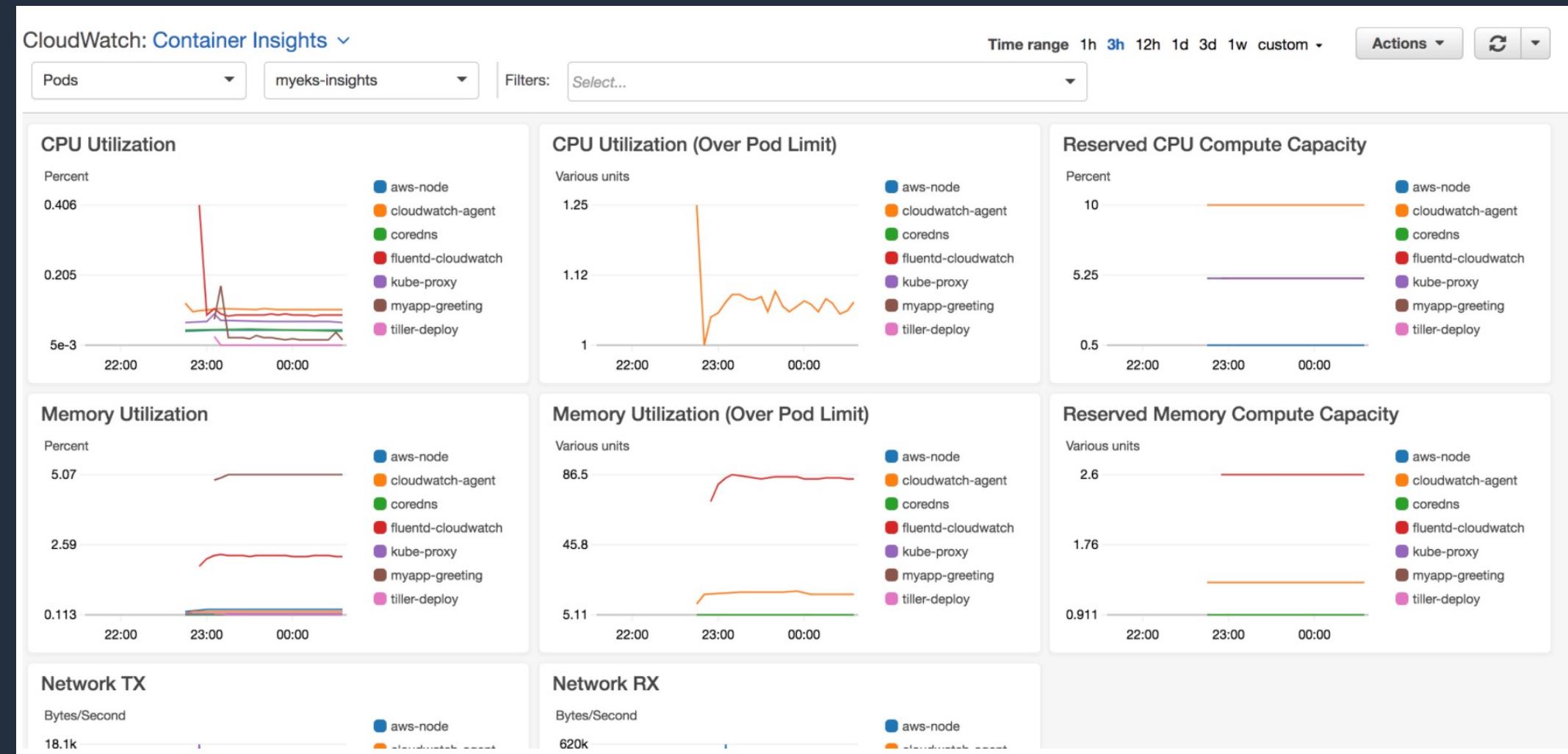
Ajustes continuos



Dimensionamiento correcto de sus aplicaciones



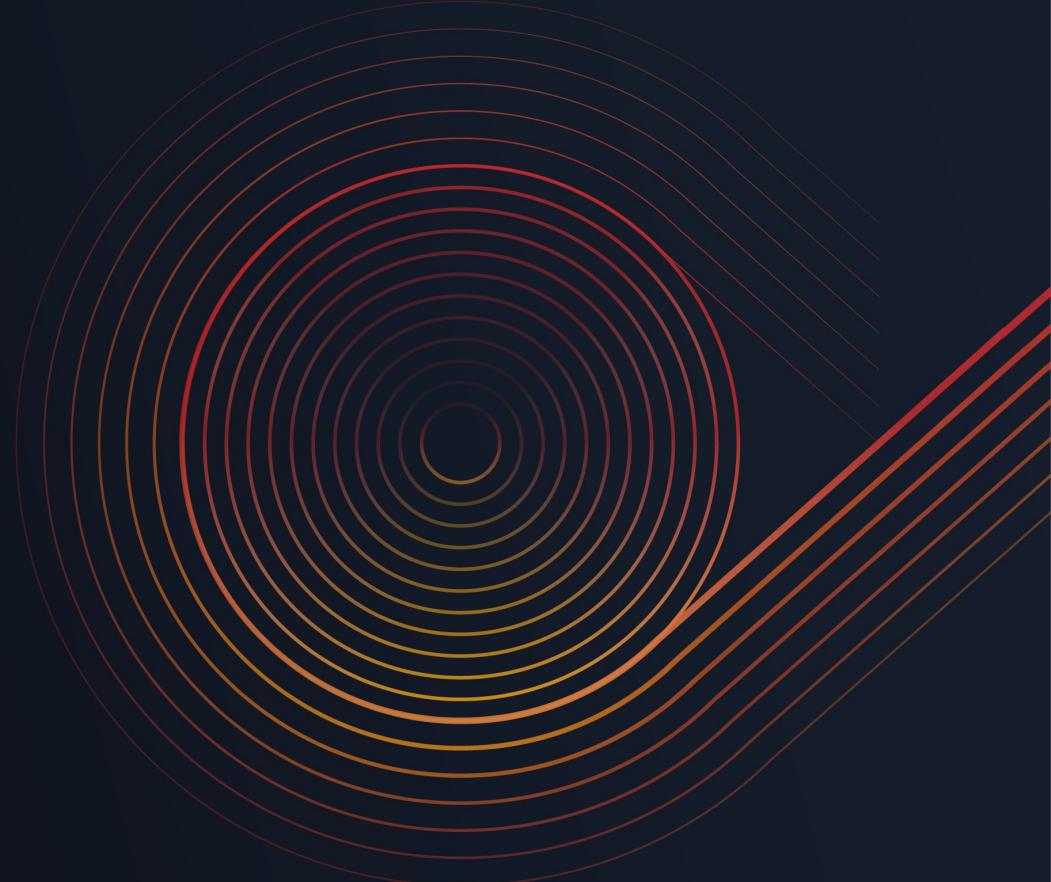
Comprenda el uso de los recursos



Mejores prácticas para la creación de *perfils de aplicaciones*

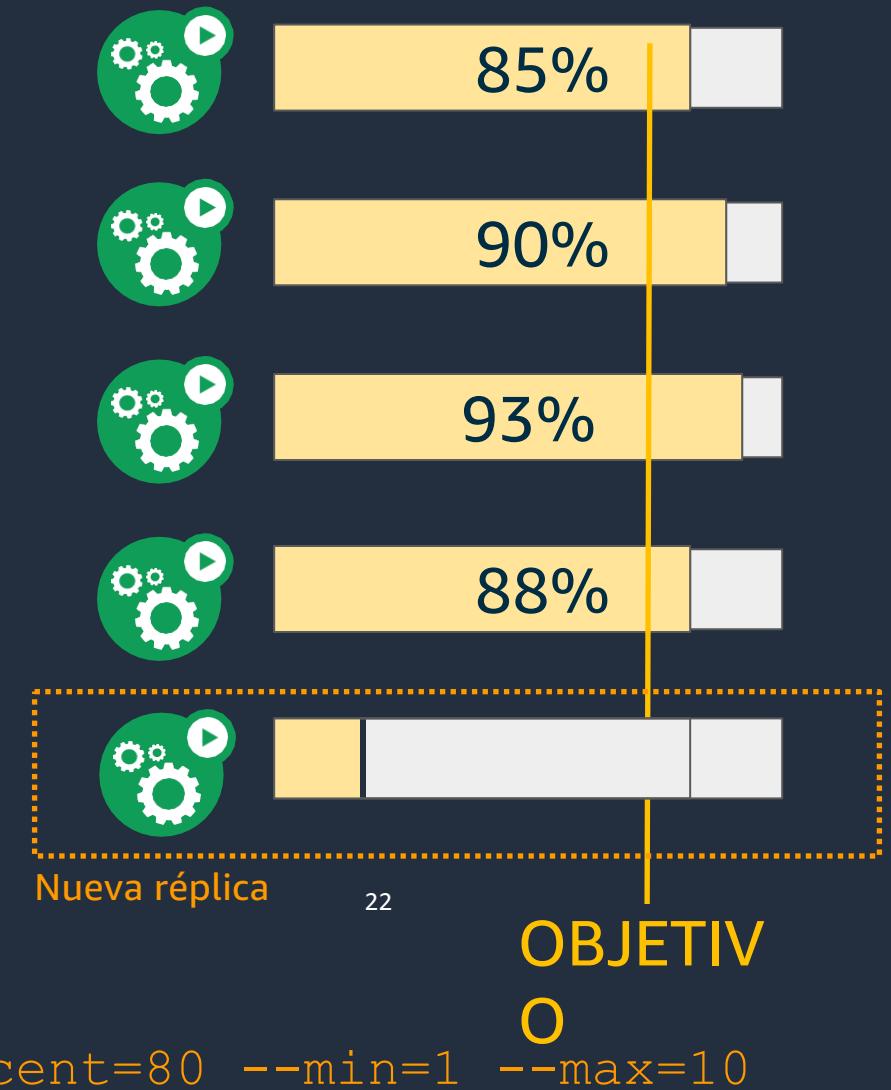
- Identificar los requisitos de recursos mediante la *creación de perfils*;
- Realice estas acciones de forma continua con ajustes;
- Utilice pruebas de carga para determinar los requisitos a escala;
- Implemente *límites* cuando sea necesario.

Aprovechar la elasticidad



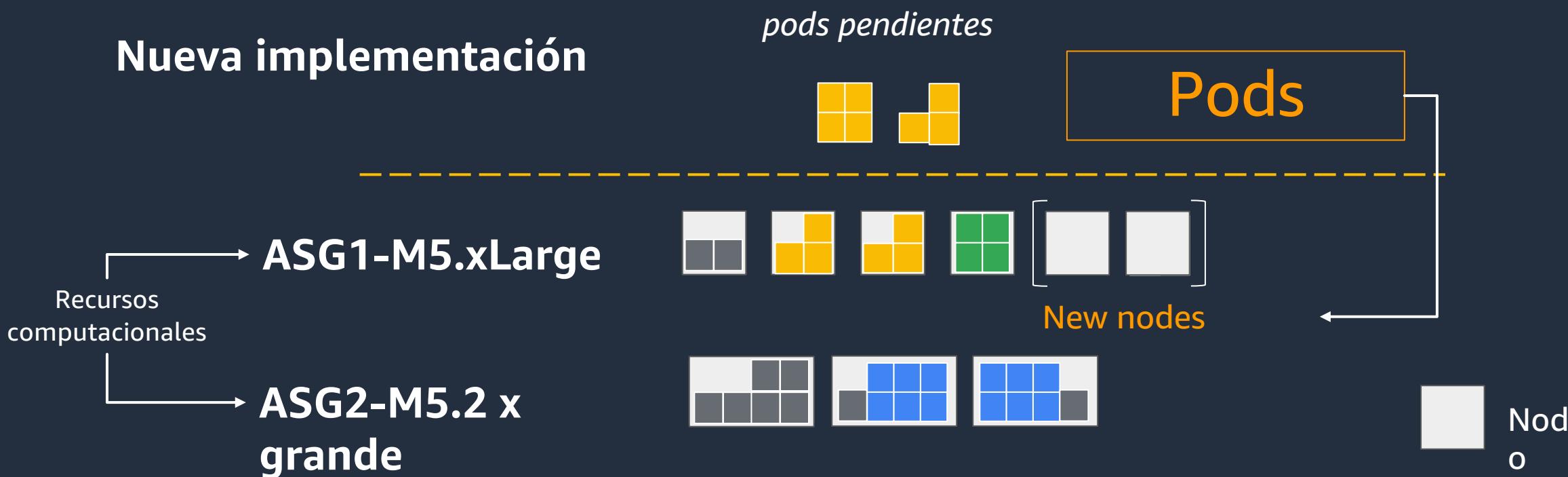
Autoscaler de pods horizontales (HPA)

- Si las *pods* están muy cargadas, poner en marcha nuevas *pods* puede reducir la carga promedio;
- Si los *pods* están inactivos, detenerlos liberará recursos;
- Especifique la **métrica de carga**: uso de CPU, solicitudes HTTP por minuto... ;
- Las métricas pueden estar **relacionadas con recursos, métricas personalizadas o métricas externas** (de una cola de Amazon SQS, por ejemplo).



implementación automática de `kubectl scale --cpu-percent=80 --min=1 --max=10`

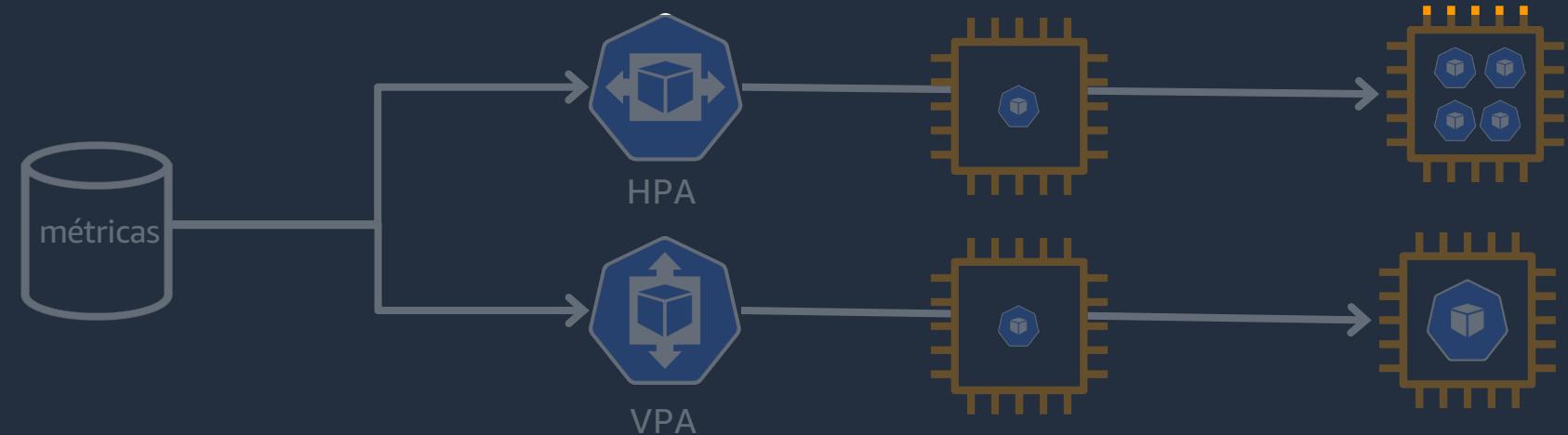
Ejemplo: Cluster Autoscaler Scale-Up



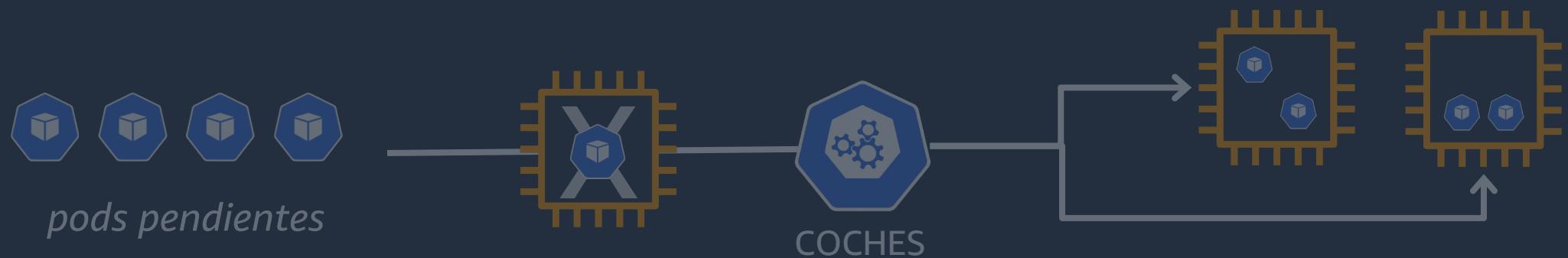
23

autoscaler con Kubernetes

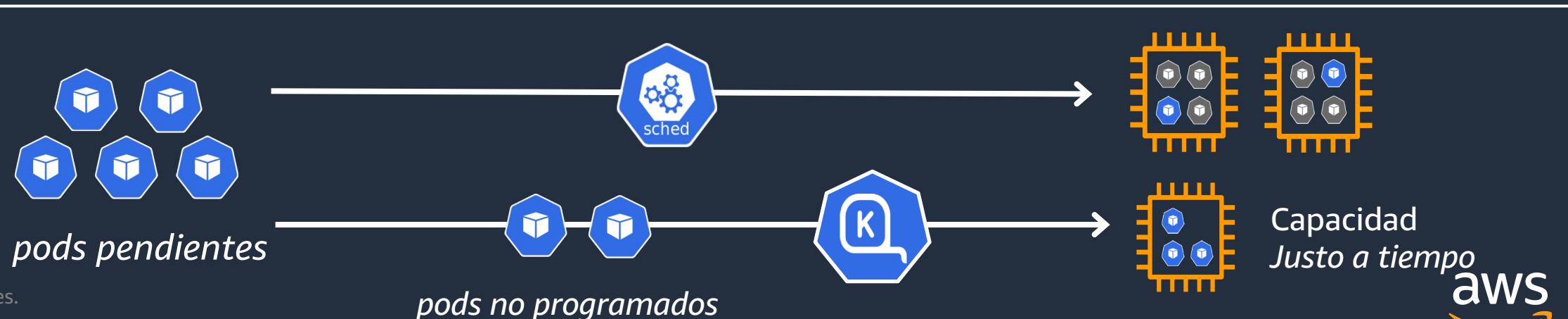
1. autoscaler de pods horizontales (HPA)
2. autoscaler de pods verticales (VPA)



3. Autoscaler de clústeres (CAS)

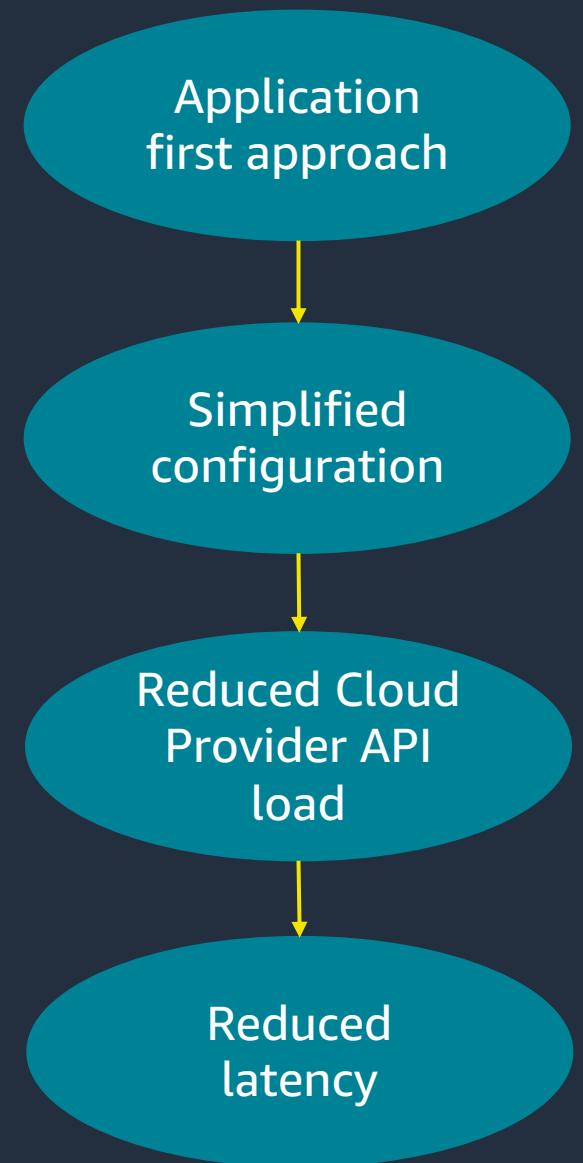


4. Karpenter



Karpenter para aprovisionamiento y autoscaler sin grupos de nodos

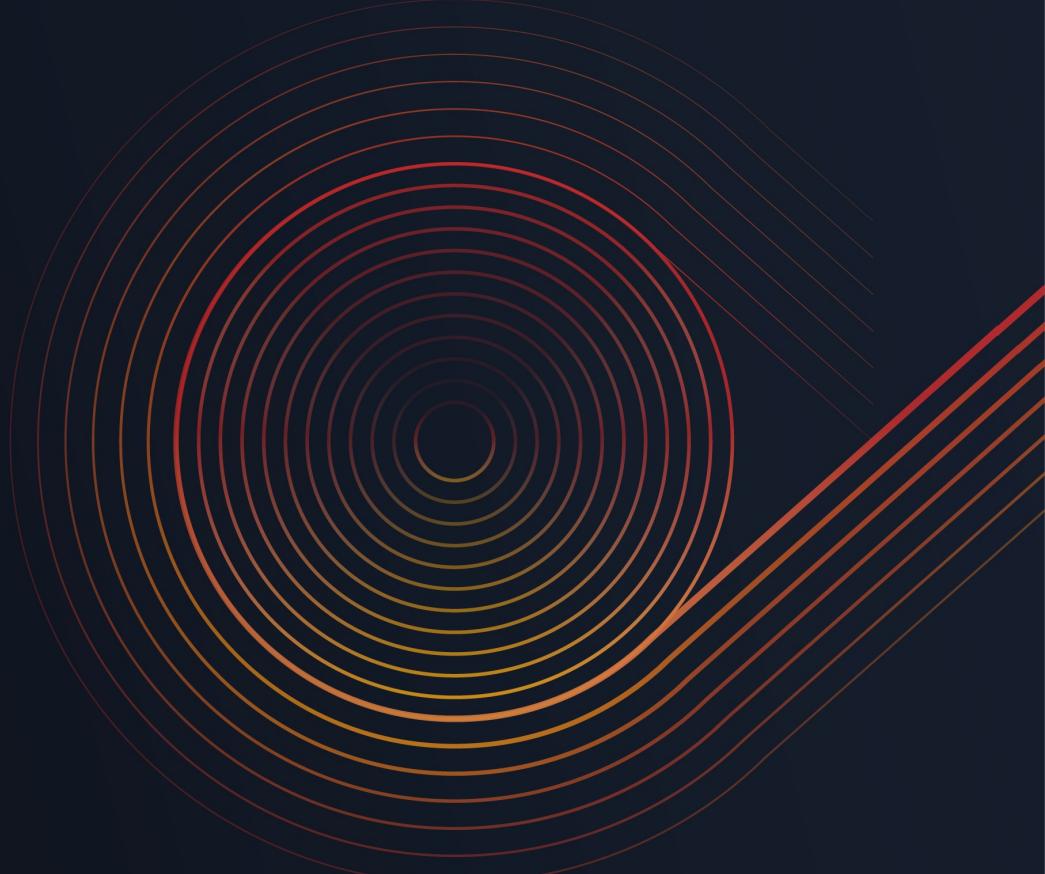
¿Qué pasa si eliminamos el concepto de *grupos de nodos*?



Mejores prácticas para Cluster Autoscaler

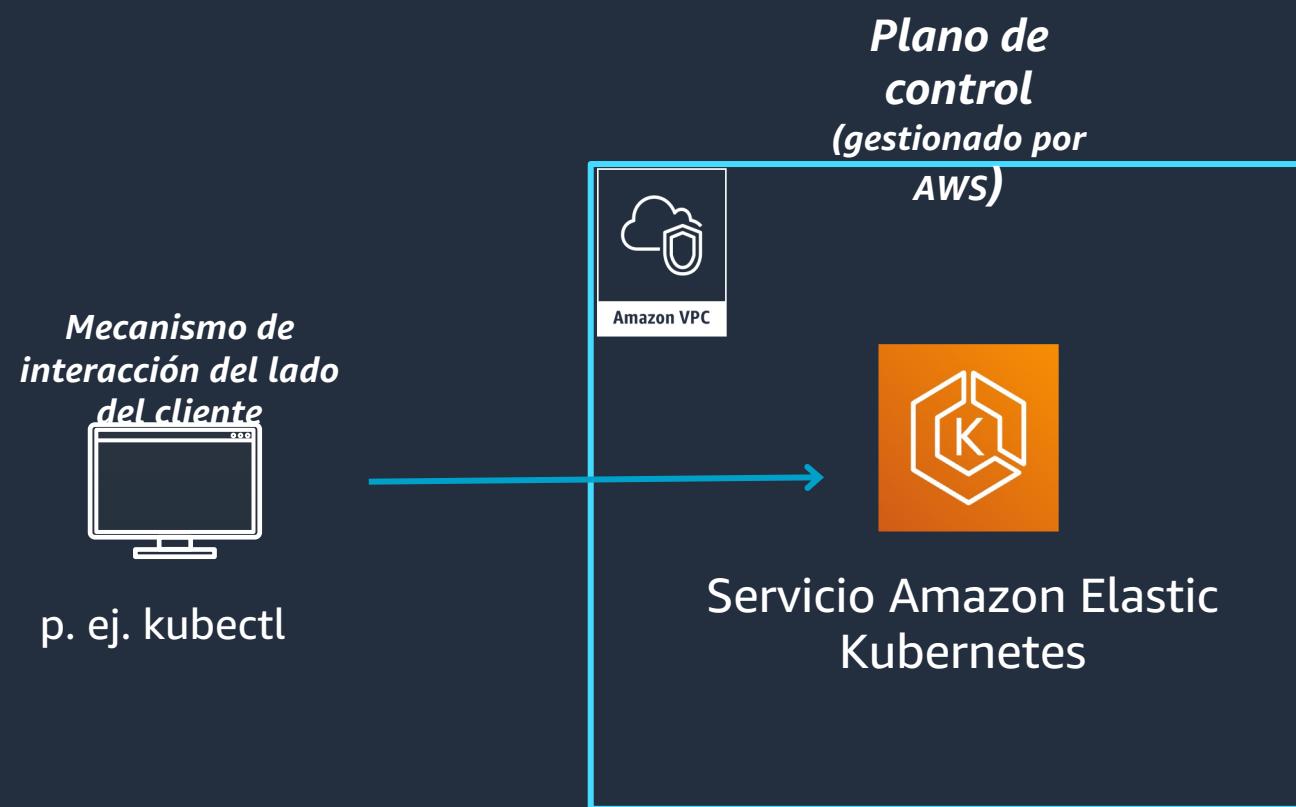
- Una estrategia con HPA y CAS ayuda a **optimizar los recursos informáticos para los clústeres**;
- Dimensionar correctamente las asignaciones de recursos de tu aplicación es un **paso fundamental para que los *pods* se escalen correctamente**;
- El HPA es más común que el VPA y **no se usan ambos al mismo tiempo**;
- Karpenter es más **flexible** que **CAS**;
- Otras buenas prácticas: **presupuestos de interrupción** de módulos, **réplicas múltiples para la producción**.

Optimización en la capa de cómputos

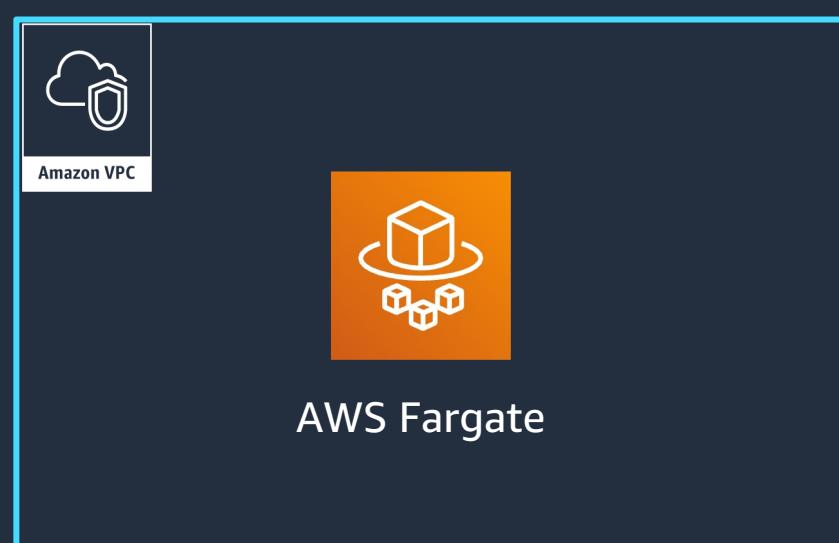
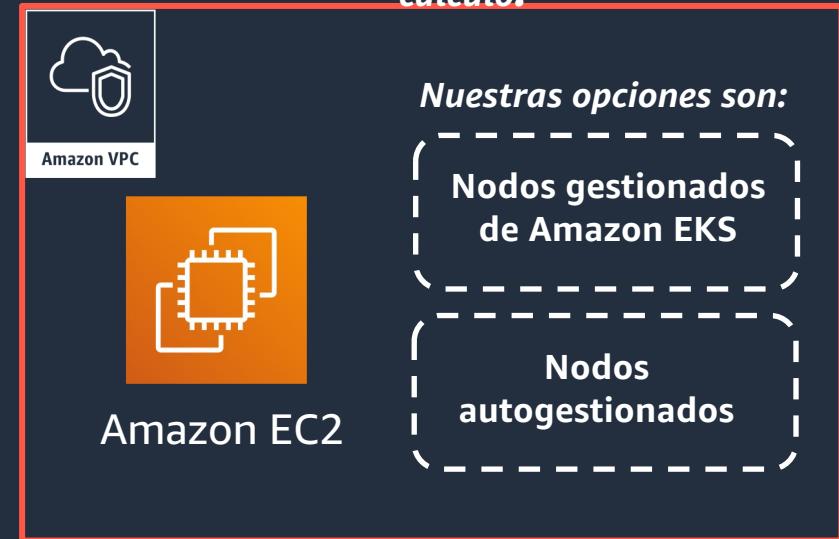


Amazon EKS

PLANO DE CONTROL + PLANO DE DATOS: OPCIONES
INFORMÁTICAS



Plano de datos y opciones de cálculo:



AWS Fargate: descarga de OPEX

- AWS Fargate permite a los clientes evitar la carga operativa que supone administrar las instancias de Amazon EC2 en clústeres;
- La reducción de la carga operativa permite reducir los costos operativos;
- El uso de AWS Fargate es una relación de compromiso importante que hay que evaluar.

Optimización de los costos de Amazon EC2



Rendimiento

Habilite las recomendaciones de recursos de AWS para reducir los costos y mejorar el rendimiento



Precio

Obtenga un precio y un rendimiento óptimos con diferentes modelos de compra



Capacidad

Administración sencilla de la capacidad en la plataforma informática más amplia y profunda

AWS Graviton



PRECIO-RENDIMIENTO

Ofrece una relación precio-rendimiento hasta un 40% superior en comparación con instancias comparables basadas en x86



EXTENSO ECOSISTEMA

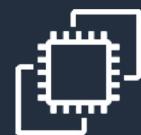
Compatible con los sistemas operativos Linux más populares, junto con muchas aplicaciones y servicios populares de AWS e ISV



SEGURIDAD MEJORADA

Proporcione funciones importantes para la seguridad de las aplicaciones, incluyendo siempre el cifrado DRAM de 256 bits

GRAVITÓN 2



C6G (D)



6 MG (D)



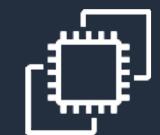
R6G (D)



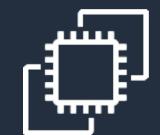
C6GN



X2GD



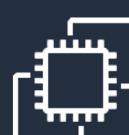
T4G



IM4GN



ES 4GEN



C7G

GRAVITÓN 3

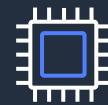
AWS Graviton

Ofrecer el mejor precio y rendimiento para las cargas de *trabajo* en la nube

Procesador Graviton



Disponible el primer procesador basado en ARM



Basado en núcleos Arm Neoverse de 64 bits con silicio diseñados por AWS



Hasta 16 vCPU, red de 10 Gbps, ancho de banda EBS de 3,5 Gbps

Procesador Graviton2



Rendimiento 7 veces mayor, 4 núcleos de computación y memoria 5 veces más rápida



Uso de tecnología de fabricación de 7 nm



Hasta 64 vCPU, red de 25 Gbps, ancho de banda EBS de 18 Gbps

Procesador Graviton 3



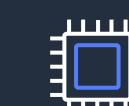
Hasta un 25% más de rendimiento en comparación con Graviton2



Un 60% más eficiente desde el punto de vista energético en comparación con las instancias EC2 comparables



Primero en la nube con memoria DDR5



Las instancias C7g ofrecen la mejor relación precio-rendimiento para cargas de trabajo con uso intensivo de cómputos en Amazon EC2.

Amazon EC2: Opciones de compra

Piense en la flexibilidad y concéntrese en esos puntos para optimizar el rendimiento y los costos

Bajo demanda



Pague la capacidad informática **por segundo** sin compromisos a largo plazo



Cargas de trabajo con altas varianzas, definición de la capacidad de las cargas de trabajo con estado

Planes de ahorro



Haga un **compromiso de 1 o 3 años** y reciba un descuento significativo en los precios a pedido

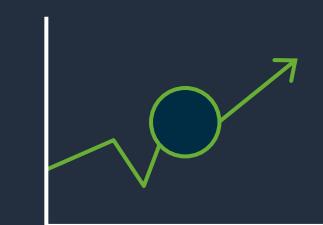


Uso comprometido y estable

EC2 Spot



Capacidad sobrante de Amazon EC2 con **ahorros de hasta un 90% en precios** bajo demanda



Cargas de trabajo tolerantes a errores, flexibles y apátridas

Planes de ahorro: tipos



Planes de ahorro de cómputos

Más flexibilidad, hasta un 66% de descuento

- ✓ Familia de instancias: pase de C5 a M5
- ✓ Región: traslado de Ohio a Londres
- ✓ Sistema operativo: Windows para Linux
- ✓ Arrendamiento: de dedicado a impago
- ✓ Opciones de procesamiento: desde Amazon EC2 hasta AWS Fargate

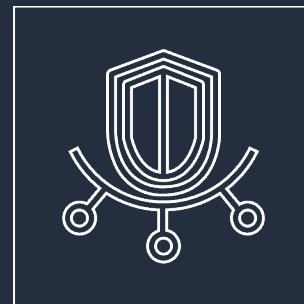


Planes de ahorro de instancias de EC2

Los precios más bajos, con hasta un 72% de descuento, en la familia de instancias seleccionada en una región de AWS específica

- ✓ Tamaño: de m5.xl a m5.4xl
- ✓ Sistema operativo: desde m5.xl Windows hasta m5.xl Linux
- ✓ Arrendamiento: de m5.xl dedicado a m5.xl predeterminado

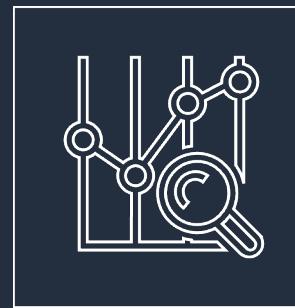
Herramientas de optimización en Amazon EC2



Asesor de confianza de AWS

Verificación de instancias de Amazon EC2 de alta utilización

Verificación de instancias de Amazon EC2 con bajo uso



Explorador de costos de AWS

recomendaciones de redimensionamiento

Reducir el tamaño dentro de la misma familia de instancias de Amazon EC2 para ahorrar costes



Optimizador de cómputos de AWS

Recomendaciones de tipos de instancias de EC2 para instancias de EC2 individuales o grupos de Auto Scaling

Buenas prácticas

- Identifique la combinación de opciones de compra y computación en función de sus *cargas de trabajo*;
- Al adoptar Spot, aproveche las herramientas para obtener los mejores ahorros de costos y disponibilidad de sus servicios;
- Adopte AWS Fargate para compensar los gastos operativos.

Referencias

Blog: [Optimización de costes para Kubernetes en AWS](#)

Documentación: [Uso de las etiquetas de recursos EC2 de EKS para la facturación](#)

Blog: [AWS y Kubecost colaboran para ofrecer un monitoreo de costos a los clientes de EKS](#)

Documentación: [Guía del usuario de EKS para grupos de nodos gestionados](#)

Taller: [Uso de instancias de EC2 SPOT con EKS](#)

Taller: [EKS y Karpenter](#)

¡Gracias!

¡Sus comentarios son muy bienvenidos!

<https://pulse.buildon.aws/survey/HIM23W8Q>

Day 2

¡Gracias!

¡Sus comentarios son muy bienvenidos!



<https://pulse.buildon.aws/survey/HIM23W8Q>