# EC2

## ¿Qué es Amazon EC2?

Amazon Elastic Compute Cloud (Amazon EC2) proporciona capacidad de computación escalable en la nube de Amazon Web Services (AWS). El uso de Amazon EC2 elimina la necesidad de invertir inicialmente en hardware, de manera que puede desarrollar e implementar aplicaciones en menos tiempo. Puede usar Amazon EC2 para lanzar tantos servidores virtuales como necesite, configurar la seguridad y las redes, y administrar el almacenamiento. Amazon EC2 le permite escalar hacia arriba o hacia abajo para controlar los cambios en los requisitos o los picos de popularidad, con lo que se reduce la necesidad de prever el tráfico.

Amazon EC2 ofrece la plataforma informática más amplia y profunda con elección de procesador, almacenamiento, red, sistema operativo y modelo de compra. Se Ofrecen los procesadores más rápidos de la nube y somos la única nube con 400 Gbps de red de Ethernet, instancias de GPU más poderosas para la capacitación de machine learning y las cargas de trabajo gráficas, así como las instancias de costo por inferencia más bajas de la nube. En AWS se ejecutan más cargas de trabajo de SAP, HPC, machine learning y Windows que en cualquier otra nube.

## Características de Amazon EC2

Amazon EC2 ofrece las siguientes características:

* Entornos informáticos virtuales, conocidos como instancias.
* Plantillas preconfiguradas para las instancias, conocidas como imágenes de máquina de Amazon (AMI), que empaquetan las partes que necesita para el servidor (incluido el sistema operativo y el software adicional).
* Varias configuraciones de CPU, memoria, almacenamiento y capacidad de red de las instancias, conocidos como tipos de instancias.
* Información de inicio de sesión segura para las instancias con pares de claves (AWS almacena la clave pública y usted guarda la clave privada en un lugar seguro).
* Volúmenes de almacenamiento para datos temporales que se eliminan cuando una instancia se detiene, se termina o se pone en hibernación, lo que se conoce como volúmenes del almacén de instancias.
* Volúmenes de almacenamiento persistente para los datos usando Amazon Elastic Block Store (Amazon EBS), conocidos como volúmenes de Amazon EBS.
* Varias ubicaciones físicas para los recursos, como las instancias y los volúmenes de Amazon EBS, conocidas como regiones y zonas de disponibilidad.
* Un firewall que permite especificar los protocolos, los puertos y los rangos de direcciones IP que pueden alcanzar las instancias mediante el uso de grupos de seguridad.
* Direcciones IPv4 estáticas para informática en la nube dinámica, conocidas como direcciones IP elásticas.
* Metadatos, conocidos como etiquetas, que se pueden crear y asignar a los recursos de Amazon EC2.
* Redes virtuales que puede crear que están aisladas lógicamente del resto de la nube de AWS y que, opcionalmente, puede conectar a su propia red, conocidas como nubes virtuales privadas (VPC).

## Instancias dedicadas

Las instancias dedicadas de Amazon EC2 brindan a las aplicaciones acceso directo al procesador y a la memoria del servidor subyacente. Estas instancias son ideales para cargas de trabajo que necesitan obtener acceso al conjunto de características del hardware (como Intel® VT-x) o para aplicaciones que deben ejecutarse en entornos no virtualizados debido a requisitos relacionados con licencias o compatibilidad. Las instancias dedicadas se basan en el sistema Nitro, un conjunto de componentes desarrollados por AWS para la protección del hardware y la reducción de la carga del hardware. Estos componentes se unen para proporcionar recursos de almacenamiento y redes de alto desempeño a las instancias EC2 de forma segura. Las instancias dedicadas son instancias EC2 y, por lo tanto, ofrecen el mismo nivel sólido de seguridad, fiabilidad, elasticidad de capacidad y compatibilidad para diferentes sistemas operativos y paquetes de software que otras instancias EC2 virtuales. También es posible usar las instancias dedicadas con servicios de AWS, como Amazon Virtual Private Cloud (VPC), Elastic Block Store (EBS), Elastic Load Balancing (ELB), entre otros.

## Optimización de costos y el desempeño informático con la flota de Amazon EC2

Mediante una única llamada a la API, la flota de Amazon EC2 le permite aprovisionar capacidad informática en tipos de instancias de EC2, zonas de disponibilidad y modelos de compra a fin de optimizar la escala, el desempeño y el costo. Puede especificar qué capacidad se lanzará mediante la flota de EC2 con las metodologías bajo demanda y spot. También puede definir qué tipos de instancias prefiere y si desea ajustar la capacidad en función de los núcleos, las instancias o la memoria. Lea las preguntas frecuentes y el blog de AWS para obtener más información. Ahora puede obtener acceso a las capacidades de flota de EC2 mediante Amazon EC2 Auto Scaling para aprovisionar y escalar automáticamente la capacidad de cómputo entre los tipos de instancia EC2, las zonas de disponibilidad y las opciones de compra en un solo grupo de Auto Scaling.

## Pausar y reanudar las instancias

Puede poner en hibernación sus instancias Amazon EC2 respaldadas por Amazon EBS y reanudarlas desde estado más adelante. Las aplicaciones que tardan un poco en arrancar y persisten en su estado en memoria (RAM) pueden beneficiarse de esta característica. La hibernación le ofrece todos los beneficios de detener y comenzar; además, los datos de la memoria (RAM) también se mantienen de una sesión a otra. No se le cobrará por el uso de la instancia mientras esta esté hibernada. El almacenamiento se cobra según las tarifas estándar de EBS. Para obtener más información sobre la hibernación y los tipos de instancia y sistemas operativos compatibles.

## Instancias de informática con GPU

Los clientes que necesitan capacidad de procesamiento masiva con punto flotante se aprovecharán los beneficios de la próxima generación de instancias informáticas con GPU de uso general de AWS, instancias P3 de Amazon EC2 con hasta 8 GPU con núcleo NVIDIA® V100 Tensor. Las instancias P3 ofrecen hasta 1 petaFLOPS de precisión mixta, 125 TeraFLOPS de precisión simple y 62 TeraFLOPS de precisión doble de rendimiento de punto flotante. Una interconexión de NVLink de segunda generación de 300 GB/s hace posible la comunicación de GPU a GPU a alta velocidad y baja latencia. Las instancias P3 también incluyen hasta 96 vCPU basadas en procesadores Intel personalizados, 768 GB de DRAM y 100 Gbps de ancho de banda de red adicional dedicado mediante el uso de Elastic Network Adapter (ENA). Las instancias P3 son ideales para el aprendizaje automático, la informática de alto rendimiento, la dinámica fluida computacional, las finanzas computacionales, el análisis sísmico, el modelado molecular, la genómica y las cargas de trabajo de renderizado.

## Instancias de gráficos con GPU

Los clientes que necesiten una capacidad de gráficos alta se beneficiarán de las instancias de gráficos con GPU. La instancia de gráficos con GPU de la generación actual, la instancia G3, proporciona acceso a las GPU Tesla M60 de NVIDIA, cada una de las cuales cuenta con hasta 2048 núcleos de procesamiento en paralelo, 8 GiB de memoria de GPU y un codificador de hardware que admite hasta 10 transmisiones H.265 (HEVC) 1080p30 y hasta 18 transmisiones H.264 1080p30. Con las versiones más recientes de los controladores, estas GPU son compatibles con OpenGL, DirectX, CUDA, OpenCL y Capture SDK (anteriormente conocido como GRID SDK). Las instancias de GPU Graphics están indicadas para visualizaciones 3D, estaciones de trabajo remotas con un uso intensivo de gráficos, renderización 3D, transmisión de aplicaciones, codificación de vídeo y otros trabajos de gráficos en el lado del servidor.

## Instancias de E/S de alto desempeño

Los clientes que necesitan un acceso de E/S, aleatorio, de baja latencia y frecuente a sus datos se pueden beneficiar de las instancias de E/S elevada. Las instancias de E/S de alto desempeño son un tipo de instancia de Amazon EC2 que puede ofrecer a los clientes tasas de E/S aleatorias de más de 3 millones de IOPS. Las instancias I3 y I3en de E/S de alto desempeño están respaldadas por un SSD basado en memoria no volátil exprés (NVMe) y son ideales para los clientes que ejecutan bases de datos NoSQL de muy alto desempeño, sistemas transaccionales y cargas de trabajo de Elastic Search. Las instancias de E/S de alto desempeño ofrecen un desempeño de disco secuencial de hasta 16 GB/s, que es ideal para las cargas de trabajo de análisis.

## Instancias de almacenamiento HDD denso

Los clientes que necesitan un nivel muy alto de volumen de almacenamiento por instancia y un alto nivel de E/S secuenciales para aplicaciones con grandes volúmenes de datos, como el almacén de datos Massively Parallel Processing (MPP), la informática distribuida de MapReduce y Hadoop, así como aplicaciones de registro y procesamiento de datos, pueden beneficiarse de las instancias de alta densidad de almacenamiento. Las instancias de almacenamiento denso son un tipo de instancia de Amazon EC2 que puede suministrar a los clientes un rendimiento de E/S secuencial de hasta 3,9 GB/s y hasta 48 TB de almacenamiento de instancias en 24 unidades de disco duro. También pueden ofrecer un equilibrio con menos almacenamiento y capacidad de memoria por vCPU con una red basada en ENA para hasta 25 Gb/s de ancho de banda de red dentro de un grupo de ubicación.

## Configuraciones de CPU optimizadas

La característica de optimización de CPU le otorga un mayor control sobre las instancias de Amazon EC2 en dos aspectos. En primer lugar, puede especificar una cantidad personalizada de vCPU al momento de lanzar instancias nuevas para ahorrar costos de licencias basadas en vCPU. En segundo lugar, puede deshabilitar los subprocesos múltiples para cargas de trabajo que se ejecuten correctamente con las CPU de subproceso único, como determinadas aplicaciones de informática de alto rendimiento (HPC). Si desea obtener más información acerca de cómo puede ayudarlo la característica de optimización de CPU.

## Costos Amazon EC2

Probar Amazon EC2 es gratis. Existen cinco modelos de compra de instancias de Amazon EC2: bajo demanda, Savings Plans, instancias reservadas e instancias de spot. También puede pagar por alojamientos dedicados, que le proporcionan instancias EC2 en servidores físicos exclusivos.

* **Capa gratuita:** La capa gratuita de AWS incluye 750 horas de instancias t2.micro de Linux y Windows, (t3.micro para las regiones en las que t2.micro no está disponible) cada mes durante un año.
* **Bajo demanda:** Con las instancias bajo demanda, paga por la capacidad informática por hora o por segundo, según las instancias que se usen. Ya no serán necesarios los contratos a largo plazo ni los pagos iniciales. Puede aumentar o reducir la capacidad informática en función de las exigencias de la aplicación y pagar únicamente la tarifa por hora específica de la instancia que use.
* **Instancias de spot:** Las instancias de spot de Amazon EC2 le permiten solicitar capacidad informática sobrante de Amazon EC2 con descuentos de hasta el 90 % en comparación con el precio de las instancias bajo demanda.
* **Savings Plans:** El programa Savings Plans es un modelo de precios flexible que ofrece precios económicos por el uso de EC2 y Fargate, a cambio de comprometerse a una cantidad constante de uso (medida en USD/hora) durante el término de 1 o 3 años.
* **Hosts dedicados:** Un host dedicado es un servidor físico de EC2 exclusivo para su uso. Los hosts dedicados pueden ayudarle a reducir costos porque le permiten usar sus licencias existentes de software enlazado al servidor, incluidos Windows Server, SQL Server y SUSE Linux Enterprise Server (en función de los términos de su licencia).