AMAZON EC2

CONTENIDO

AMAZON EC2 2
INTRODUCCIÓN 2
AMAZON ELASTIC COMPUTE CLOUD (EC2) 3
BENEFICIOS DE AMAZON EC2 4
PRECIOS DE AMAZON EC2 6
INSTANCIA BAJO DEMANDA 6
INSTANCIA RESERVADA 6
INSTANCIA PUNTUAL 7
SAVINGS PLANS 7
SERVIDORES DEDICADOS 8
FUNCIONAMIENTO DE AMAZON EC2 8
TIPOS DE INSTANCIAS DE AMAZON EC2 9
INSTANCIAS DE PROPÓSITO GENERAL 9
INSTANCIAS DE CÓMPUTO OPTIMIZADO 9
INSTANCIAS OPTIMIZADAS DE MEMORIA 9
INSTANCIAS DE CÓMPUTO ACELERADO 10
INSTANCIAS OPTIMIZADAS DE ALMACENAMIENTO 10
ESCALADO DE AMAZON EC2 10
ESCALABILIDAD 10
AMAZON EC2 AUTO SCALING 11
BENEFICIOS DEL AUTO SCALING 11
EJEMPLO DE AUTO SCALING 13
DIRECCIÓN DE TRÁFICO CON ELASTIC LOAD BALANCING 15
EJEMPLO DE ELASTIC LOAD BALANCING 15

AMAZON EC2

INTRODUCCIÓN

Las empresas, ya sean de servicios, fabricación, seguros, distribución de contenido, etcétera, utilizan el modelo cliente/servidor, en el que un cliente envía una solicitud al servidor, el servidor realiza un trabajo y luego envía una respuesta.

Dicho modelo permite a las empresas ofrecer sus productos, recursos o datos a los usuarios finales y requiere servidores para impulsar el negocio y las aplicaciones, por lo tanto, el modelo cliente/servidor necesita capacidad de cómputo para alojar las aplicaciones empresariales y proporcionar la potencia de cómputo que las empresas necesitan. Cuando trabajamos con AWS, esos servidores son virtuales y el servicio que se utiliza para acceder a ellos se llama EC2.
El uso de EC2 para cómputo es muy flexible, rentable y rápido comparado con la ejecución de los propios servidores en un centro de datos propio.

El tiempo y el dinero necesarios para poner en marcha los recursos en las instalaciones son bastante elevados. Cuando se posee una flota propia de servidores físicos, primero hay que investigar mucho para ver qué tipo de servidores se quieren comprar y cuánto se van a necesitar. A continuación, compra ese hardware por adelantado y tendrá que esperar varias semanas o meses a que un proveedor le entregue esos servidores. Luego los lleva a un centro de datos de su propiedad o alquilado para instalarlos, montarlos, apilarlos y cablearlos. Luego se asegura de que estén protegidos y encendidos y ya están listos para ser utilizados. Solo entonces podrá empezar a alojar sus aplicaciones sobre estos servidores.

Lo peor es que una vez comprados estos servidores, se queda con ellos, tanto si los utiliza como si no. Con EC2 es mucho más fácil comenzar, AWS ya se encargó de la parte difícil, ya creó y aseguró los centros de datos, ya compró los servidores, los montó y apiló y ya están en línea listos para ser utilizados. AWS opera de forma constante una cantidad masiva de cómputo y puede utilizar cualquier parte de esa capacidad cuando la necesite.

Todo lo que tiene que hacer es solicitar las instancias de EC2 que desee y se lanzarán e iniciarán listas para ser utilizadas en pocos minutos. Una vez que haya terminado, puede detener o terminar de forma sencilla las instancias de EC2. No está limitado a servidores que no necesita o no desea, ni depende de estos. El uso de instancias de EC2 puede variar mucho con el tiempo.

Solo paga por lo que usa, porque con EC2 solo se paga por las instancias en ejecución, no por las instancias detenidas o terminadas. EC2 se ejecuta sobre máquinas host físicas administradas por AWS mediante tecnología de virtualización.

AMAZON ELASTIC COMPUTE CLOUD (EC2) proporciona capacidad de cómputo segura y de tamaño modificable en la nube como instancias de Amazon EC2.
Un cliente puede elegir entre una amplia variedad de instancias. Algunas son de uso intensivo de CPU, otras de memoria, algunas son de computación acelerada optimizada, algunas están optimizadas para almacenamiento, otras son instancias de entrada/salida (E/S) y algunas son de propósito general. Según el caso de uso, el cliente puede elegir entre una variedad de tipos de instancias, por ejemplo, si está ejecutando una carga de trabajo de base de datos que necesita mucha memoria, puede elegir una instancia de memoria intensiva, y si planea ejecutar aprendizaje automático (ML), puede elegir una instancia de computación acelerada.

Amazon Elastic Compute Cloud (EC2) proporciona una capacidad informática casi infinita en la nube, y no solo es confiable, sino también segura. Puede ejecutar cualquier tipo de carga de trabajo en la nube de Amazon y no tiene que invertir muchos gastos de capital para obtener recursos informáticos. El modelo de computación en la nube es de pago por uso, lo que significa que solo paga por los recursos que va a utilizar por horas o por segundo, según el tipo de instancia. Por lo tanto, para adquirir nuevos servidores, no es necesario esperar meses para que se apruebe un presupuesto, sino que puede implementar sus aplicaciones más rápido e innovar rápidamente.

El ecosistema EC2 de Amazon está diseñado para escalar, por lo tanto, siempre que se produzca un aumento en el tráfico de su carga de trabajo, puede obtener rápidamente servidores adicionales casi instantáneamente, y cuando el tráfico se reduce, puede deshacerse de esos servidores. Por ejemplo, digamos que para su negocio normal durante los días de semana necesita un servidor con 16 CPU, pero el fin de semana espera el doble de tráfico. Puede aprovisionar otro servidor adicional con 16 CPU solo para el fin de semana, y cuando llegue el lunes, puede deshacerse de ese servidor adicional. Tiene que pagar por el servidor adicional solo por las horas del sábado y domingo.

Una instancia EC2 es similar en tierra a una máquina virtual (VMWare, Hyper-V) y el cliente tiene total administración de la máquina.
La máquina virtual no es algo nuevo, pero la facilidad de aprovisionamiento de instancias de EC2 les permite a los programadores y a las empresas innovar más rápidamente. AWS hace que sea mucho más fácil y más rentable que adquiera servidores a través de este modelo de cómputo como servicio.

Imagine que es la persona responsable de la arquitectura de los recursos de su empresa y necesita dar soporte a sitios web nuevos. Con los recursos en las instalaciones tradicionales, debe hacer lo siguiente:

- Gastar dinero por adelantado para comprar hardware.
- Esperar a que le entreguen los servidores.

- Instalar los servidores en su centro de datos físico.
- Realizar todas las configuraciones necesarias.

En comparación, con una instancia de Amazon EC2 puede utilizar un servidor virtual para ejecutar aplicaciones en la nube de AWS, con ventajas como que:

- Puede aprovisionar y lanzar una instancia de Amazon EC2 en cuestión de minutos.
- Puede dejar de usarla cuando haya terminado de ejecutar una carga de trabajo.
- Solo paga por el tiempo de cómputo que utiliza cuando una instancia se está ejecutando, no cuando está detenida o finalizada.
- Puede ahorrar costos si solo paga por la capacidad del servidor que necesite o desee.

BENEFICIOS DE AMAZON EC2

- Tiempo de comercialización: La mayor ventaja de ejecutar una instancia EC2 es el tiempo de comercialización. Puede implementar cualquier servidor casi instantáneamente y no tiene que esperar semanas o meses para obtener un nuevo servidor. Esto también facilita la innovación porque puede obtener rápidamente los recursos para su nuevo proyecto. Si el proyecto finaliza, simplemente puede deshacerse de los servidores y comenzar un nuevo proyecto con nuevos recursos.

- Escalabilidad: Otro beneficio de ejecutar EC2 es la escalabilidad, puede escalar hacia arriba y hacia abajo en cualquier momento dependiendo de su carga de trabajo. En el pasado, siempre tenía que aprovisionar en exceso los recursos solo para asegurarse de poder soportar la demanda máxima. Pero con las instancias EC2 no es necesario que sobre aprovisione los recursos, simplemente aprovisiona los recursos necesarios para su negocio y siempre que haya un crecimiento adicional o un pico, puede implementar rápidamente servidores adicionales que puedan atender la demanda adicional. La tecnología EC2 Auto Scaling le permite escalar hacia arriba o hacia abajo automáticamente las aplicaciones según las necesidades. De esta manera, no solo maximiza el rendimiento, sino que también minimiza el costo.

- Control: Usted tiene control total sobre los servidores al igual que tiene control sobre los servidores en su centro de datos. Puede iniciar y detener el servicio en cualquier momento y tiene acceso de root a los servidores. Puede interactuar con las instancias como interactúa con cualquier otra máquina. Puede controlar o reiniciar los servidores de forma remota y también puede acceder a ellos utilizando la consola de Amazon.

- EC2 ofrece un entorno **confiable** donde las instancias de reemplazo se pueden poner en marcha de manera rápida y predecible. El acuerdo de nivel de servicio de EC2 tiene una disponibilidad del 99,99 por ciento para cada región.

- **Seguro**: Toda la infraestructura de Amazon es segura, de hecho, la seguridad es un trabajo de máxima prioridad para Amazon, todo lo que opera bajo la nube EC2 es seguro. Puede crear un recurso EC2 junto con Amazon VPC para proporcionar seguridad y funcionalidad de red para sus recursos informáticos.

- **Varios tipos de instancias**: Amazon EC2 le permite seleccionar entre una variedad de instancias. Puede elegir la instancia en función del sistema operativo, el paquete de software, el tamaño de almacenamiento de la CPU o la memoria, etc. También puede elegir una instancia de Amazon Marketplace donde varios proveedores externos ofrecen sus servidores preempaquetados (AMI).

- **Integración:** Amazon EC2 está integrado con la mayoría de los servicios de AWS, como S3, VPC, Lambda, RedShift, RDS, EMR, etc. Con EC2 y los demás servicios de AWS, puede obtener una solución completa para todas sus necesidades de TI.

- Rentable: Dado que solo paga por el uso del servidor por horas o por segundos, según la instancia que ejecute, realmente no tiene que pagar un gran gasto de capital cuando aprovisiona servidores en EC2.

Cuando separa servidores en Amazon EC2, tiene control total sobre el tipo de almacenamiento que utiliza, las configuraciones de red, la configuración de seguridad, etc. La interfaz web de EC2 le permite configurar un servidor en un tiempo mínimo. Imagine un modelo de implementación tradicional en el que debe aprovisionar un servidor. La instalación del sistema operativo tarda un par de horas, sin incluir el tiempo que lleva configurar el almacenamiento y la red.
Si suma todo este tiempo, sería un par de días de esfuerzo. Con Amazon EC2, el tiempo necesario para obtener e iniciar el nuevo servidor es cuestión de minutos. Dado que ahora solo toma unos minutos implementar un servidor, en realidad puede implementar cientos o miles de servidores casi instantáneamente, y dado que este modelo es muy escalable, puede escalar hacia arriba o hacia abajo rápidamente.

Puede elegir entre varios tipos de instancias, sistemas operativos y paquetes de software. Amazon EC2 le permite seleccionar una configuración de memoria, CPU, almacenamiento de instancia y tamaño de partición de arranque que sea óptima para su elección de sistema operativo y aplicación. Por ejemplo, su elección de sistemas operativos incluye numerosas distribuciones de Linux y Microsoft Windows Server. Los siguientes son algunos de los sistemas operativos compatibles con EC2:

- Windows
- Amazon Linux
- Debian

- SUSE
- CentOS
- Red Hat Enterprise Linux
- Ubuntu

PRECIOS DE AMAZON EC2

Amazon ofrece múltiples opciones de precios, según la duración de la instancia y su flexibilidad de pago. Las instancias se dividen en varias categorías desde una perspectiva de precios:

INSTANCIA BAJO DEMANDA: Este es el modelo de precios más popular de una instancia EC2. En este modelo, solo paga por el uso con una tarifa fija por hora o facturación por segundo. No hay costos iniciales ni costos ocultos ni nada más. Digamos que si crea una instancia y la usa durante diez horas y luego descarta la instancia, debe pagar solo por diez horas. No hay compromiso ni plazo mínimo involucrado y puede escalar hacia arriba o hacia abajo en cualquier momento. Si está desarrollando una nueva aplicación y no sabe cuántos recursos va a tomar, una instancia bajo demanda es una excelente manera de comenzar.

INSTANCIA RESERVADA: Ofrece hasta un 75 por ciento de descuento en comparación con una instancia bajo demanda. Si ya sabe cuántos recursos va a tomar su carga de trabajo y durante cuánto tiempo, una instancia reservada proporcionará el máximo beneficio de costo. Uno de los mejores casos de uso de instancias reservadas es ejecutar la carga de trabajo de producción. Supongamos que sabe que su servidor de producción va a ocupar 16 CPU y necesita esta configuración de servidor de producción durante al menos un año, puede reservar la capacidad durante un año y obtener el descuento en comparación con una instancia bajo demanda. Una instancia reservada es ideal cuando sabe que su aplicación tiene un estado bastante estable o es predecible en términos de rendimiento. Dado que las reservas requieren un compromiso de uno o tres años, es importante conocer la naturaleza de la carga de trabajo antes de comprometerse con una instancia reservada.
Hay varias opciones de pago disponibles cuando reserva una instancia. Puede pagarla en su totalidad, lo que se llama reservado por adelantado, o hacer un pago parcial por adelantado, lo que se denomina reservado parcial por adelantado, o no puede pagar nada por adelantado y todo se factura en un ciclo de facturación mensual que se denomina sin reserva inicial.

La instancia reservada se divide en dos subcategorías: Instancia reservada estándar e instancia reservada convertible. La instancia reservada estándar es la normal que acaba de estudiar. La instancia reservada convertible proporciona una mayor flexibilidad si su requisito de cómputo cambia durante el período de tiempo dado. Una instancia reservada convertible le brinda la capacidad y flexibilidad de intercambiar la instancia de una clase de familia a otra clase si sus necesidades de computación cambian.

INSTANCIA PUNTUAL (SPOT): AWS tiene las capacidades informáticas más grandes entre todos los diferentes proveedores de nube y, a menudo, parte del exceso de capacidad informática no se utiliza. AWS le brinda la posibilidad de ofertar por la capacidad no utilizada y puede obtener hasta un 90 por ciento de descuento en comparación con los precios bajo demanda. Este modelo de precios se denomina precio al contado y la instancia creada con el modelo de precios al contado se denomina instancia al contado. La instancia se ejecuta en el modelo de licitación y puede ofertar por el precio al contado, el precio al contado fluctúa en función de la oferta y la demanda, y si alguien le sobreoferta, pierde la instancia en muy poco tiempo. Las instancias de spot ofrecen las mismas funciones a las que está acostumbrado con EC2, pero por una fracción del costo. Sin embargo, debe tener cuidado al elegir el tipo de carga de trabajo que va a ejecutar en la instancia puntual. Las instancias de spot son excelentes para cargas de trabajo que pueden reiniciarse desde donde fallaron, en otras palabras, son excelentes para ejecutar proyectos que no son de misión crítica.
A menudo, los clientes agregan algunas instancias puntuales junto con instancias bajo demanda para proporcionar potencia adicional. Estos son algunos de los casos de uso en los que se puede aprovechar la instancia puntual:

- Procesamiento por lotes.
- Aprendizaje automático (PyTorch, Tensorflow o trabajos que requieren mucho entrenamiento).
- Integración continua y despliegue continuo (CI/CD).
- Todo lo que sea tolerante a fallas o sin estado que pueda ser de instancia flexible.

El caso de uso más común en el que las instancias puntuales tienen éxito es la creación de una carga de trabajo tolerante a fallas porque EC2 puede reclamar la instancia puntual con una notificación de dos minutos. Debe diseñar sus cargas de trabajo de tal manera que puedan manejar las interrupciones. Puede mezclar y combinar diferentes tipos de instancias en una flota puntual y, por lo tanto, si hay una interrupción en un tipo particular de instancia, otros tipos de instancias continúan ejecutándose que respaldan su carga de trabajo.

SAVINGS PLANS: AWS ofrece Savings Plans para algunos servicios de cómputo, incluido Amazon EC2. Los Savings Plans de las instancias de EC2 reducen los costos de las instancias de EC2 cuando se compromete a realizar gastos por hora a una familia de instancias y a una región por un periodo de 1 año o de 3 años. Este compromiso del periodo genera ahorros de hasta el 72 % comparado a las tarifas bajo demanda. Cualquier uso hasta el compromiso se cobra según la tarifa de Savings Plans con descuento, (por ejemplo, 10 USD por hora). Cualquier uso más allá del compromiso se cobra en base a las tarifas bajo demanda regulares.

Los Savings Plans de las instancias de EC2 son una buena opción si necesita flexibilidad en su uso de Amazon EC2 sobre la duración del periodo de compromiso. Tiene el beneficio de ahorrar en la ejecución de cualquier instancia de EC2 dentro de una familia de instancias de

EC2 en una región elegida (por ejemplo, el uso de M5 en Virginia del Norte) sin importar la zona de disponibilidad, el tamaño, el sistema operativo o la tenencia de la instancia.

Los ahorros con los Savings Plans de las instancias de EC2 son similares a los ahorros suministrados por las instancias reservadas Standard.
Sin embargo, al contrario de las instancias reservadas, no debe especificar de antemano el tipo y el tamaño de la instancia de EC2 (por ejemplo m5.xlarge), el sistema operativo ni la tenencia para obtener un descuento. Además, no debe comprometerse a un número determinado de instancias de EC2 durante un periodo de 1 año o de 3 años. Asimismo, los Savings Plans de las instancias de EC2 no incluyen una opción de reserva de capacidad de EC2.

SERVIDORES DEDICADOS: Son servidores físicos con capacidad de instancias de Amazon EC2 totalmente dedicados a su uso.
Puede utilizar las licencias de software por zócalo, por núcleo o por máquina virtual (VM) existentes para ayudar a mantener el cumplimiento de las licencias. Puede comprar reservas de hosts dedicados y hosts dedicados bajo demanda. De todas las opciones de Amazon EC2 cubiertas, los hosts dedicados son los de mayor costo.

FUNCIONAMIENTO DE AMAZON EC2

1. **LANZAMIENTO**: Primero se lanza una instancia, empiece por seleccionar una plantilla con configuraciones básicas para su instancia. Estas configuraciones incluyen el sistema operativo, el servidor de aplicaciones o las aplicaciones.
   También seleccione el tipo de instancia, que es la configuración de hardware específica de la instancia.
   Mientras se prepara para lanzar una instancia, debe especificar la configuración de seguridad para controlar el tráfico de red que puede entrar y salir de la instancia.

2. **CONECTAR:** A continuación, conéctese a la instancia. Puede hacerlo de varias formas. Sus programas y aplicaciones tienen varios métodos diferentes para conectarse directamente a la instancia e intercambiar datos. Los usuarios también pueden conectarse a la instancia al iniciar sesión y acceder al escritorio remoto de la computadora.

3. **USO:** Una vez que se haya conectado a la instancia, puede empezar a usarla. Puede ejecutar comandos para instalar software, añadir almacenamiento, copiar y organizar archivos y mucho más.

TIPOS DE INSTANCIAS DE AMAZON EC2

Los tipos de instancias de Amazon EC2 están optimizados para distintas tareas. Al seleccionar un tipo de instancia, tenga en cuenta las necesidades específicas de sus cargas de trabajo y aplicaciones. Es posible que incluya requisitos para las capacidades de cómputo, memoria o almacenamiento.

INSTANCIAS DE PROPÓSITO GENERAL: Brindan un equilibrio entre recursos de cómputo, de memoria y de redes. Puede utilizarlos para diversas cargas de trabajo, tales como: Servidores de aplicaciones, servidores de juegos, servidores backend para aplicaciones empresariales, bases de datos pequeñas y medianas.

Imagine que tiene una aplicación en la que las necesidades de recursos de cómputo, memoria y redes son aproximadamente equivalentes. Puede considerar ejecutarla en una instancia de uso general porque la aplicación no requiere optimización en ninguna área de recursos única.

INSTANCIAS DE CÓMPUTO OPTIMIZADO: Son ideales para aplicaciones vinculadas a cómputo que se benefician de los procesadores de alto rendimiento. Al igual que las instancias de propósito general, puede usar instancias de cómputo optimizado para cargas de trabajo como servidores web, de aplicaciones y de juegos.

Sin embargo, la diferencia es que las aplicaciones de cómputo optimizado son ideales para servidores web de alto rendimiento, servidores de aplicaciones con uso intensivo de cómputo y servidores de juegos dedicados. También puede usar instancias optimizadas para informática para cargas de trabajo de procesamiento por lotes que requieren procesar muchas transacciones en un solo grupo.

INSTANCIAS OPTIMIZADAS DE MEMORIA: Están diseñadas para ofrecer un rendimiento rápido para las cargas de trabajo que procesan grandes conjuntos de datos en la memoria.

En cómputo, la memoria es un área de almacenamiento temporal, contiene todos los datos e instrucciones que necesita una unidad central de procesamiento (CPU) para poder completar acciones. Antes de que se pueda ejecutar un programa o una aplicación de computadora, se carga desde el almacenamiento en la memoria. Este proceso de carga previa le proporciona a la CPU acceso directo al programa de computadora.

Imagine que tiene una carga de trabajo que requiere que se carguen previamente grandes cantidades de datos antes de ejecutar una aplicación. Este escenario puede ser una base de datos de alto rendimiento o una carga de trabajo que implique procesar en tiempo real una gran cantidad de datos no estructurados. En este tipo de casos de uso, considere usar una instancia optimizada para memoria. Las instancias optimizadas para memoria le permiten

ejecutar cargas de trabajo con altas necesidades de memoria y obtener un excelente rendimiento.

INSTANCIAS DE CÓMPUTO ACELERADO: Utilizan aceleradores de hardware, o coprocesadores, para realizar algunas funciones de manera más eficiente de lo que es posible en el software que se ejecuta en las CPU. Ejemplos de estas funciones incluyen cálculos numéricos de punto flotante, procesamiento de grafos y coincidencia de patrones de datos.

En cómputo, un acelerador de hardware es un componente que puede acelerar el procesamiento de datos. Las instancias de cómputo acelerado son ideales para cargas de trabajo como aplicaciones gráficas, streaming de juegos y streaming de aplicaciones.

INSTANCIAS OPTIMIZADAS DE ALMACENAMIENTO: Están diseñadas para las cargas de trabajo que necesitan realizar con gran frecuencia accesos secuenciales de lectura y escritura a grandes conjuntos de datos en el almacenamiento local. Algunos ejemplos de cargas de trabajo adecuadas para instancias optimizadas de almacenamiento incluyen sistemas de archivos distribuidos, aplicaciones de almacén de datos y sistemas de procesamiento de transacciones en línea (OLTP) de alta frecuencia.

En cómputo, el término operaciones de entrada/salida por segundo (IOPS) es una métrica que mide el rendimiento de un dispositivo de almacenamiento. Indica cuántas operaciones de entrada o salida diferentes puede realizar un dispositivo en un segundo. Las instancias optimizadas para almacenamiento están diseñadas para ofrecer decenas de miles de IOPS aleatorias y de baja latencia a las aplicaciones.

Puede considerar las operaciones de entrada como datos introducidos en un sistema, como registros ingresados en una base de datos. Una operación de salida son datos generados por un servidor. Un ejemplo de salida podría ser el análisis realizado en los registros de una base de datos. Si tiene una aplicación que requiere un alto nivel de IOPS, una instancia optimizada para almacenamiento puede proporcionar un mejor rendimiento en comparación con otros tipos de instancias no optimizadas para este tipo de casos de uso.

ESCALADO DE AMAZON EC2

ESCALABILIDAD

La escalabilidad implica comenzar solo con los recursos que necesita y diseñar la arquitectura para responder automáticamente a la demanda cambiante mediante el escalado o la reducción horizontal. Como resultado, solo paga por los recursos que utiliza.

No tiene que preocuparse por la falta de capacidad de cómputo para satisfacer las necesidades de sus clientes.

Si quisiera que el proceso de escalado se realizara automáticamente, el servicio de AWS que proporciona esta funcionalidad para las instancias de Amazon EC2 es Amazon EC2 Auto Scaling.

AMAZON EC2 AUTO SCALING

Auto Scaling es la tecnología que le permite escalar su carga de trabajo hacia arriba y hacia abajo automáticamente según las reglas que defina. Es una de las innovaciones que hace que la nube sea elástica y lo ayuda a personalizar según sus propios requisitos. Con Auto Scaling, no es necesario que sobre aprovisione los recursos para satisfacer la demanda máxima porque Auto Scaling configurará nuevos recursos automáticamente y luego los eliminará cuando disminuya la demanda.

Las implementaciones en las instalaciones locales requieren que los clientes pasen por un extenso ejercicio de dimensionamiento, esencialmente adivinando los recursos necesarios para cumplir con las cargas de trabajo máximas. La experiencia demuestra que es casi imposible calcular correctamente las estimaciones de tamaño y la mayoría de las veces, los clientes terminan con recursos infrautilizados mientras subestiman los recursos para las cargas de trabajo máximas. Otras veces, los clientes planean para el pico aprovisionando en exceso los recursos. Por ejemplo, puede aprovisionar todo el hardware para el Black Friday al comienzo del año, ya que obtiene su presupuesto de capital durante el comienzo del año. Entonces, durante todo el año, esos servidores funcionan solo con, digamos, entre un 15 y un 20 por ciento de CPU y alcanzan el pico durante la venta del Black Friday. En este caso, ha desperdiciado mucha capacidad de cómputo a lo largo del año que podría haber sido utilizado para algún otro propósito.

Con AWS tiene la capacidad de activar los servidores cuando sus cargas de trabajo requieren recursos adicionales y volver a desactivarlos cuando la demanda desciende.

BENEFICIOS DEL AUTO SCALING

Estos son los principales beneficios de Auto Scaling:

• **ESCALADO DINÁMICO**: La mayor ventaja de Auto Scaling es el escalado dinámico de recursos en función de la demanda. No hay límite para la cantidad de servidores a los que puede escalar. Puede escalar de dos servidores a cientos o miles o decenas de miles de servidores casi al instante. Con Auto Scaling, puede asegurarse de que su aplicación siempre

funcione de manera óptima y obtenga potencia adicional en términos de CPU y otros recursos cuando sea necesario. Puede aprovisionarlos en tiempo real.

• **MEJOR EXPERIENCIA DE USUARIO Y RENDIMIENTO:** Auto Scaling ayuda a brindar la mejor experiencia posible a sus usuarios porque nunca se queda sin recursos y su aplicación siempre funciona de manera óptima. Puede crear varias reglas dentro de Auto Scaling para proporcionar la mejor experiencia de usuario. Por ejemplo, puede especificar que, si la utilización de la CPU aumenta a más del 70 por ciento, se inicie una nueva instancia.

• **VERIFICACIÓN DE ESTADO Y ADMINITRACIÓN DE FLOTAS**: Puede monitorear las verificaciones de estado de sus instancias de EC2 mediante Auto Scaling. Si aloja su aplicación en un grupo de servidores EC2, la colección de esos servidores EC2 se denomina flota. Puede configurar verificaciones de estado con Auto Scaling y, si una verificación de estado detecta que hay una falla en una instancia, la reemplaza automáticamente. Le reduce mucha carga porque ahora no tiene que reemplazar manualmente la instancia fallida. También ayuda a mantener la capacidad deseada de la flota. Por ejemplo, si su aplicación se ejecuta en seis servidores EC2, podrá mantener la flota de seis servidores EC2 sin importar cuántas veces haya un problema con un servidor EC2. Alternativamente, si una o más instancias dejan de funcionar, Auto Scaling iniciará servidores adicionales para asegurarse de que siempre tenga seis instancias en ejecución.

• **EQUILIBRIO O BALANCEO DE CARGA**: Dado que Auto Scaling se utiliza para escalar dinámicamente hacia arriba y hacia abajo los recursos, puede encargarse de equilibrar la carga de trabajo en varias instancias de EC2 cuando usa Auto Scaling. Auto Scaling también equilibra automáticamente las instancias EC2 en varias zonas de disponibilidad cuando se configuran varias zonas de disponibilidad. Auto Scaling se asegura de que exista un equilibrio uniforme de instancias EC2 en varias zonas de disponibilidad que defina.

• **SEGUIMIENTO DE OBJETIVOS**: Puede utilizar Auto Scaling para ejecutar un objetivo en particular y Auto Scaling ajusta el número de instancias EC2 para que pueda cumplir con ese objetivo. El objetivo puede ser una métrica de escala compatible con Auto Scaling, por ejemplo, si siempre desea que la utilización de la CPU de su servidor de aplicaciones permanezca en el 65 por ciento, Auto Scaling aumentará y disminuirá la cantidad de instancias EC2 automáticamente para cumplir con la métrica de utilización de la CPU del 65 por ciento.

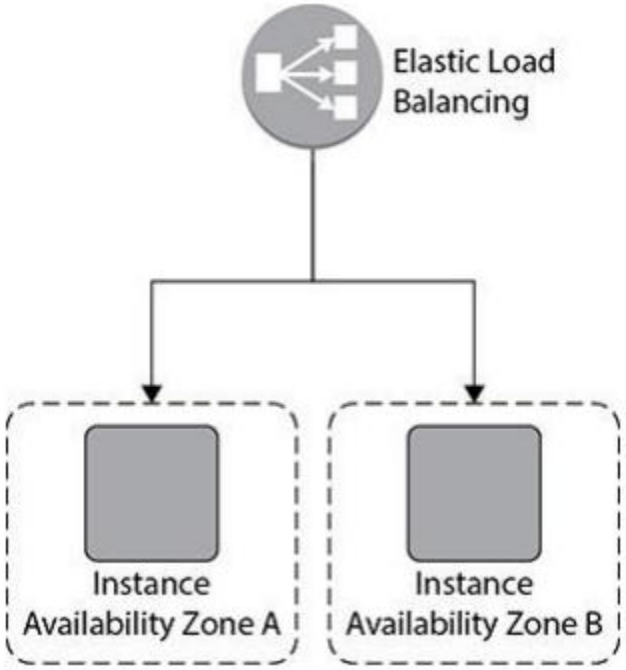
• **CONTROL DE COSTOS**: Con Auto Scaling, también puede eliminar automáticamente los recursos que no necesita para evitar un gasto excesivo. Por ejemplo, por la noche cuando los usuarios se van, Auto Scaling eliminará el exceso de recursos automáticamente y esto ayuda a mantener el presupuesto bajo control.

• **ESCALADO PREDICTIVO**: Auto Scaling ahora está integrado con el aprendizaje automático (ML) y, mediante el uso de ML Auto Scaling, puede escalar automáticamente su capacidad informática por adelantado según el aumento previsto de la demanda. La forma en que funciona es que Auto Scaling recopila los datos de su uso real de EC2 y luego usa los modelos de aprendizaje automático para predecir su tráfico esperado diario y semanal. Los datos se evalúan cada 24 horas para crear un pronóstico para las próximas 48 horas.

No hay ningún cargo adicional por usar Amazon EC2 Auto Scaling.

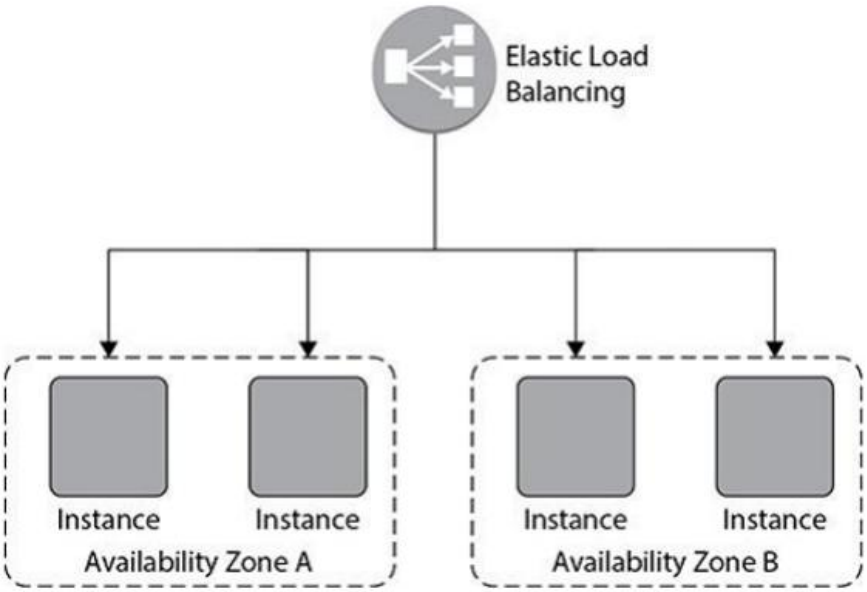
EJEMPLO DE AUTO SCALING

Supongamos que tiene una aplicación que consta de dos servidores web alojados en dos instancias EC2 independientes. Para mantener la alta disponibilidad, ha colocado los servidores web en diferentes zonas de disponibilidad. Ha integrado los servidores web con ELB y los usuarios se conectan al ELB. La arquitectura se parecerá a esta:



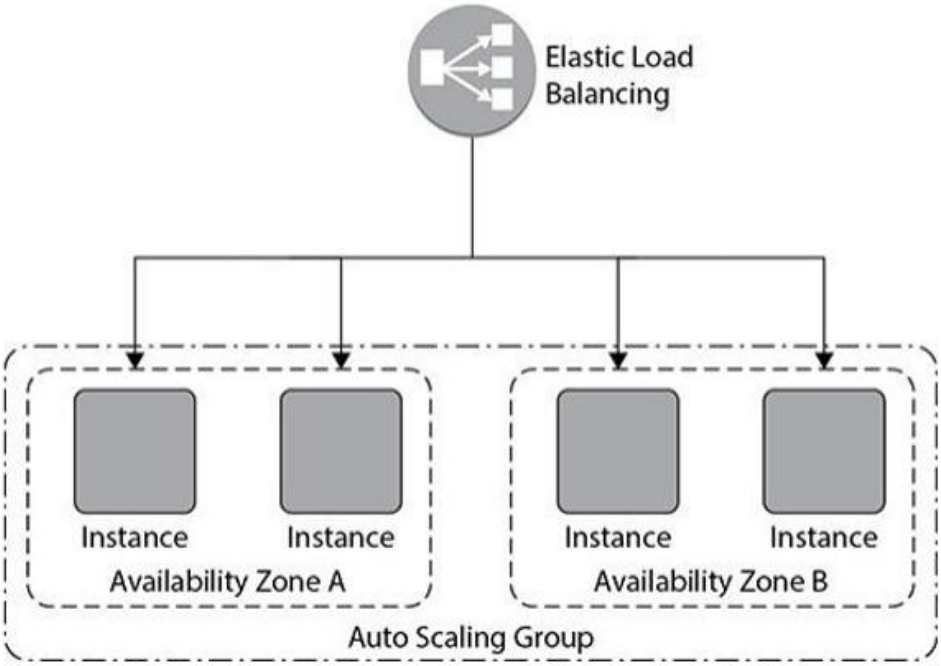
Todo va bien cuando de repente notamos que hay un aumento en el tráfico web. Para satisfacer el tráfico adicional, aprovisionamos dos servidores web adicionales y los integramos con ELB, como se muestra a continuación. Hasta este punto, estamos haciendo todo manualmente, lo que incluye agregar servidores web e integrarlos con ELB. Además, si el tráfico disminuye, debemos eliminar las instancias manualmente, ya que mantenerlas costará más.

Esto está bien y es manejable cuando tenemos una menor cantidad de servidores para administrar y podemos predecir el tráfico. Pero ¿Qué pasa si tenemos cientos o miles de servidores que alojan la aplicación? ¿Qué pasa si el tráfico es totalmente impredecible? ¿Todavía podemos agregar cientos y miles de servidores casi instantáneamente y luego integrar cada uno de ellos con ELB? ¿Qué hay de quitar esos servidores? ¿Podemos hacerlo rápido? Realmente no. Auto Scaling resuelve este problema por nosotros.



Agregar dos servidores web adicionales a la aplicación

Cuando utilizamos Auto Scaling, simplemente agregamos las instancias EC2 a un grupo de Auto Scaling, definimos el número mínimo y máximo de servidores y luego definimos la política de escalado. Auto Scaling se encarga de agregar y eliminar los servidores e integrarlos con ELB según el uso. Cuando integramos Auto Scaling, la arquitectura ahora se parece a esta:



Agregar los cuatro servidores web como parte del auto Scaling

DIRECCIÓN DE TRÁFICO CON ELASTIC LOAD BALANCING

Elastic Load Balancer es el servicio de AWS que distribuye automáticamente el tráfico de aplicaciones entrantes entre varios recursos, como las instancias de Amazon EC2.

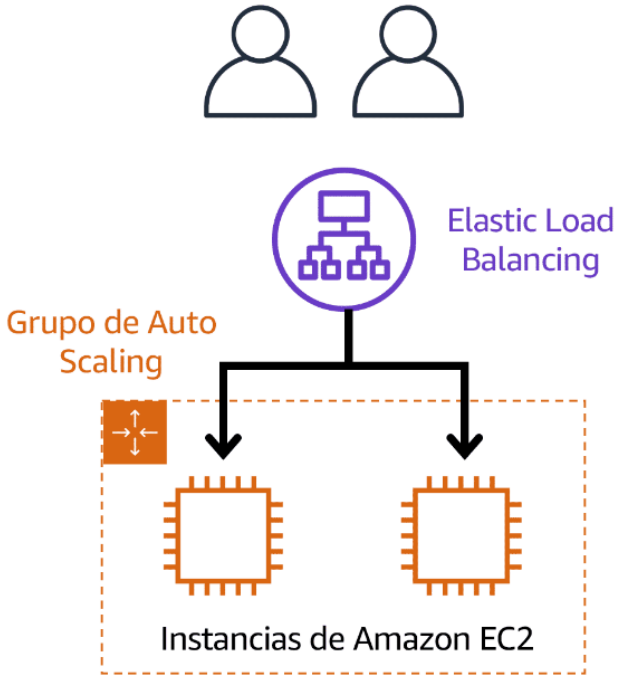
Un equilibrador de carga actúa como un único punto de contacto para todo el tráfico web que entra en el grupo de Auto Scaling. Esto significa que, a medida que agrega o elimina instancias de Amazon EC2 en respuesta a la cantidad de tráfico entrante, estas solicitudes se dirigen primero al equilibrador de carga. A continuación, las solicitudes se distribuyen en varios recursos que se encargarán de gestionarlas. Por ejemplo, si tiene varias instancias de Amazon EC2, Elastic Load Balancer distribuye la carga de trabajo entre las distintas instancias para que ninguna instancia tenga que cargar la mayor parte.

Aunque Elastic Load Balancer y Amazon EC2 Auto Scaling son servicios independientes, funcionan juntos para ayudar a garantizar que las aplicaciones que se ejecutan en Amazon EC2 puedan proporcionar un alto rendimiento y una alta disponibilidad.

EJEMPLO DE ELASTIC LOAD BALANCING

Periodo de baja demanda: Supongamos que algunos clientes están haciendo peticiones a nuestra aplicación.

Si solo hay unas pocas instancias EC2 corriendo, esto coincide con la demanda de los clientes que solicitan el servicio. Es menos probable que la aplicación tenga instancias EC2 corriendo sin peticiones de clientes.



Periodo de alta demanda: A lo largo del día, a medida que aumenta el número de peticiones de clientes, la aplicación lanza más instancias EC2 para atenderlos.
Además, un equilibrador de carga dirige las peticiones de los clientes hacia la instancia EC2 más adecuada para que el número de solicitudes pueda distribuirse uniformemente entre las instancias EC2 activas.