Lectura 4.5: Auto Scaling de Amazon EC2

La disponibilidad y la accesibilidad se mejoran al agregar un servidor más. Sin embargo, todo el sistema puede volver a no estar disponible si hay un problema de capacidad. Veamos ese problema de carga con los dos tipos de sistemas que discutimos, activo-pasivo y activo-activo.

Escala vertical

Si se envían demasiadas solicitudes a un solo sistema activo-pasivo, el servidor activo dejará de estar disponible y, con suerte, se conmutará al servidor pasivo. Pero esto no soluciona nada. Con activo-pasivo, necesita escalado vertical. Esto significa aumentar el tamaño del servidor. Con las instancias EC2, selecciona un tipo más grande o un tipo de instancia diferente. Esto solo se puede hacer mientras la instancia está detenida. En este escenario, se producen los siguientes pasos:

1. Detenga la instancia pasiva. Esto no afecta a la aplicación, ya que no recibe ningún tráfico.
2. Cambie el tamaño o el tipo de la instancia, luego vuelva a iniciar la instancia.
3. Cambie el tráfico a la instancia pasiva, activándola.
4. El último paso es detener, cambiar el tamaño e iniciar la instancia activa anterior, ya que ambas instancias deben coincidir.

Cuando la cantidad de solicitudes se reduce, se debe realizar la misma operación. Aunque no hay tantos pasos involucrados, en realidad es mucho trabajo manual. Otra desventaja es que un servidor solo puede escalar verticalmente hasta cierto límite.

Una vez que se alcanza ese límite, la única opción es crear otro sistema activo-pasivo y dividir las solicitudes y funcionalidades entre ellos. Esto podría requerir una reescritura masiva de la aplicación. Aquí es donde el sistema activo-activo puede ayudar. Cuando hay demasiadas solicitudes, este sistema se puede escalar horizontalmente agregando más servidores.

Escala horizontal

Como se mencionó anteriormente, para que la aplicación funcione en un sistema activo-activo, ya está creada como sin estado, sin almacenar ninguna sesión de cliente en el servidor. Esto significa que tener dos servidores o tener cuatro no requerirá ningún cambio en la aplicación. Solo sería cuestión de crear más instancias cuando sea necesario y cerrarlas cuando el tráfico disminuya.

El servicio Amazon EC2 Auto Scaling puede encargarse de esa tarea creando y eliminando automáticamente instancias EC2 en función de las métricas de Amazon CloudWatch.

Puede ver que hay muchas más ventajas al usar un sistema activo-activo en comparación con un activo-pasivo. La modificación de su aplicación para que no tenga estado permite la escalabilidad.

Integre ELB con EC2 Auto Scaling

El servicio ELB se integra a la perfección con EC2 Auto Scaling. Tan pronto como se agrega o elimina una nueva instancia de EC2 del grupo de EC2 Auto Scaling, se notifica a ELB. Sin embargo, antes de que pueda enviar tráfico a una nueva instancia EC2, debe validar que la aplicación que se ejecuta en esa instancia EC2 está disponible.

Esta validación se realiza a través de la función de controles de salud de ELB. El monitoreo es una parte importante de los balanceadores de carga, ya que debe enrutar el tráfico solo a instancias EC2 saludables. Es por eso que ELB admite dos tipos de controles de salud.

- Establecer una conexión a una instancia EC2 de back-end mediante TCP y marcar la instancia como disponible si la conexión se realiza correctamente.
- Realizar una solicitud HTTP o HTTPS a una página web que especifique y validar que se devuelva un código de respuesta HTTP.

Diferenciar entre escalado tradicional y escalado automático

Con un enfoque tradicional de escalado, usted compra y aprovisiona suficientes servidores para manejar el tráfico en su punto máximo. Sin embargo, esto significa que durante la noche hay más capacidad que tráfico. Esto también significa que estás desperdiciando dinero. Apagar esos servidores por la noche o en momentos en que el tráfico es menor solo ahorra electricidad.

La nube funciona de manera diferente, con un modelo de pago por uso. Es importante desactivar los servicios no utilizados, especialmente las instancias EC2 que paga bajo demanda. Uno podría agregar y eliminar servidores manualmente en un momento previsto. Pero con picos inusuales en el tráfico, esta solución conduce a un desperdicio de recursos con un aprovisionamiento excesivo o con una pérdida de clientes debido a un aprovisionamiento insuficiente.

Aquí se necesita una herramienta que agregue y elimine automáticamente instancias EC2 de acuerdo con las condiciones que usted defina; eso es exactamente lo que hace el servicio EC2 Auto Scaling.

Utilice el escalado automático de Amazon EC2

El servicio EC2 Auto Scaling funciona para agregar o eliminar capacidad para mantener un rendimiento constante y

predecible al menor costo posible. Al ajustar la capacidad exactamente a lo que usa su aplicación, solo paga por lo que su aplicación necesita. E incluso con aplicaciones que tienen un uso constante, EC2 Auto Scaling puede ayudar con la gestión de flotas. Si hay un problema con una instancia EC2, EC2 Auto Scaling puede reemplazar automáticamente esa instancia. Esto significa que EC2 Auto Scaling ayuda tanto a escalar su infraestructura como a garantizar una alta disponibilidad.

Configurar componentes de EC2 Auto Scaling

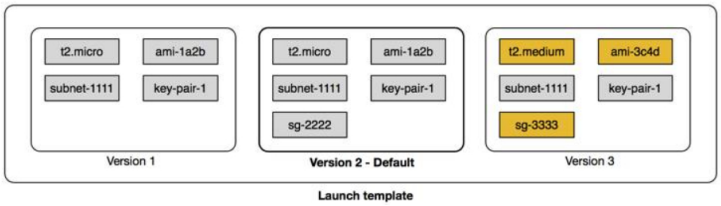
Hay tres componentes principales para EC2 Auto Scaling.

- **Plantilla o configuración de lanzamiento:** ¿Qué recurso se debe escalar automáticamente?
- **Grupo de EC2 Auto Scaling:** ¿Dónde deben implementarse los recursos?
- **Políticas de escalado:** ¿Cuándo deben agregarse o eliminarse los recursos?

Más información sobre las plantillas de lanzamiento

Se requieren varios parámetros para crear instancias EC2: ID de imagen de máquina de Amazon (AMI), tipo de instancia, grupo de seguridad, volúmenes adicionales de Amazon Elastic Block Store (EBS) y más. EC2 Auto Scaling también requiere toda esta información para crear la instancia EC2 en su nombre cuando sea necesario escalar. Esta información se almacena en una plantilla de lanzamiento.

Puede utilizar una plantilla de lanzamiento para lanzar manualmente una instancia EC2. También puede usarlo con EC2 Auto Scaling. También es compatible con el control de versiones, lo que permite retroceder rápidamente si hubo un problema o especificar una versión predeterminada de su plantilla de lanzamiento. De esta forma, mientras iteran en una nueva versión, otros usuarios pueden continuar lanzando instancias EC2 utilizando la versión predeterminada hasta que realice los cambios necesarios.



Puede crear una plantilla de lanzamiento de tres maneras.

- La forma más rápida de crear una plantilla es utilizar una instancia EC2 existente. Todos los ajustes ya están definidos.
- Otra opción es crear uno a partir de una plantilla ya existente o una versión anterior de una plantilla de lanzamiento.
- La última opción es crear una plantilla desde cero. Será necesario definir las siguientes opciones: ID de AMI, tipo de instancia, par de claves, grupo de seguridad, almacenamiento y etiquetas de recursos.

Nota : Otra forma de definir lo que Amazon EC2 Auto Scaling necesita escalar es mediante una **configuración de lanzamiento** . Es similar a la plantilla de lanzamiento, pero no permite el control de versiones usando una configuración de lanzamiento creada previamente como plantilla. Tampoco permite crear uno a partir de una instancia EC2 ya existente. Por estos motivos y para asegurarse de obtener las funciones más recientes de Amazon EC2, utilice una plantilla de lanzamiento en lugar de una configuración de lanzamiento.

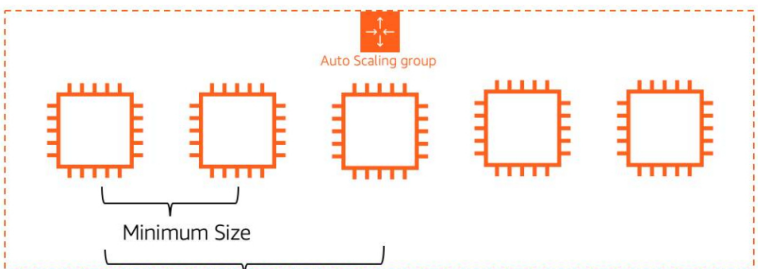
Conozca los grupos de Auto Scaling de EC2

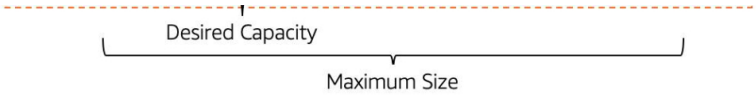
El siguiente componente que EC2 Auto Scaling necesita es un EC2 Auto Scaling Group (ASG). Un ASG le permite definir dónde EC2 Auto Scaling implementa sus recursos. Aquí es donde se especifica la nube privada virtual (VPC) de Amazon y las subredes en las que se debe lanzar la instancia EC2.

EC2 Auto Scaling se encarga de crear las instancias EC2 en las subredes, por lo que es importante seleccionar al menos dos subredes que se encuentren en diferentes zonas de disponibilidad.

Los ASG también le permiten especificar el tipo de compra para las instancias EC2. Puede usar solo bajo demanda, solo de spot o una combinación de ambos, lo que le permite aprovechar las instancias de spot con una sobrecarga administrativa mínima. Para especificar cuántas instancias debe lanzar EC2 Auto Scaling, hay tres configuraciones de capacidad para configurar por el tamaño del grupo.

- **Mínimo:** la cantidad mínima de instancias que se ejecutan en su ASG incluso si se alcanza el umbral para reducir la cantidad de instancias.
- **Máximo:** la cantidad máxima de instancias que se ejecutan en su ASG incluso si se alcanza el umbral para agregar nuevas instancias.
- **Capacidad deseada:** la cantidad de instancias que debe haber en su ASG. Este número solo puede estar dentro o igual al mínimo o al máximo. EC2 Auto Scaling agrega o elimina automáticamente instancias para que coincidan con el número de capacidad deseado.



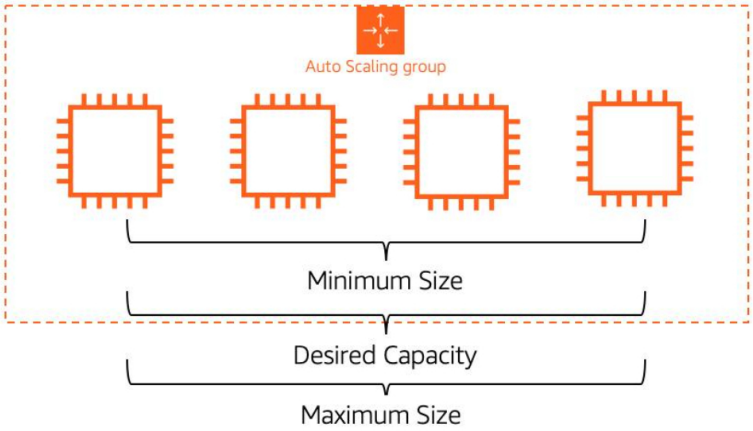


Cuando EC2 Auto Scaling elimina instancias EC2 porque el tráfico es mínimo, sigue eliminando instancias EC2 hasta que alcanza una capacidad mínima. Dependiendo de su aplicación, usar un mínimo de dos es una buena idea para garantizar una alta disponibilidad, pero sabe cuántas instancias EC2 como mínimo requiere su aplicación en todo momento. Al llegar a ese límite, incluso si se le indica a EC2 Auto Scaling que elimine una instancia, no lo hace, para garantizar que se mantenga el mínimo.

Por otro lado, cuando el tráfico sigue creciendo, EC2 Auto Scaling sigue agregando instancias EC2. Esto significa que el costo de su aplicación también seguirá creciendo. Por eso es importante establecer una cantidad máxima para asegurarse de que no supere su presupuesto.

La capacidad deseada es la cantidad de instancias EC2 que EC2 Auto Scaling crea en el momento en que se crea el grupo. Si ese número disminuye, EC2 Auto Scaling elimina la instancia más antigua de forma predeterminada. Si ese número aumenta, EC2 Auto Scaling crea nuevas instancias utilizando la plantilla de lanzamiento.

Garantice la disponibilidad con EC2 Auto Scaling



El uso de diferentes números para la capacidad mínima, máxima y deseada se usa para ajustar dinámicamente la capacidad. Sin embargo, si prefiere utilizar EC2 Auto Scaling para la gestión de flotas, puede configurar los tres ajustes en el mismo número, por ejemplo, cuatro. EC2 Auto Scaling se asegurará de que, si una instancia EC2 se vuelve inestable, la reemplace para garantizar que siempre haya cuatro instancias EC2 disponibles. Esto garantiza una alta disponibilidad para sus aplicaciones.

Habilite la automatización con políticas de escalado

De forma predeterminada, un ASG se mantendrá en su capacidad inicial deseada. Aunque es posible cambiar manualmente la capacidad deseada, también puede usar políticas de escalado.

En el módulo de monitoreo de AWS, aprendió sobre las métricas y alarmas de Amazon CloudWatch. Utiliza **métricas** para mantener información sobre diferentes atributos de su instancia EC2, como el porcentaje de CPU. **Las alarmas** se utilizan para especificar una acción cuando se alcanza un umbral. Las métricas y las alarmas son lo que usan las políticas de escalado para saber cuándo actuar. Por ejemplo, configura una alarma que dice que cuando la utilización de la CPU supera el 70 % en toda la flota de instancias EC2, active una política de escalado para agregar una instancia EC2.

Hay tres tipos de políticas de escalado: escalado simple, por pasos y de seguimiento de destino.

Política de escalado simple

Una política de escalado simple le permite hacer exactamente lo que se describe arriba. Utiliza una alarma de CloudWatch y especifica qué hacer cuando se activa. Esto puede ser una cantidad de instancias EC2 para agregar o eliminar, o una cantidad específica para establecer la capacidad deseada. Puede especificar un porcentaje del grupo en lugar de usar una cantidad de instancias EC2, lo que hace que el grupo crezca o se reduzca más rápidamente.

Una vez que se activa esta política de escalado, espera un período de recuperación antes de realizar cualquier otra acción. Esto es importante ya que las instancias EC2 tardan en iniciarse y es posible que la alarma de CloudWatch aún se active mientras se inicia la instancia EC2. Por ejemplo, podría decidir agregar una instancia EC2 si la utilización de la CPU en todas las instancias es superior al 65 %. No desea agregar más instancias hasta que la nueva instancia EC2 acepte tráfico.

Sin embargo, ¿qué pasaría si la utilización de la CPU estuviera ahora por encima del 85 % en el ASG? Solo agregar una instancia puede no ser el movimiento correcto aquí. En su lugar, es posible que desee agregar otro paso en su política de escalado. Desafortunadamente, una política de escalado simple no puede ayudar con eso.

Política de escalamiento escalonado

Aquí es donde ayuda una política de escalamiento escalonado. Las políticas de escalado por pasos responden a alarmas adicionales incluso mientras se está realizando una actividad de escalado o un reemplazo de verificación de estado. De manera similar al ejemplo anterior, decide agregar dos instancias más en caso de que la utilización de la CPU sea del 85 % y cuatro instancias más cuando sea del 95 %.

Decidir cuándo agregar y quitar instancias según las alarmas de CloudWatch puede parecer una tarea difícil. Por eso existe el tercer tipo de política de escalado: el seguimiento de objetivos.

Política de escalado de seguimiento de objetivos

Si su aplicación se escala según el uso promedio de la CPU, el uso promedio de la red (dentro o fuera) o según el recuento de solicitudes, entonces este tipo de política de escalamiento es el que debe usar. Todo lo que necesita proporcionar es el valor objetivo para rastrear y crea automáticamente las alarmas de CloudWatch requeridas.

Recursos

- *Sitio externo:* AWS: Amazon EC2 Auto Scaling
- *Sitio externo:* AWS: Preguntas frecuentes sobre Amazon EC2 Auto Scaling
- *Sitio externo:* AWS: configuración de límites de capacidad para su grupo de Auto Scaling
- *Sitio externo:* AWS: políticas de escalado sencillas y por pasos para Amazon EC2 Auto Scaling
- *Sitio externo:* AWS: Políticas de escalado de seguimiento de destino para Amazon EC2 Auto Scaling
- *Sitio externo:* AWS: creación de un grupo de Auto Scaling mediante una plantilla de lanzamiento