Lectura 4.4: Enrutar el tráfico con Amazon Elastic Load Balancing

¿QUÉ ES UN EQUILIBRADOR DE CARGA?

El equilibrio de carga se refiere al proceso de distribución de tareas entre un conjunto de recursos. En el caso de la aplicación de directorio corporativo, los recursos son instancias EC2 que alojan la aplicación y las tareas son las diferentes solicitudes que se envían. Es hora de distribuir las solicitudes entre todos los servidores que alojan la aplicación mediante un equilibrador de carga.

Para hacer esto, primero debe habilitar el balanceador de carga para que tome todo el tráfico y lo redirija a los servidores back-end según un algoritmo. El algoritmo más popular es el round-robin, que envía el tráfico a cada servidor uno tras otro.

Una solicitud típica de la aplicación comenzaría desde el navegador del cliente. Se envía a un equilibrador de carga. Luego, se envía a una de las instancias EC2 que aloja la aplicación. El tráfico de retorno pasaría por el balanceador de carga y volvería al navegador del cliente. Por lo tanto, el balanceador de carga está directamente en la ruta del tráfico.

Aunque es posible instalar su propia solución de balanceo de carga de software en instancias EC2, AWS proporciona un servicio para eso llamado Elastic Load Balancing (ELB).

CARACTERÍSTICAS DE ELB

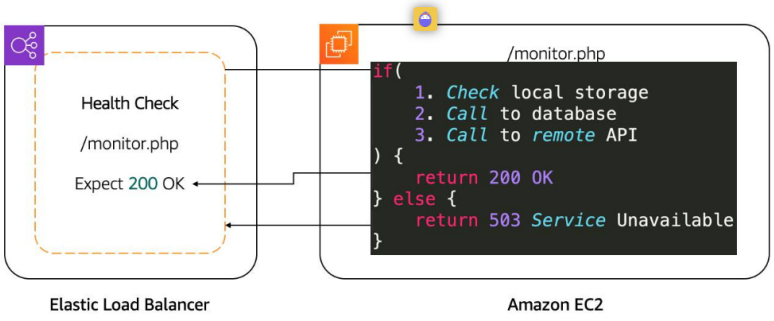
El servicio ELB ofrece una gran ventaja sobre el uso de su propia solución para equilibrar la carga, ya que no necesita administrarlo ni operarlo. Puede distribuir el tráfico de aplicaciones entrantes a través de instancias EC2, así como contenedores, direcciones IP y funciones de AWS Lambda.

- El hecho de que ELB pueda equilibrar la carga en las direcciones IP significa que también puede funcionar en un modo híbrido, donde también equilibra la carga en los servidores locales.
- ELB es altamente disponible. La única opción que debe asegurarse es que el balanceador de carga se implemente en varias zonas de disponibilidad.
- En términos de escalabilidad, ELB escala automáticamente para satisfacer la demanda del tráfico entrante. Maneja el tráfico entrante y lo envía a su aplicación de back-end.

CONTROLES DE SALUD

Tomarse el tiempo para definir un control de salud apropiado es fundamental. Solo verificar que el puerto de una aplicación esté abierto no significa que la aplicación esté funcionando. Tampoco significa que simplemente hacer una llamada a la página de inicio de una aplicación sea la forma correcta.

Por ejemplo, la aplicación de directorio de empleados depende de una base de datos y S3. El control de salud debe validar todos esos elementos. Una forma de hacerlo sería crear una página web de monitoreo como "/monitor" que hará una llamada a la base de datos para asegurarse de que pueda conectarse y obtener datos, y hacer una llamada a S3. Luego, dirige la comprobación de estado del equilibrador de carga a la página "/monitor".

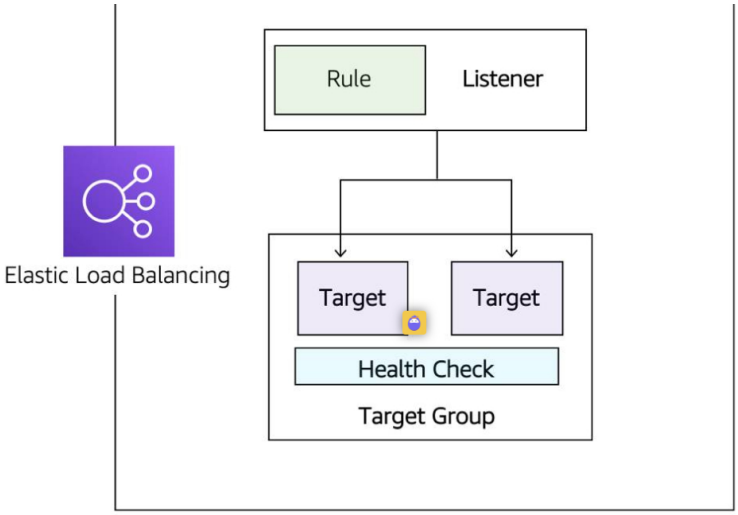


Después de determinar la disponibilidad de una nueva instancia EC2, el balanceador de carga comienza a enviarle tráfico. Si ELB determina que una instancia EC2 ya no funciona, deja de enviarle tráfico y se lo comunica a EC2 Auto Scaling. La responsabilidad de EC2 Auto Scaling es eliminarlo del grupo y reemplazarlo con una nueva instancia de EC2. El tráfico solo se envía a la nueva instancia si pasa la verificación de estado.

En el caso de una acción de reducción que EC2 Auto Scaling deba realizar debido a una política de escalado, le informa a ELB que las instancias de EC2 se cancelarán. ELB puede evitar que EC2 Auto Scaling finalice la instancia EC2 hasta que finalicen todas las conexiones a esa instancia, al tiempo que evita cualquier conexión nueva. Esa característica se llama **drenaje de conexión** .

COMPONENTES ELB

El servicio ELB se compone de tres componentes principales.



- **Oyentes:** El cliente se conecta al oyente. Esto a menudo se denomina del lado del cliente. Para definir un agente de escucha, se debe proporcionar un puerto, así como el protocolo, según el tipo de equilibrador de carga. Puede haber muchos oyentes para un solo balanceador de carga.
- **Grupos objetivo:** los servidores backend, o del lado del servidor, se definen en uno o más grupos objetivo. Aquí es donde define el tipo de backend al que desea dirigir el tráfico, como instancias EC2, funciones de AWS Lambda o direcciones IP. Además, se debe definir un control de salud para cada grupo objetivo.
- **Reglas:** para asociar un grupo objetivo a un oyente, se debe usar una regla. Las reglas se componen de una condición que puede ser la dirección IP de origen del cliente y una condición para decidir a qué grupo objetivo enviar el tráfico.

EQUILIBRADOR DE CARGA DE APLICACIONES

Estas son algunas de las características principales de Application Load Balancer (ALB).

ALB enruta el tráfico en función de los datos de solicitud. Toma decisiones de enrutamiento basadas en el protocolo HTTP, como la ruta de la URL (/upload) y el host, los encabezados y el método HTTP, así como la dirección IP de origen del cliente. Esto permite el enrutamiento granular a los grupos objetivo.

Enviar respuestas directamente al cliente. ALB tiene la capacidad de responder directamente al cliente con una respuesta fija como una página HTML personalizada. También tiene la capacidad de enviar una redirección al cliente, lo cual es útil cuando necesita redirigir a un sitio web específico o redirigir la solicitud de HTTP a HTTPS, eliminando ese trabajo de sus servidores backend.

ALB admite la descarga de TLS. Hablando de HTTPS y ahorrando trabajo de los servidores back-end, ALB entiende el tráfico HTTPS. Para poder pasar el tráfico HTTPS a través de ALB, se proporciona un certificado SSL importando un certificado a través de los servicios Identity and Access Management (IAM) o AWS Certificate Manager (ACM), o creando uno de forma gratuita con ACM. Esto garantiza que el tráfico entre el cliente y el ALB esté cifrado.

Autenticar usuarios. En cuanto al tema de la seguridad, ALB tiene la capacidad de autenticar a los usuarios antes de que se les permita pasar por el balanceador de carga. ALB utiliza el protocolo OpenID Connect y se integra con otros servicios de AWS para admitir más protocolos como SAML, LDAP, Microsoft AD y más.

Tráfico seguro. Para evitar que el tráfico llegue al equilibrador de carga, configure un grupo de seguridad para especificar los intervalos de direcciones IP admitidos.

ALB utiliza el algoritmo de enrutamiento por turnos. ALB garantiza que cada servidor reciba la misma cantidad de solicitudes en general. Este tipo de enrutamiento funciona para la mayoría de las aplicaciones.

ALB utiliza el algoritmo de enrutamiento de solicitudes menos sobresaliente. Si las solicitudes al back-end varían en complejidad, donde una solicitud puede necesitar mucho más tiempo de CPU que otra, entonces el algoritmo de solicitud menos sobresaliente es más apropiado. También es el algoritmo de enrutamiento correcto para usar si los objetivos varían en las capacidades de procesamiento. Una solicitud pendiente es cuando se envía una solicitud al servidor backend y aún no se ha recibido una respuesta.

Por ejemplo, si las instancias EC2 de un grupo de destino no tienen el mismo tamaño, la utilización de la CPU de un servidor será mayor que la del otro si se envía la misma cantidad de solicitudes a cada servidor mediante el algoritmo de enrutamiento por turnos. Ese mismo servidor también tendrá más solicitudes pendientes. El uso del algoritmo de enrutamiento de solicitudes menos sobresaliente garantizaría un uso equitativo entre los objetivos.

ALB tiene sesiones pegajosas. En el caso de que las solicitudes deban enviarse al mismo servidor back-end porque la aplicación tiene estado, utilice la función de sesión permanente. Esta característica utiliza una cookie HTTP para recordar a través de las conexiones a qué servidor enviar el tráfico. Finalmente, ALB es específicamente para el tráfico HTTP y HTTPS. Si su aplicación utiliza un protocolo diferente, considere el Network Load Balancer (NLB).

EQUILIBRADOR DE CARGA DE RED

Estas son algunas de las características principales de Network Load Balancer (NLB). **Network Load Balancer admite los protocolos TCP, UDP y TLS.** HTTPS utiliza TCP y TLS como protocolo. Sin embargo, NLB opera en la capa de conexión, por lo que no comprende qué es una solicitud HTTPS. Eso significa que todas las funciones mencionadas anteriormente que se requieren para comprender el protocolo HTTP y HTTPS, como las reglas de enrutamiento basadas en ese protocolo, la autenticación y el algoritmo de enrutamiento de solicitudes menos sobresalientes, no están disponibles con NLB.

NLB utiliza un algoritmo de enrutamiento hash de flujo. El algoritmo se basa en:

- El protocolo
- La dirección IP de origen y el puerto de origen
- La dirección IP de destino y el puerto de destino
- El número de secuencia TCP

Si todos estos parámetros son iguales, los paquetes se envían exactamente al mismo destino. Si alguno de ellos es diferente en los siguientes paquetes, entonces la solicitud puede enviarse a un objetivo diferente.

NLB tiene sesiones pegajosas. A diferencia de ALB, estas sesiones se basan en la dirección IP de origen del cliente en lugar de una cookie.

NLB admite la descarga de TLS. NLB entiende el protocolo TLS. También puede descargar TLS de los servidores back-end de forma similar a como funciona ALB.

NLB maneja millones de solicitudes por segundo. Si bien ALB también puede admitir esta cantidad de solicitudes, debe escalar para alcanzar esa cantidad. Esto lleva tiempo. NLB puede manejar instantáneamente esta cantidad de solicitudes.

NLB admite direcciones IP estáticas y elásticas. Hay algunas situaciones en las que el cliente de la aplicación necesita enviar solicitudes directamente a la dirección IP del balanceador de carga en lugar de usar DNS. Por ejemplo, esto es útil si su aplicación no puede usar DNS o si los clientes que se conectan requieren reglas de firewall basadas en direcciones IP. En este caso, NLB es el tipo correcto de balanceador de carga para usar.

NLP conserva la dirección IP de origen. NLB conserva la dirección IP de origen del cliente cuando envía el tráfico al backend. Con ALB, si observa la dirección IP de origen de las solicitudes, encontrará la dirección IP del balanceador de carga. Mientras esté con NLB, verá la dirección IP real del cliente, que en algunos casos es requerida por la aplicación de back-end.

SELECCIONE ENTRE TIPOS DE ELB

La selección entre los tipos de servicio ELB se realiza determinando qué característica se requiere para su aplicación. A continuación puede encontrar una lista de las funciones principales que aprendió en esta unidad y en la anterior.

Característica	Equilibrador de carga de aplicaciones	Equilibrador de carga de red
protocolos	HTTP, HTTPS	TCP, UDP, TLS
Drenaje de conexión (retraso de baja)	✓	
Direcciones IP como objetivos	✓	✓
IP estática y dirección IP elástica		✓
Conservar la dirección IP de origen		✓
Enrutamiento basado en la dirección IP de origen, la ruta, el host, los encabezados HTTP, el método HTTP y la cadena de consulta	✓	
Redirecciones	✓	
Respuesta fija	✓	
Autenticacion de usuario	✓	

Recursos :

- *Sitio externo:* AWS: Comparación de productos de Elastic Load Balancer
- *Sitio externo:* AWS: Administrador de certificados de AWS
- *Sitio externo:* AWS: autentica a los usuarios mediante un Balanceador de carga de aplicaciones
- *Sitio externo:* AWS: Cómo funciona AWS WAF

Marcar como completo