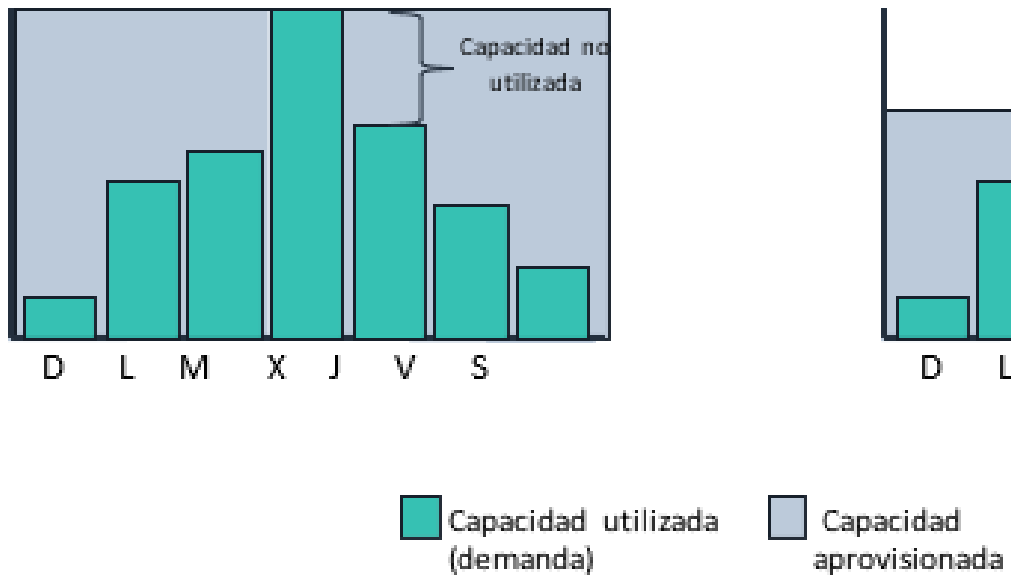
LECCIÓN 3:

AMAZON EC2

AUTO SCALING

Cuando ejecute sus aplicaciones en AWS, es conveniente que se asegure de que su arquitectura pueda escalar a fin de gestionar los cambios en la demanda. En esta Lección, aprenderá a escalar de forma automática sus instancias EC2 con Amazon EC2 Auto Scaling.

¿POR QUÉ ES IMPORTANTE EL ESCALADO?



El escalado es la capacidad de aumentar o reducir la capacidad de cómputo de una aplicación. Para comprender por qué es importante el escalado, considere este ejemplo de una carga de trabajo que tiene diferentes requisitos de recursos. En este ejemplo, la mayor capacidad de recursos se requiere el miércoles y, la menor capacidad de recursos, los domingos.

Una opción consiste en asignar más capacidad de la necesaria de manera de poder satisfacer siempre la demanda más alta, que, en este caso, es la de los miércoles. Sin embargo, esta situación implica que está ejecutando recursos que no se utilizarán por completo la mayoría de los días de la semana. Con esta opción, los costos no están optimizados.

Otra opción consiste en asignar menos capacidad para reducir los costos. Esta situación implica que ciertos días no dispondrá de la capacidad suficiente. Si no resuelve el problema de la capacidad, la aplicación podría tener un rendimiento inferior o, incluso, podría dejar de estar disponible para los usuarios.

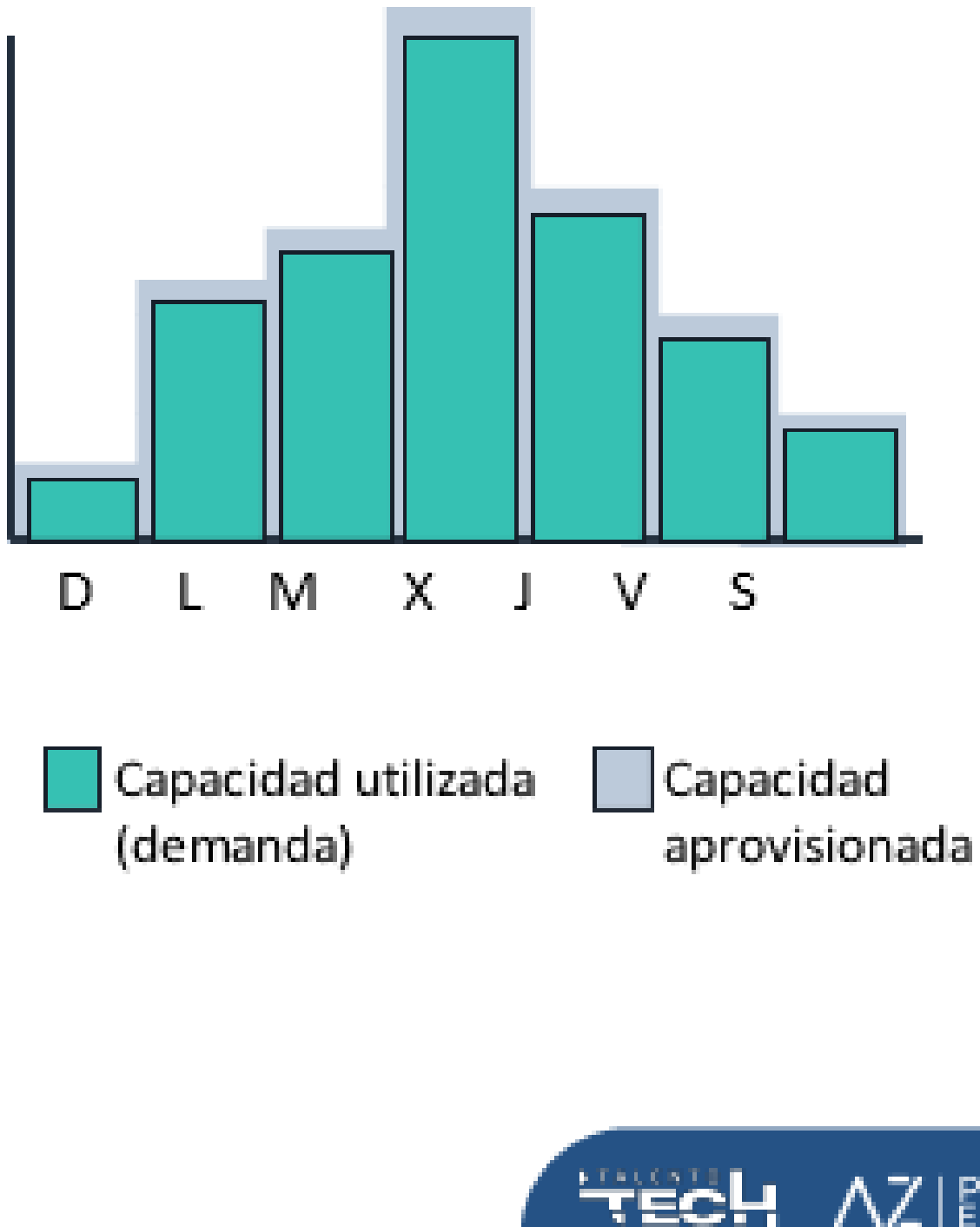
AMAZON EC2 AUTO SCALING

- **Lo ayuda a mantener la disponibilidad de las aplicaciones**

- **Le permite agregar o eliminar automáticamente instancias EC2 de acuerdo con las condiciones que defina**

- **Detecta las instancias EC2 dañadas y las aplicaciones en mal estado, y reemplaza las instancias sin su intervención**

- **Ofrece varias opciones de escalado: manual, programado, dinámico o bajo demanda y predictivo**

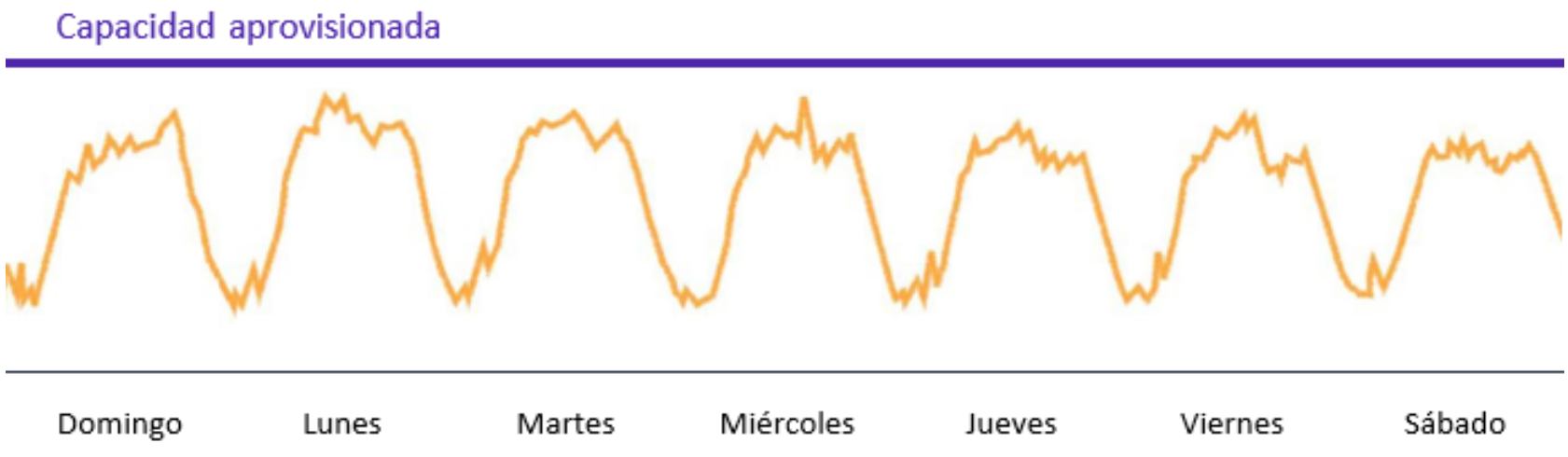


En la nube, como la capacidad de cómputo es un recurso programático, es posible enfocar el tema del escalado de una manera flexible. Amazon EC2 Auto Scaling es un servicio de AWS que lo ayuda a mantener la disponibilidad de la aplicación y le permite agregar o eliminar instancias EC2 de forma automática según las condiciones que defina. Puede utilizar las características de administración de flotas de EC2 Auto Scaling para mantener el estado y la disponibilidad de la suya.

Amazon EC2 Auto Scaling ofrece varias formas de ajustar el escalado para satisfacer las necesidades de sus aplicaciones de la mejor manera. Puede agregar o eliminar instancias EC2 manualmente, según una programación, en respuesta a los cambios en la demanda o en combinación con AWS Auto Scaling para el escalado predictivo. El escalado dinámico y el escalado predictivo se pueden utilizar juntos para escalar con mayor rapidez.

Para obtener más información acerca de Amazon EC2 Auto Scaling, consulte la página de producto de Amazon EC2 Auto Scaling.

TRÁFICO SEMANAL HABITUAL EN AMAZON.COM

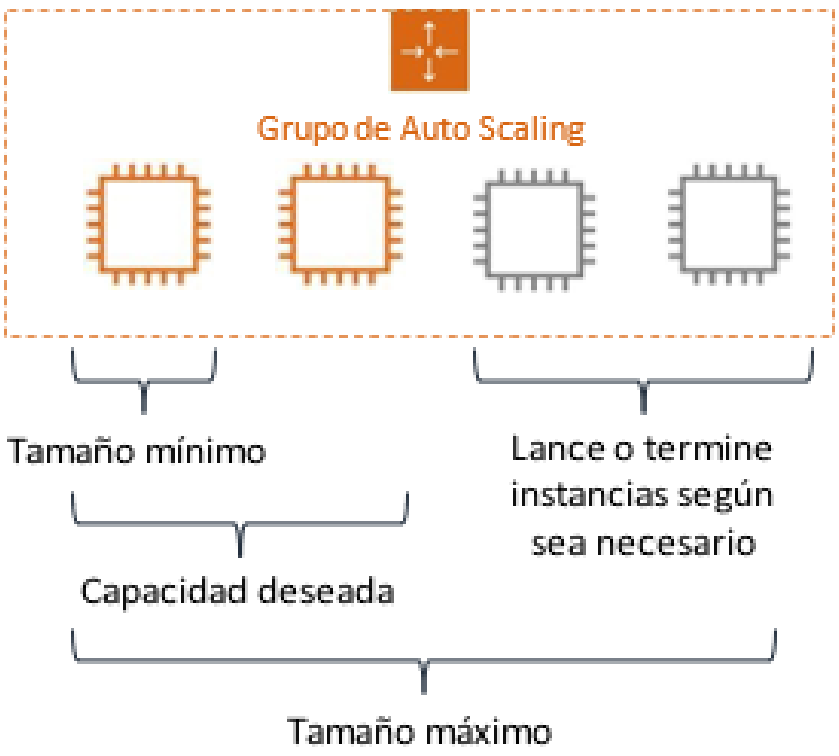


El escalado automático resulta útil para las cargas de trabajo predecibles, como, por ejemplo, el tráfico semanal en la empresa de venta minorista Amazon.com.

TRÁFICO QUE LLEGA A AMAZON.COM EN NOVIEMBRE

El escalado automático también es útil para el escalado dinámico bajo demanda. Amazon.com experimenta un pico de tráfico estacional en noviembre (el Black Friday y el Cyber Monday, que son días a finales de noviembre en los que los vendedores minoristas estadounidenses tienen ventas importantes). Si Amazon aprovisiona una capacidad fija que se ajuste al uso máximo, no se utilizará el 76 % de los recursos durante la mayor parte del año. El escalado de la capacidad es necesario para admitir las demandas de servicio fluctuantes. Sin el escalado, los servidores podrían bloquearse debido a la saturación y la empresa perdería la confianza de los clientes.

GRUPOS DE AUTO SCALING



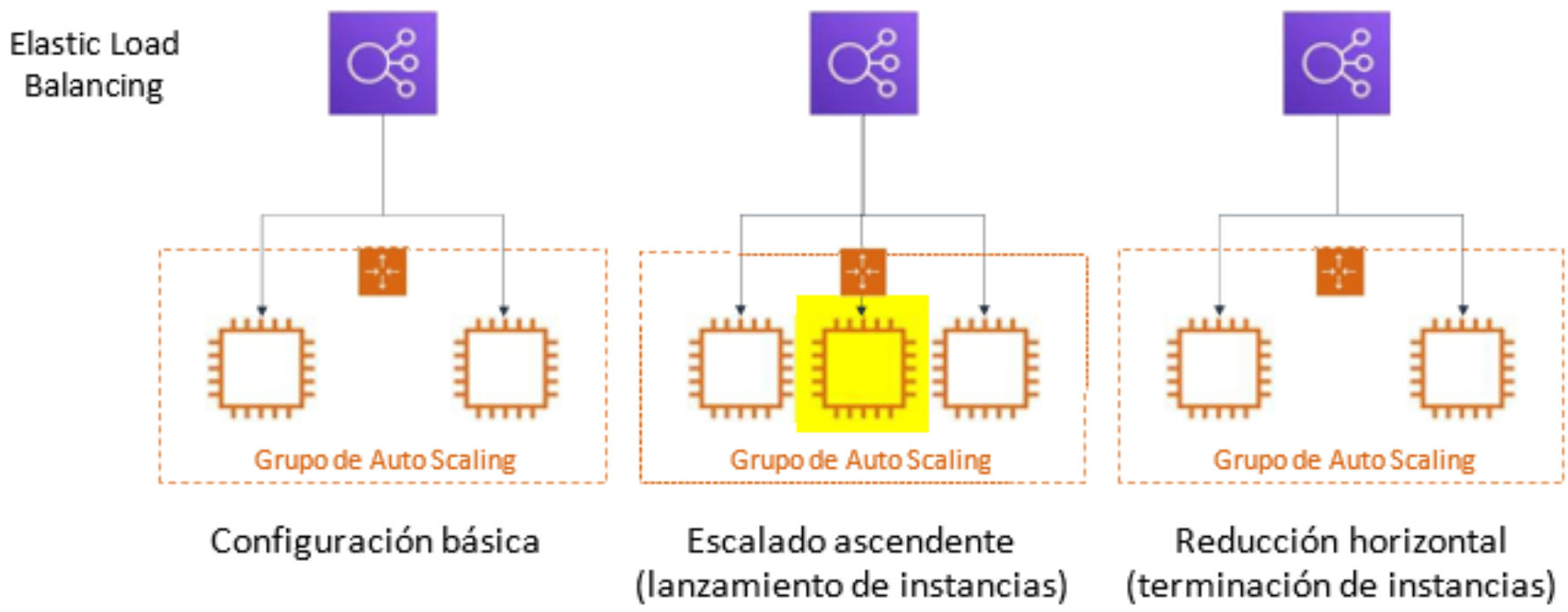
Un grupo de Auto Scaling consiste en una colección de instancias EC2 que se tratan como una agrupación lógica a efectos de la administración y el escalado automático.

Un grupo de Auto Scaling es una colección de instancias de Amazon EC2 que se tratan como una agrupación lógica a efectos de la administración y el escalado automático. El tamaño de un grupo de Auto Scaling depende de la cantidad de instancias que se establecen como la capacidad deseada. Puede ajustar su tamaño para satisfacer la demanda, ya sea de forma manual o mediante el escalado automático.

Puede especificar la cantidad mínima de instancias en cada grupo de Auto Scaling y, al estar diseñado para ello, Amazon EC2 Auto Scaling evitará que su grupo se reduzca por debajo de este tamaño. Puede especificar la cantidad máxima de instancias en cada grupo de
Auto Scaling y, al estar diseñado para ello, Amazon EC2 Auto Scaling evitará que su grupo supere este tamaño. Si especifica la capacidad deseada, ya sea al crear el grupo o en cualquier momento posterior, Amazon EC2 Auto Scaling está diseñado para ajustar el tamaño de su grupo de manera que tenga la cantidad especificada de instancias. Si especifica las políticas de escalado, Amazon EC2 Auto Scaling puede lanzar o terminar instancias en función de los aumentos o las disminuciones en las demandas de la aplicación.

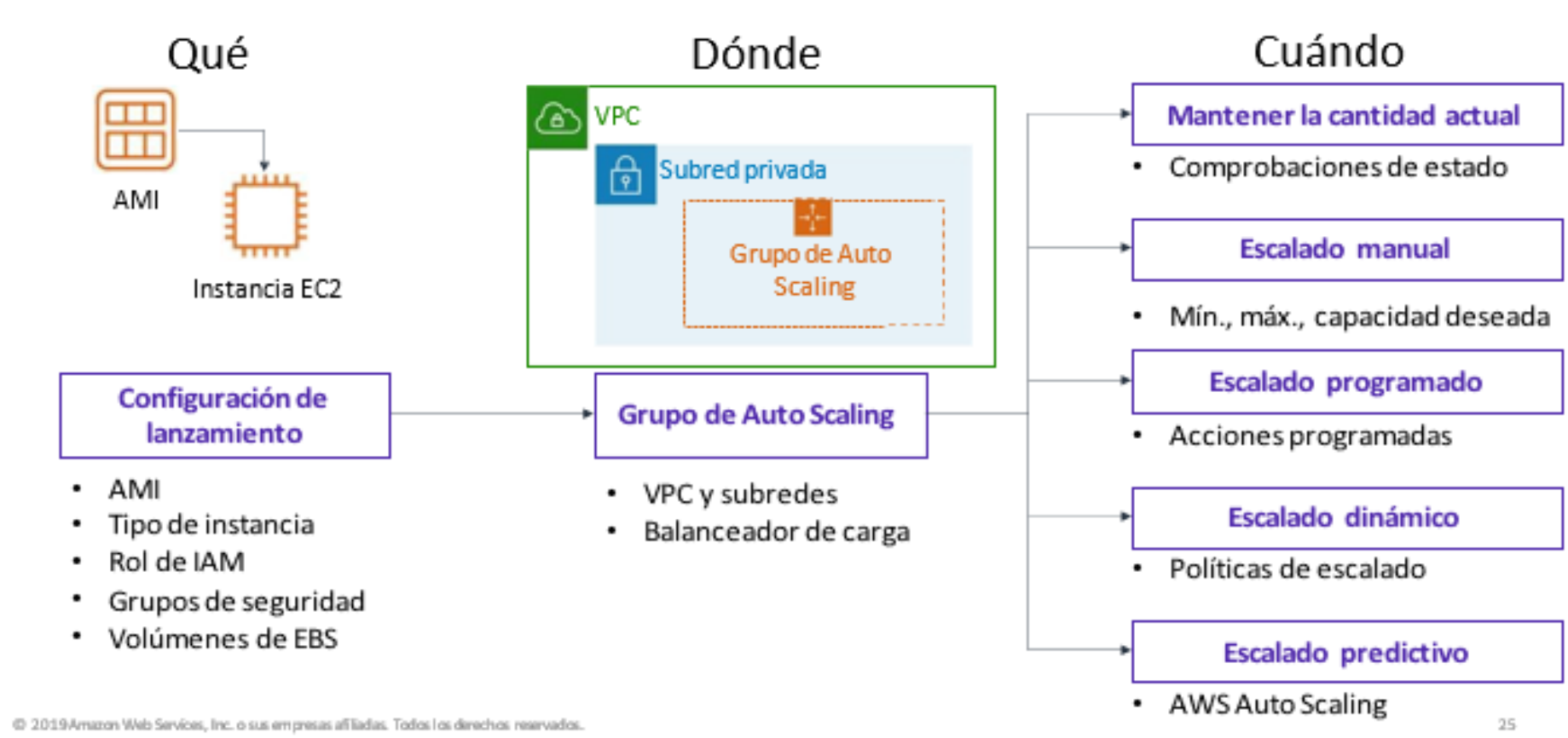
Por ejemplo, este grupo de Auto Scaling tiene un tamaño mínimo de una instancia, una capacidad deseada de dos instancias y un tamaño máximo de cuatro instancias. Las políticas de escalado que defina ajustan la cantidad de instancias dentro del número mínimo y máximo de instancias, en función de los criterios que especifique.

COMPARACIÓN ENTRE EL ESCALADO ASCENDENTE Y EL ESCALADO DESCENDENTE



Con Amazon EC2 Auto Scaling, el lanzamiento de instancias se denomina escalado ascendente y la terminación de instancias se denomina escalado descendente.

CÓMO FUNCIONA AMAZON EC2 AUTO SCALING



Para lanzar instancias EC2, un grupo de Auto Scaling utiliza una configuración de lanzamiento, que es una plantilla de configuración de instancias. Puede considerar la configuración de lanzamiento como qué es lo que está escalando. Cuando se crea una configuración de lanzamiento, se especifica información sobre las instancias. La información que especifica incluye el ID de la Imagen de Amazon Machine (AMI), el tipo de instancia, el rol de AWS Identity and Access Management (IAM), el almacenamiento adicional, uno o más grupos de seguridad y cualquier volumen de Amazon Elastic Block Store (Amazon EBS).

Usted define las cantidades mínima y máxima de instancias y la capacidad deseada de su grupo de Auto Scaling. A continuación, lo lanza a una subred dentro de una VPC (puede considerar esto como dónde está escalando). Amazon EC2 Auto Scaling se integra con Elastic Load Balancing a fin de permitirle asociar uno o más balanceadores de carga a un grupo de Auto Scaling existente. Tras asociar el balanceador de carga, registra automáticamente las instancias del grupo y distribuye el tráfico entrante entre las instancias.

Por último, debe especificar cuándo desea que se produzca el evento de escalado. Existen muchas opciones de escalado:

Mantener los niveles actuales de la instancia en todo momento:

puede configurar su grupo de Auto Scaling para mantener un número específico de instancias en ejecución en todo momento. Para mantener los niveles de instancia actuales, Amazon EC2 Auto Scaling efectúa una comprobación de estado periódica de las instancias en ejecución dentro un grupo de Auto Scaling. Cuando Amazon EC2 Auto Scaling encuentra una instancia en mal estado, la termina y lanza una nueva.

Escalado manual:

con el escalado manual solo debe especificar el cambio en la capacidad máxima, mínima o deseada de su grupo de Auto Scaling.

Escalado programado:

Con el escalado programado, las acciones de escalado se efectúan automáticamente en función de la fecha y la hora. Esto resulta útil para las cargas de trabajo predecibles cuando sabe exactamente cuándo aumentar o reducir la cantidad de instancias de su grupo. Por ejemplo, consideremos que cada semana, el tráfico de su aplicación web comienza a aumentar el miércoles, permanece alto el jueves y comienza a disminuir el viernes. Puede programar las acciones de escalado en función de los patrones de tráfico predecibles de su aplicación web. Para implementar el escalado programado,cree una acción programada.

Escalado dinámico bajo demanda:

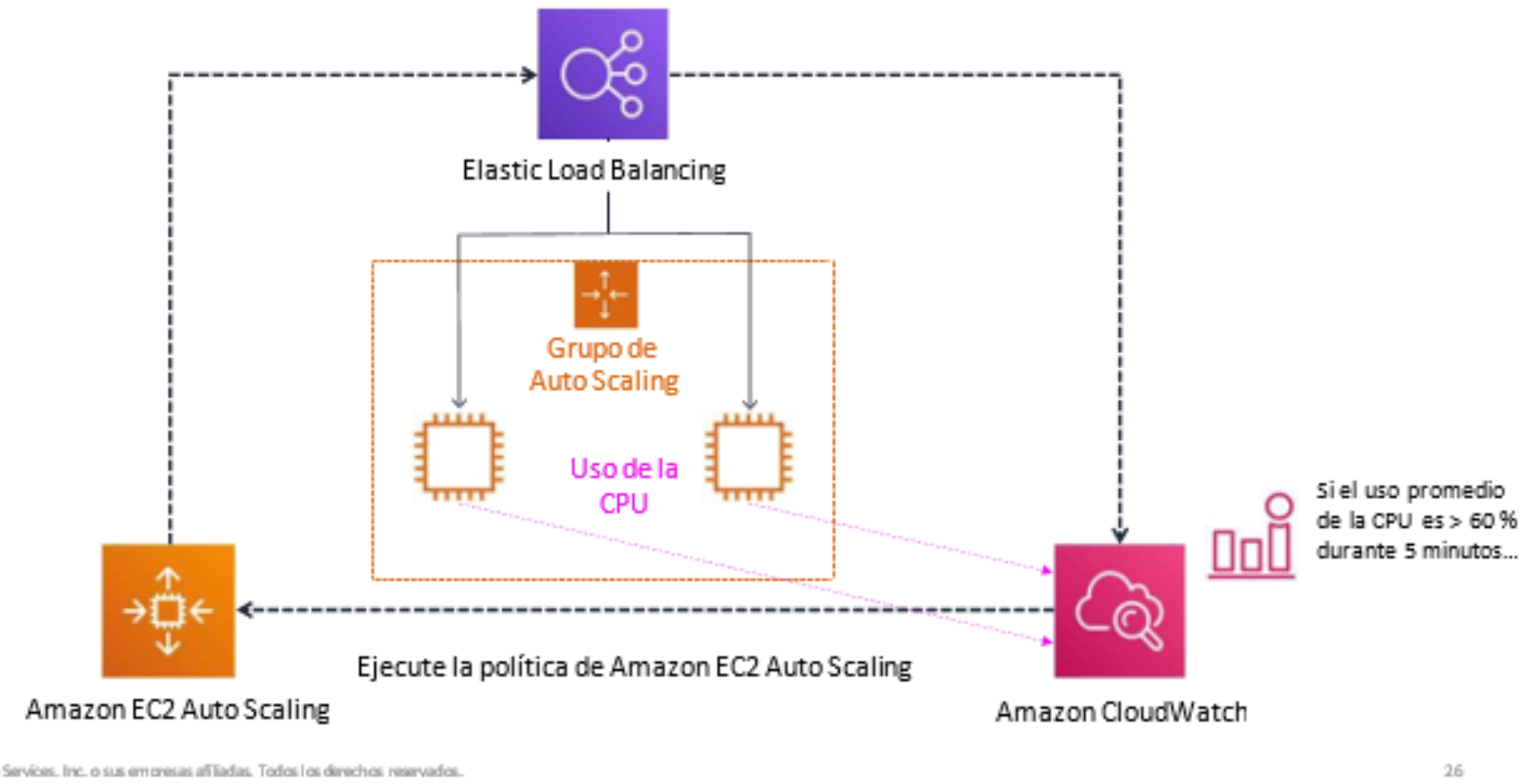
Una forma más avanzada de escalar sus recursos le permite definir los parámetros que controlan el proceso de escalado. Por ejemplo, tiene una aplicación web que actualmente se ejecuta en dos instancias y desea que el uso de la CPU del grupo de Auto Scaling permanezca cerca del 50 % cuando la carga de la aplicación cambia. Esta opción resulta útil cuando el escalado responde a los cambios en las condiciones, pero no se sabe en qué momento cambiarán estas condiciones. El escalado dinámico le brinda capacidad adicional para manejar los picos de tráfico sin tener que mantener una cantidad excesiva de recursos sin utilizar. Puede configurar el grupo de Auto Scaling a fin de que el escalado se realice automáticamente para cubrir esta necesidad. El tipo de política de escalado determina cómo se implementa la acción de escalado. Puede utilizar Amazon EC2 Auto Scaling con Amazon CloudWatch para activar la política de escalado en respuesta a una alarma.

Escalado predictivo:

Puede usar Amazon EC2 Auto Scaling con AWS Auto Scaling para implementar el escalado predictivo, según el cual su capacidad escala en función de la demanda prevista. El escalado predictivo utiliza datos que se recopilan a partir del uso real de EC2, y los datos además se informan mediante miles de millones de puntos de datos extraídos de nuestras propias observaciones. Luego, AWS utiliza modelos de aprendizaje automático bien entrenados para predecir el tráfico esperado (y el uso de EC2), incluidos los patrones diarios y semanales. El modelo necesita datos históricos de al menos 1 día para comenzar a hacer predicciones. El modelo se revalúa cada 24 horas para crear una predicción de las siguientes 48 horas. El proceso de predicción genera un plan de escalado que puede impulsar a uno o más grupos de instancias EC2 escaladas automáticamente.

Para obtener más información acerca de estas opciones, consulte Escalar el tamaño de su grupo de Auto Scaling en la documentación de AWS.

IMPLEMENTACIÓN DEL ESCALADO DINÁMICO



Una configuración común para implementar el escalado dinámico consiste en crear una alarma de CloudWatch basada en la información de rendimiento de sus instancias EC2 o del balanceador de carga. Cuando se infringe un límite de rendimiento, una alarma de CloudWatch desencadena un evento de escalado automático que genera un escalado ascendente o descendente en las instancias EC2 del grupo de Auto Scaling.

Para comprender cómo funciona, considere este ejemplo:

Se crea una alarma de Amazon CloudWatch para monitorear el uso de la CPU en su flota de instancias EC2 y ejecutar políticas de escalado automático si el uso promedio de la CPU en la flota supera el 60 % durante 5 minutos.

Se crea una alarma de Amazon CloudWatch para monitorear el uso de la CPU en su flota de instancias EC2 y ejecutar políticas de escalado automático si el uso promedio de la CPU en la flota supera el 60 % durante 5 minutos.

Una vez agregada la nueva instancia, Amazon EC2 Auto Scaling realiza una llamada a Elastic Load Balancing para registrar la nueva instancia EC2 de ese grupo de Auto Scaling.

Luego, Elastic Load Balancing lleva a cabo las comprobaciones de estado necesarias y comienza a distribuir tráfico a dicha instancia. Elastic Load Balancing dirige el tráfico entre las instancias EC2 y envía métricas a Amazon CloudWatch.

Amazon CloudWatch, Amazon EC2 Auto Scaling y Elastic Load Balancing funcionan bien individualmente. Sin embargo, juntos se vuelven más potentes y aumentan el control sobre cómo su aplicación gestiona la demanda de los clientes y la flexibilidad con la que la aplicación efectúa esta gestión.

AWS AUTO SCALING

• Monitorea sus aplicaciones y ajusta automáticamente la capacidad para mantener un rendimiento estable y predecible al menor costo posible.

• Proporciona una interfaz de usuario simple y potente que le permite crear planes de escalado para los siguientes recursos:

- **InstanciasdeAmazon EC2 y flotas de spot**
- **Tareas de Amazon ElasticContainerService (Amazon ECS)**
- **Tablas e índices deAmazon DynamoDB**
- **Réplicas de Amazon Aurora**

Hasta ahora, ha aprendido a escalar instancias EC2 con Amazon EC2 Auto Scaling. También ha aprendido que puede usar Amazon EC2 Auto Scaling con AWS Auto Scaling para efectuar el escalado predictivo.

AWS Auto Scaling es un servicio independiente que monitorea sus aplicaciones. Ajusta automáticamente la capacidad para mantener un rendimiento estable y predecible al menor costo posible. El servicio proporciona una interfaz de usuario sencilla y potente que le permite crear planes de escalado para los recursos, entre los que se incluyen los siguientes:

• InstanciasdeAmazon EC2 y flotas de spot

• Tareas de Amazon ElasticContainerService (Amazon ECS)

• Tablas e índices deAmazon DynamoDB

• Réplicas de Amazon Aurora

Si ya utiliza Amazon EC2 Auto Scaling para escalar las instancias EC2 de manera dinámica, ahora puede combinarlo con AWS Auto Scaling a fin de escalar recursos adicionales para otros servicios de AWS.

Para obtener más información acerca de AWS Auto Scaling, consulte AWS Auto Scaling.

ESTOS SON ALGUNOS DE LOS APRENDIZAJES CLAVE DE ESTA LECCIÓN:

- **El escalado le permite responder rápidamente a los cambios en las necesidades de recursos.**
- **Amazon EC2 Auto Scaling lo ayuda a mantener la disponibilidad de la aplicación y le permite agregar o eliminar instancias EC2 de forma automática según las cargas de trabajo.**
- **Un grupo de Auto Scaling es una colección de instancias EC2.**
- **Una configuración de lanzamiento es una plantilla de configuración de instancias.**
- **Puede implementar el escalado dinámico con Amazon EC2 Auto Scaling, Amazon CloudWatch y Elastic Load Balancing.**

- **AWS Auto Scaling es un servicio independiente que monitorea sus aplicaciones y ajusta automáticamente la capacidad para los siguientes recursos:**

Instancias de Amazon EC2 y flotas de spot
Tareas de Amazon ECS
Tablas e índices de Amazon DynamoDB
Réplicas de Amazon Aurora