11: Implementación de elasticidad, y alta disponibilidad

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

aws academy

Información general sobre el módulo

Secciones

1. Necesidad de arquitectura
2. Escalado de los recursos informáticos
3. Escalado de las bases de datos
4. Diseño de un entorno de alta disponibilidad

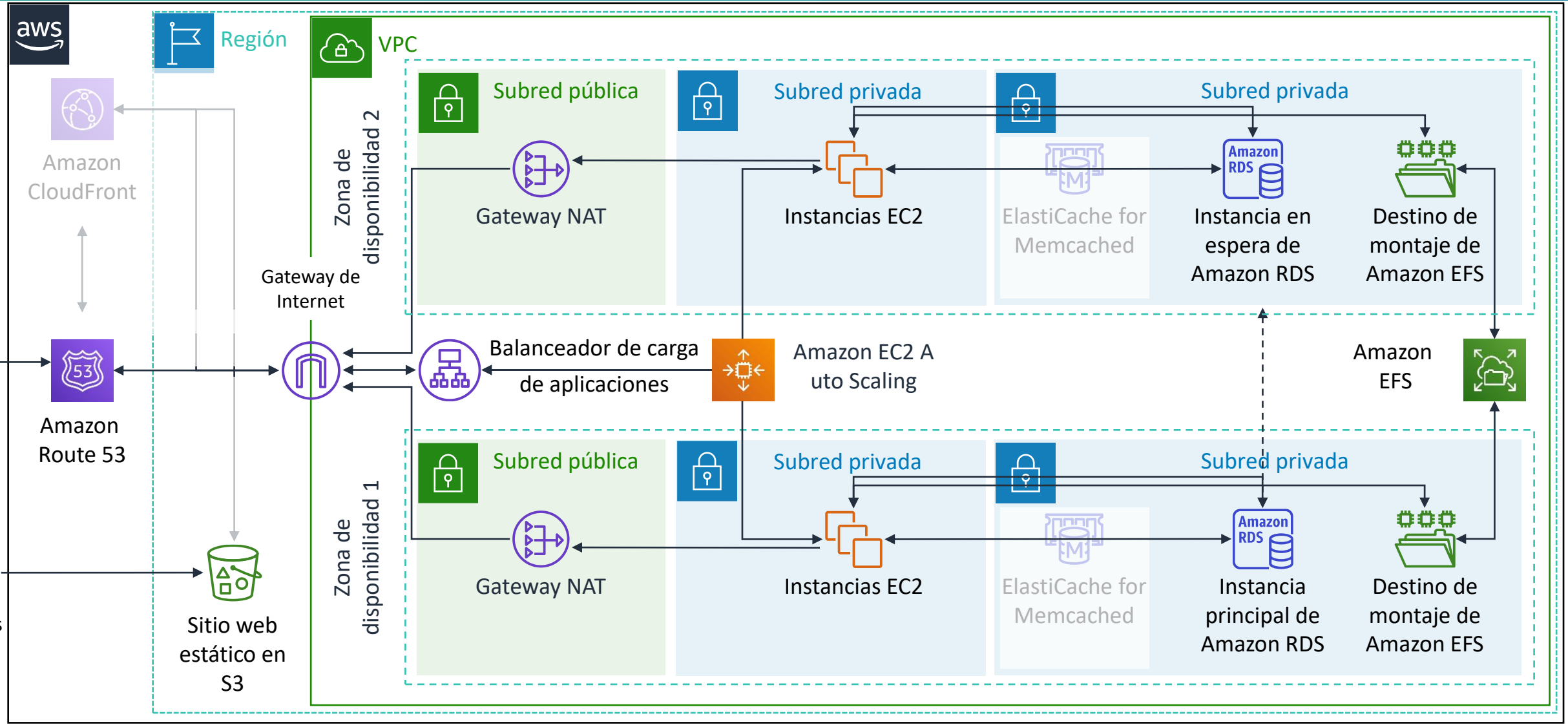
Objetivos del módulo

Una vez finalizado este módulo, debería ser capaz de lo siguiente:

- Utilizar Amazon EC2 Auto Scaling dentro de una arquitectura para fomentar la elasticidad
- Explicar cómo escalar los recursos de las bases de datos
- Implementar un balanceador de carga de aplicaciones para crear un entorno de alta disponibilidad
- Utilizar Amazon Route 53 para la conmutación por error de sistemas de nombres de dominio (DNS).
- Crear un entorno de alta disponibilidad

Necesidad de arquitectura

Implementar la alta disponibilidad como parte de una arquitectura más grande



Diseño de un entorno de alta disponibilidad

Sistemas de alta disponibilidad

- El sistema puede resistir cierto grado de degradación sin dejar de estar disponible
- Puede minimizar el tiempo de inactividad
- Requiere una intervención humana mínima
- Recuperación de errores o paso a origen secundario al cabo de una cantidad aceptable de tiempo de rendimiento degradado

Porcentaje de tiempo de actividad	Tiempo de inactividad máximo por año	Tiempo de inactividad equivalente por día
90 %	36,5 días	2,4 horas
99 %	3,65 días	14 minutos
99,9 %	8,76 horas	86 segundos
99,99 %	52,6 minutos	8,6 segundos
99,999 %	5,25 minutos	0,86 segundos

Elastic Load Balancing

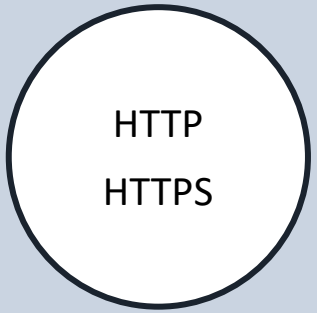
Elastic Load Bal
ancing

Se trata de un servicio de balanceo de carga administrado que distribuye el tráfico entrante de las aplicaciones en varias instancias EC2, direcciones IP y funciones Lambda.

- Puede ser externo o interno
- Cada balanceador de carga recibe un nombre DNS
- Reconoce instancias en mal estado y responde a ellas.

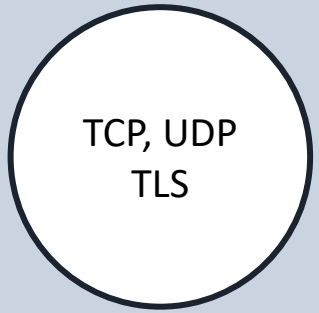
Tipos de balanceadores de carga

Balanceador de carga de aplicaciones



- Administración flexible de aplicaciones
- Balanceo de carga avanzado del tráfico de HTTP y HTTPS
- Funciona a nivel de solicitud (capa 7)

Balanceador de carga de red



- Dirección IP estática y de rendimiento sumamente alto para su aplicación
- Balanceo de carga del tráfico TCP, UDP y TLS
- Funciona a nivel de conexión (capa 4)

Balanceador de carga **clásico**

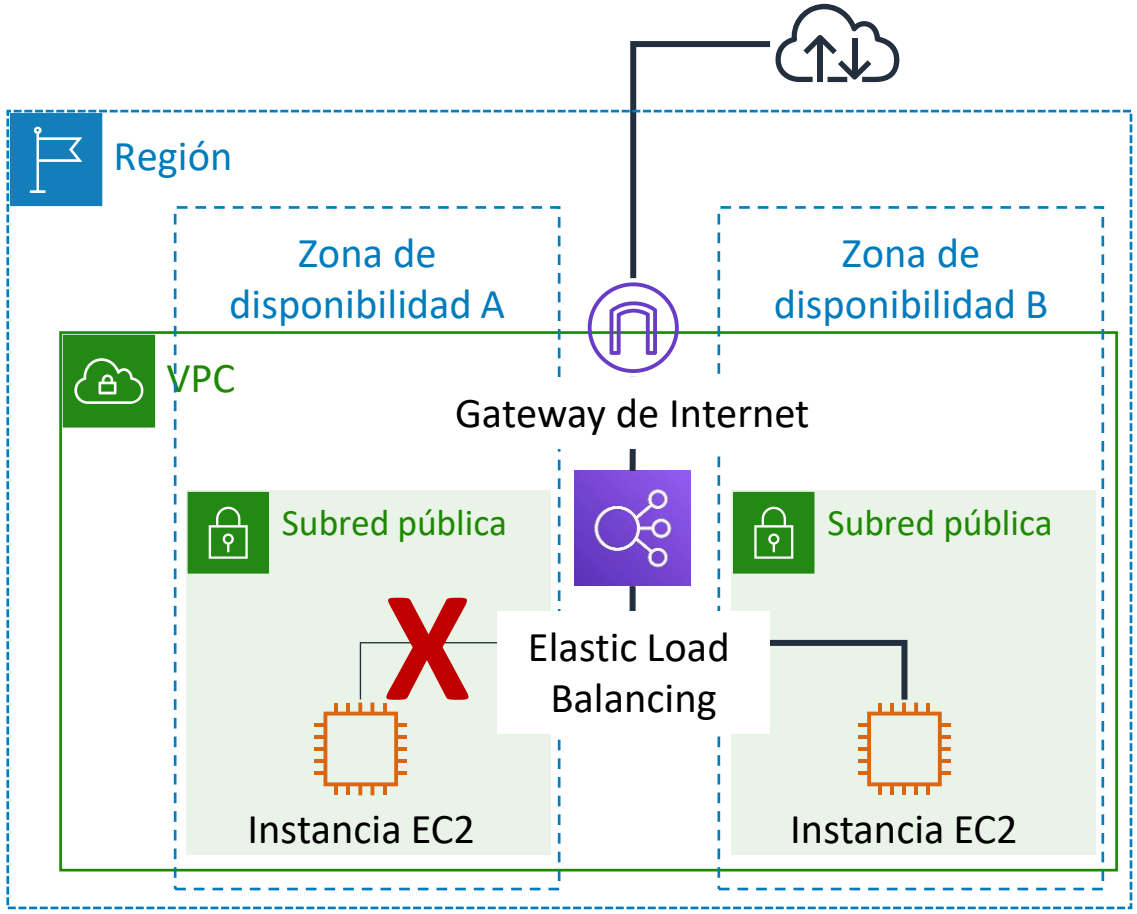
GENERACIÓN ANTERIOR para HTTP, HTTPS, TCP y SSL

- Balanceo de carga en varias instancias EC2
- Funciona tanto a nivel de solicitud como a nivel de conexión

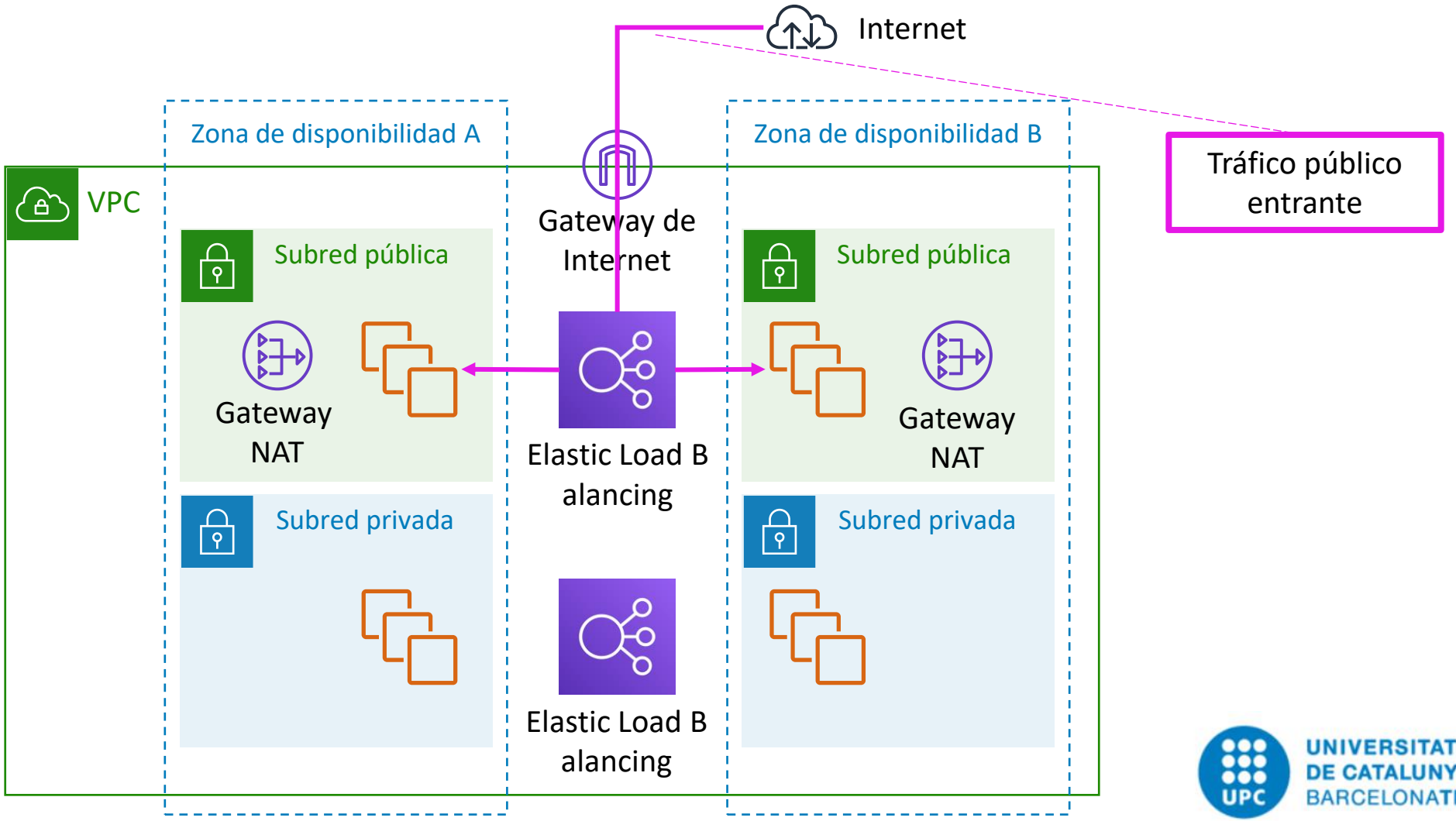
Implementación de la disponibilidad alta

Comience con dos zonas de disponibilidad por región de AWS.

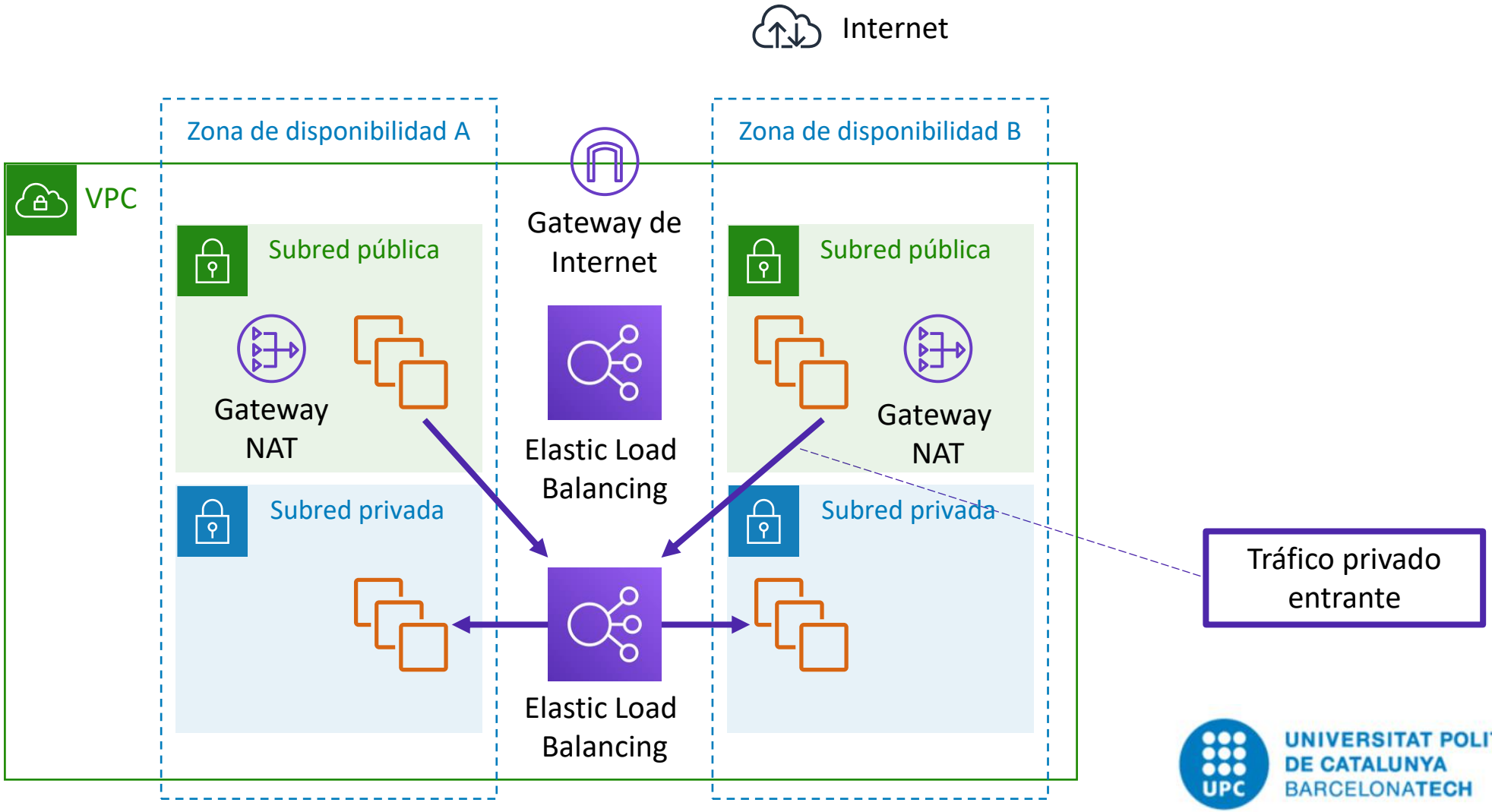
Si los recursos de una zona de disponibilidad son inaccesibles, la aplicación no debería fallar.



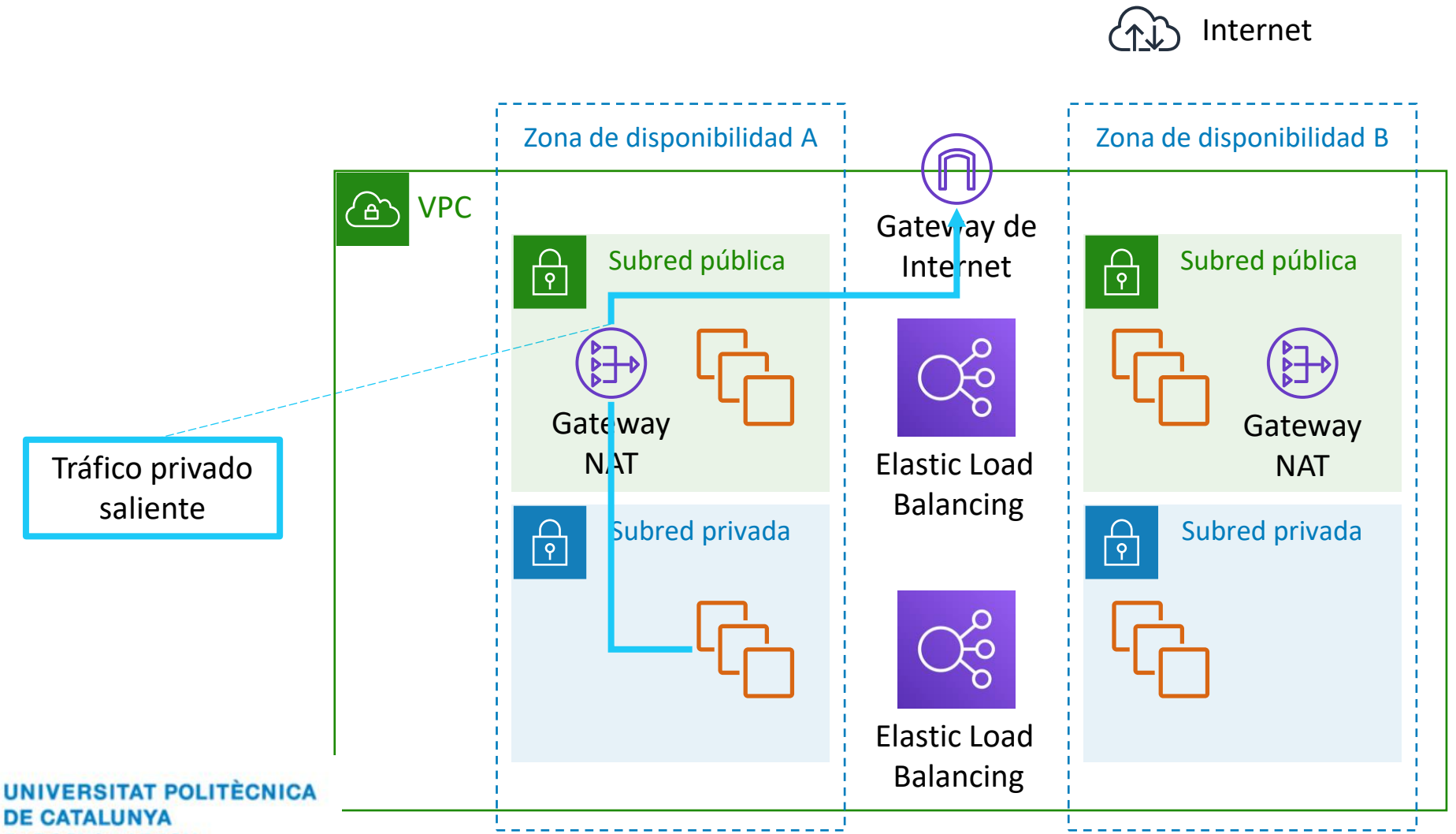
Ejemplo de una arquitectura de disponibilidad alta (1 de 3)



Ejemplo de una arquitectura de disponibilidad alta (2 de 3)



Ejemplo de una arquitectura de disponibilidad alta (3 de 3)



Amazon Route 53

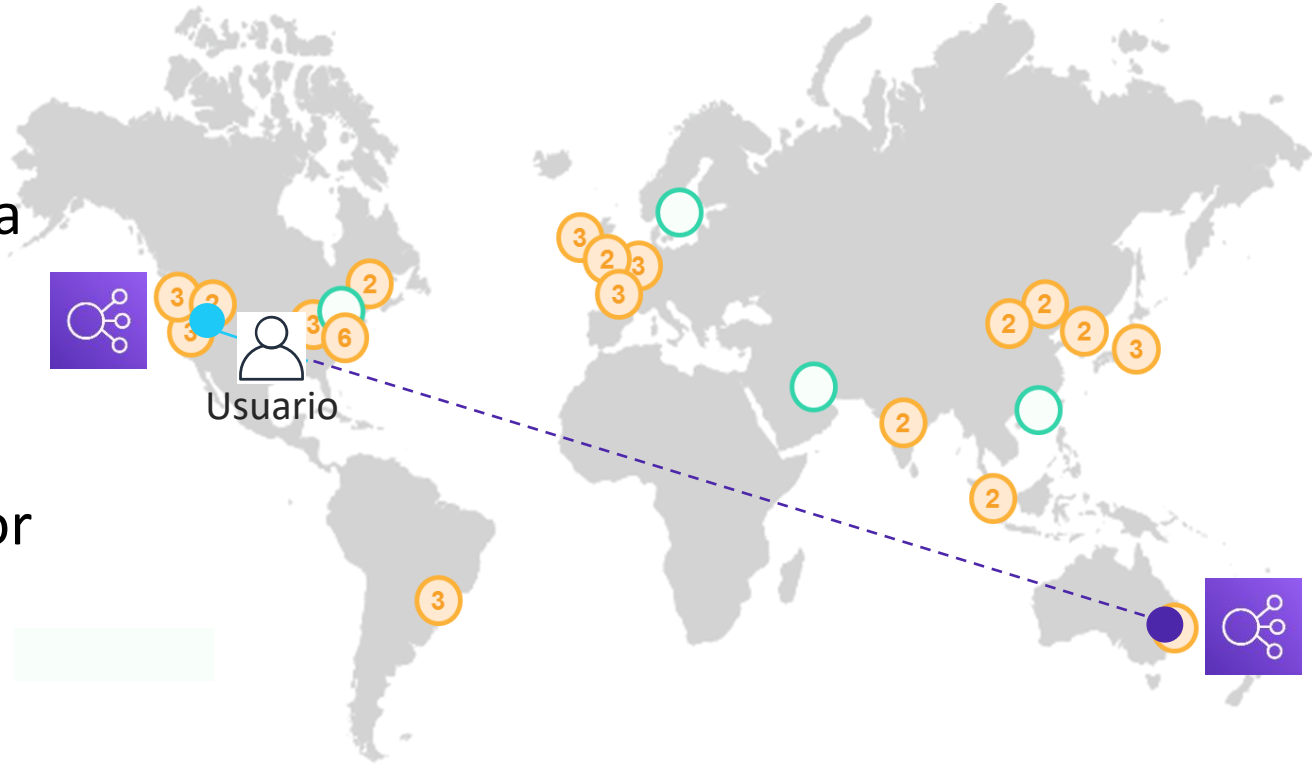
Amazon
Route 53

Amazon Route 53 es un servicio de DNS escalable y de disponibilidad alta en la nube.

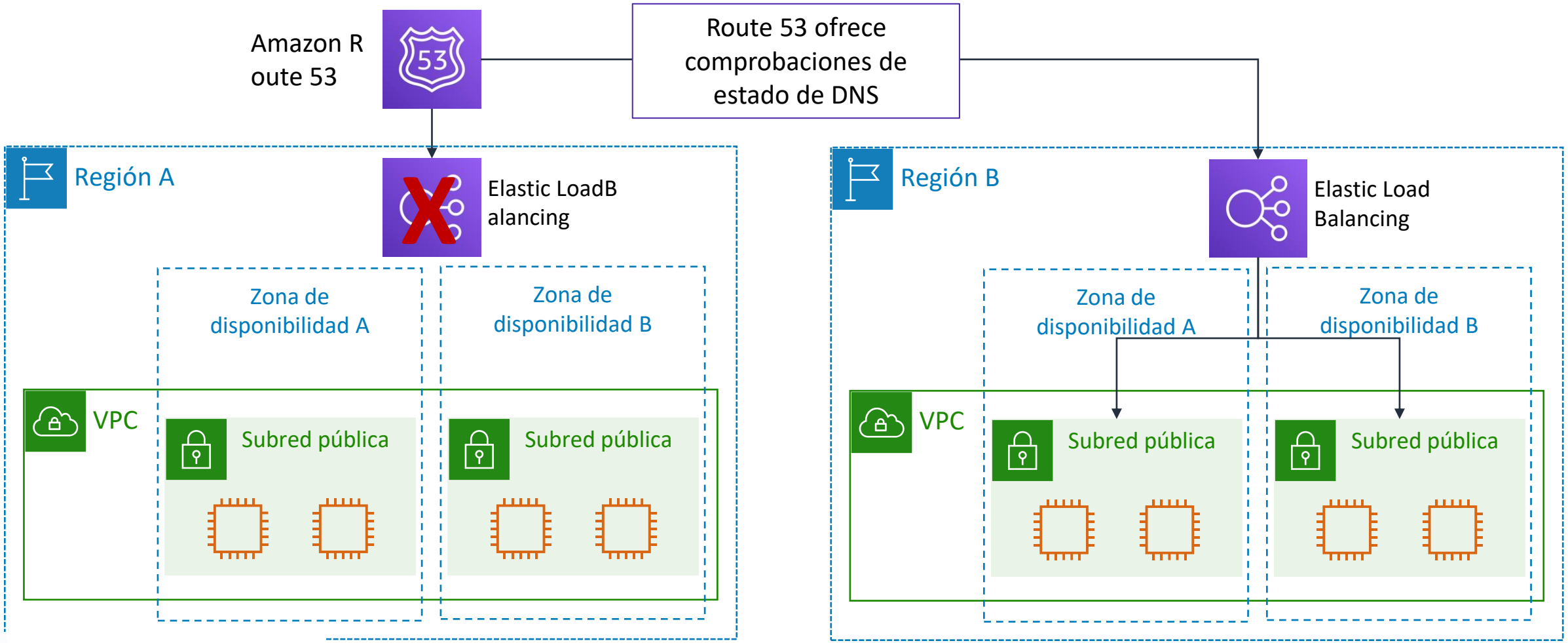
- Traduce los nombres de los dominios a direcciones IP
- Conecta las solicitudes de los usuarios a infraestructura que se ejecuta en AWS y fuera de ella
- Se puede configurar para que direccione el tráfico a puntos de enlace en buen estado o para que monitoree el estado de la aplicación y sus puntos de enlace
- Ofrece el registro de nombres de dominio
- Tiene varias opciones de direccionamiento

Direccionamiento admitido por Amazon Route 53

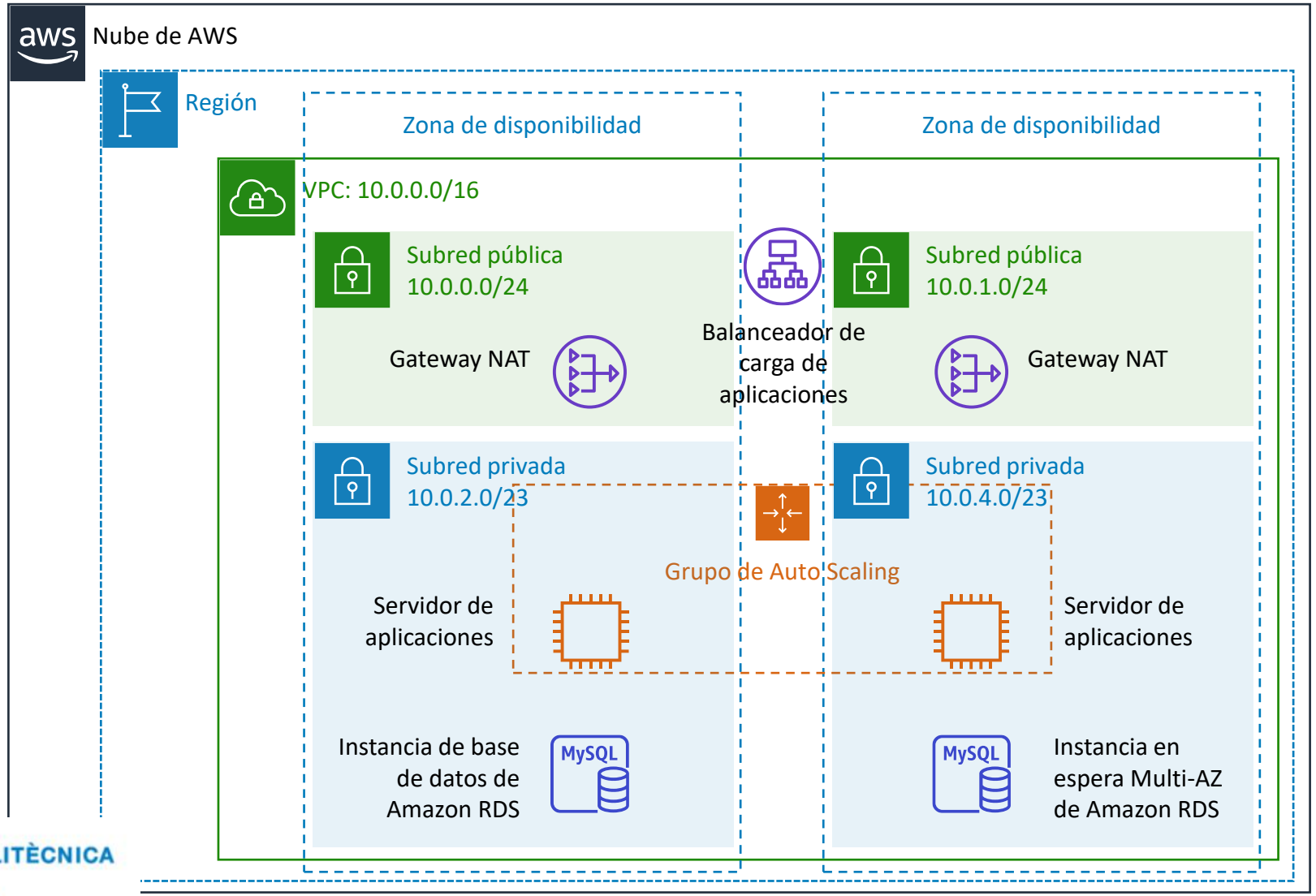
- Direccionamiento simple
- Direccionamiento de turno rotativo ponderado
- Direccionamiento basado en la latencia
- Direccionamiento de geolocalización
- Direccionamiento de geoproximidad
- Direccionamiento tras conmutación por error
- Direccionamiento de respuesta con varios valores



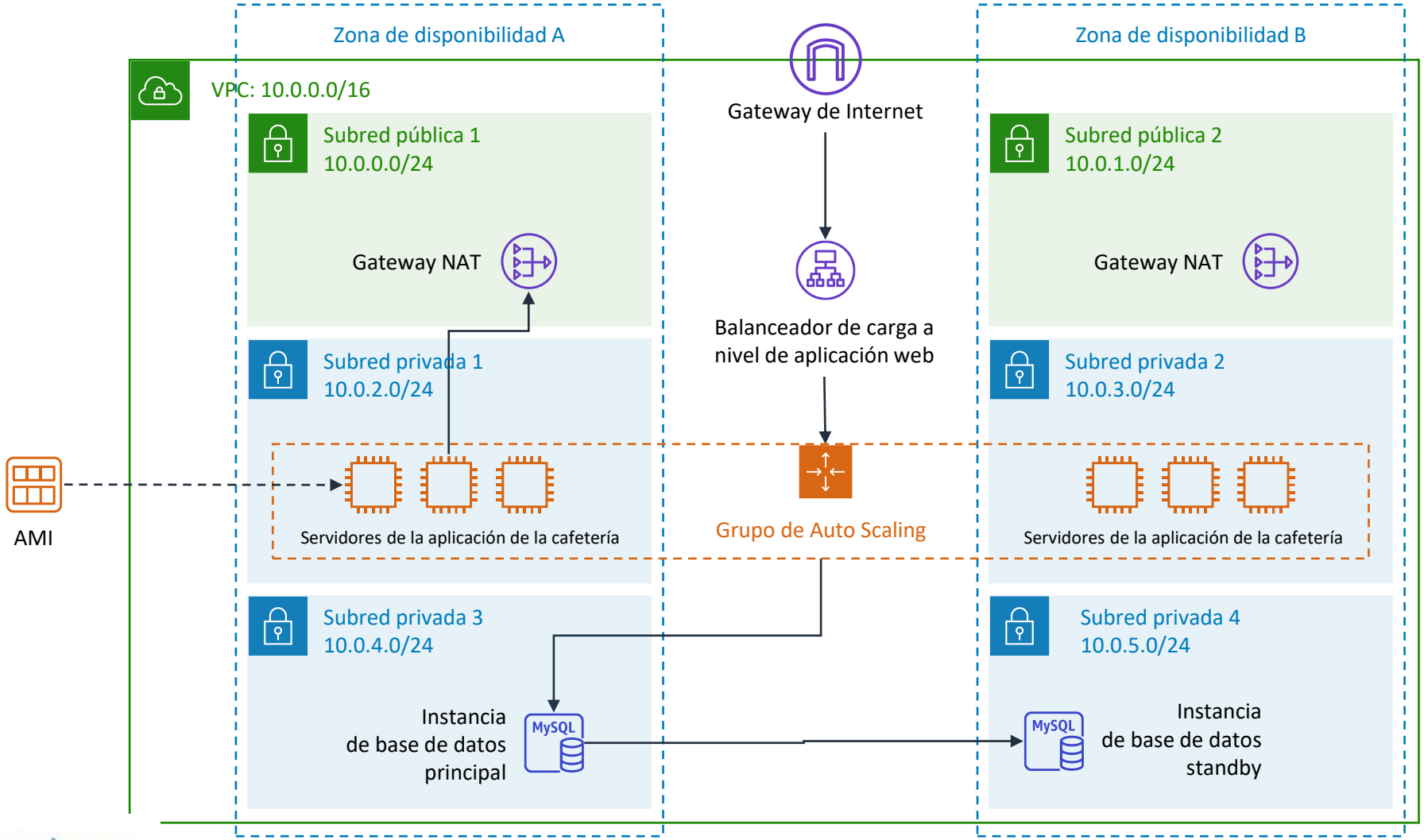
Disponibilidad alta y DNS en varias regiones



Architectura de 3 capas



Architectura 3 capas



Escalado de los recursos informáticos

¿Qué es la elasticidad?

Una infraestructura elástica puede ampliarse y contraerse a medida que cambian las necesidades de capacidad.

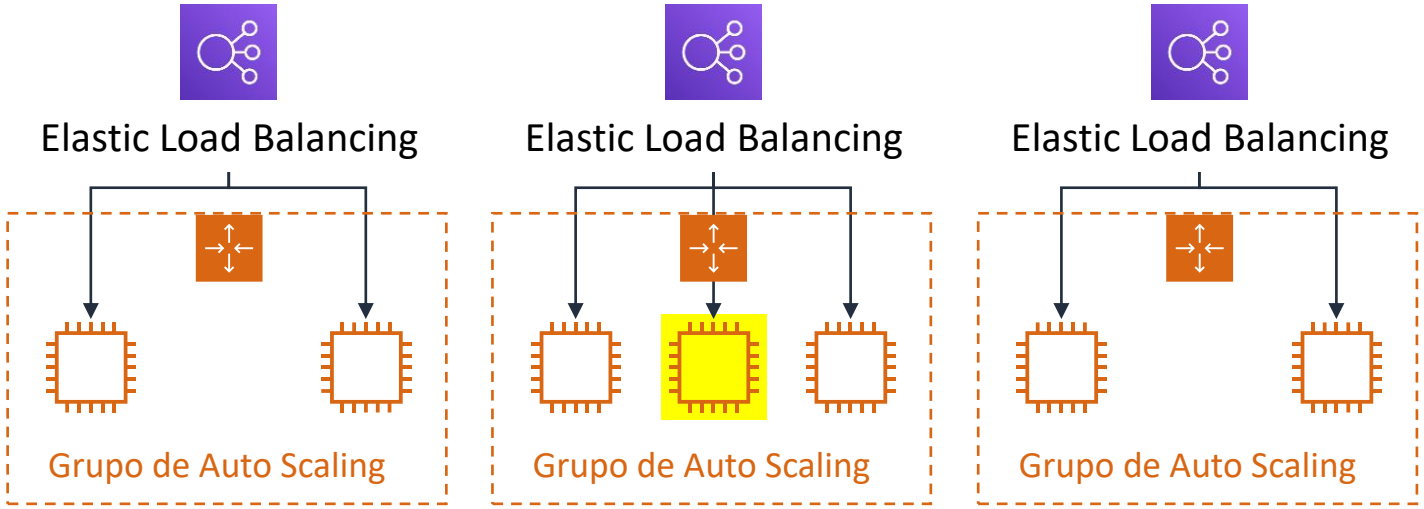
Ejemplos:

- Aumento de la cantidad de servidores web si se incrementa el tráfico
- Reducción de la capacidad de escritura en la base de datos si el tráfico disminuye
- Manejo de la fluctuación diaria de la demanda en toda la arquitectura

¿Qué es el escalado?

Una técnica que se utiliza para lograr elasticidad

Escalado horizontal

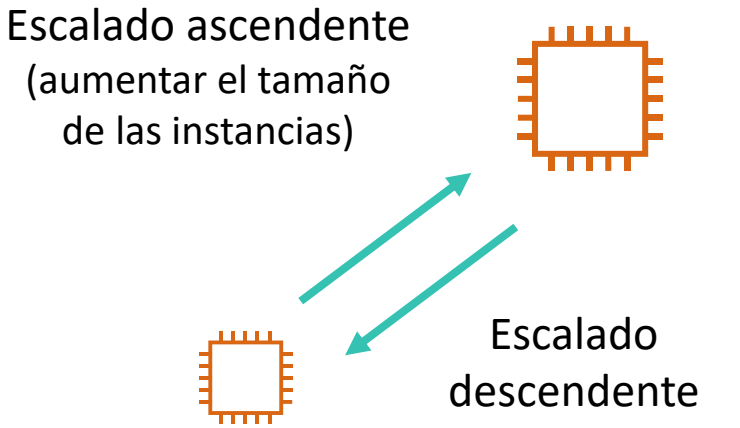


Configuración básica

Escalado ascendente (lanzar instancias)

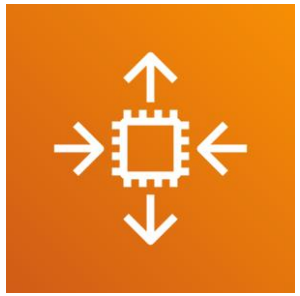
Escalado descendente (terminar instancias)

Escalado vertical



UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Amazon EC2 Auto Scaling



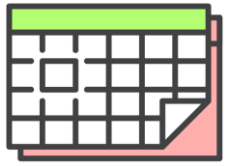
Amazon EC2 A
uto Scaling

- Lanza o termina instancias en función de las condiciones especificadas
- De forma automática, registra nuevas instancias con balanceadores de carga si así se especifica
- Puede llevar a cabo lanzamientos en las zonas de disponibilidad

Opciones de escalado

Programado

Adecuado para cargas de trabajo predecibles



Escalado según la fecha y la hora

Caso de uso: Desactivación de las instancias de desarrollo y prueba durante la noche

Dinámico

Adecuado para cambios en las condiciones

Admite el seguimiento de valores objetivo

Caso de uso: Escalado según el uso de la CPU

Predictivo

Adecuado para la demanda prevista

Escalado según el aprendizaje automático (ML)

Caso de uso: Manejo de un aumento en la carga de trabajo para el sitio web de comercio electrónico durante un evento importante de ventas

Tipos de políticas de escalado dinámico

- Escalado sencillo: ajuste de escalado sencillo
  - Ejemplos de casos de uso: cargas de trabajo nuevas, picos de cargas de trabajo
- Escalado por pasos: el ajuste depende del tamaño de la interrupción de alarma
  - Ejemplo de caso de uso: cargas de trabajo previsibles
- Escalado de seguimiento de valores objetivo: valor objetivo para una métrica determinada
  - Ejemplo de caso de uso: aplicaciones escalables horizontalmente, como aplicaciones de carga balanceada y aplicaciones de procesamiento de datos por lotes

Grupos de Auto Scaling

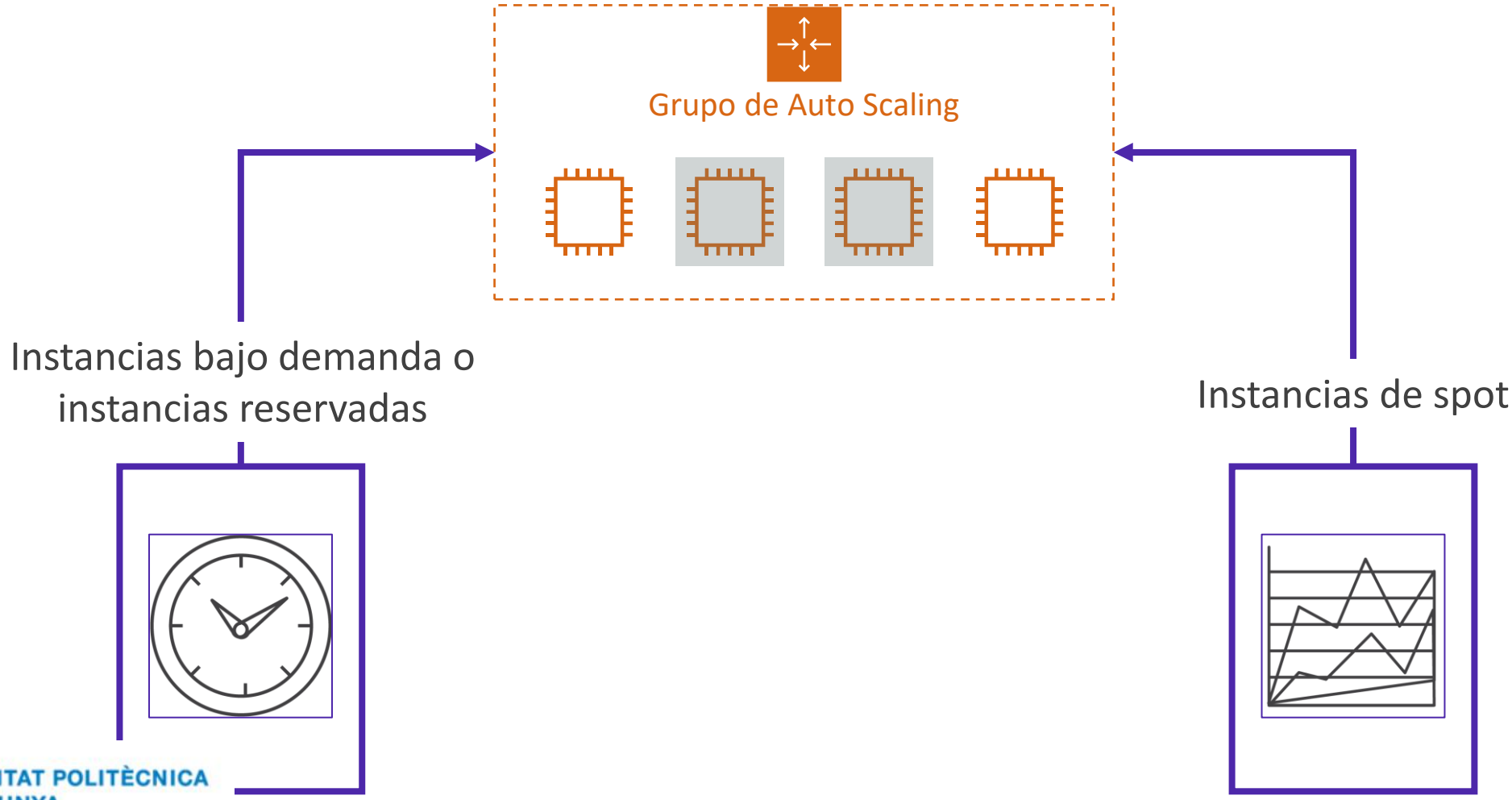
Un grupo de Auto Scaling define lo siguiente:

- La capacidad mínima
- La capacidad máxima
- La capacidad deseada*

¿Capacidad?

* La capacidad deseada refleja la cantidad de instancias que se están ejecutando y puede fluctuar en respuesta a eventos.

Amazon EC2 Auto Scaling: opciones de compra



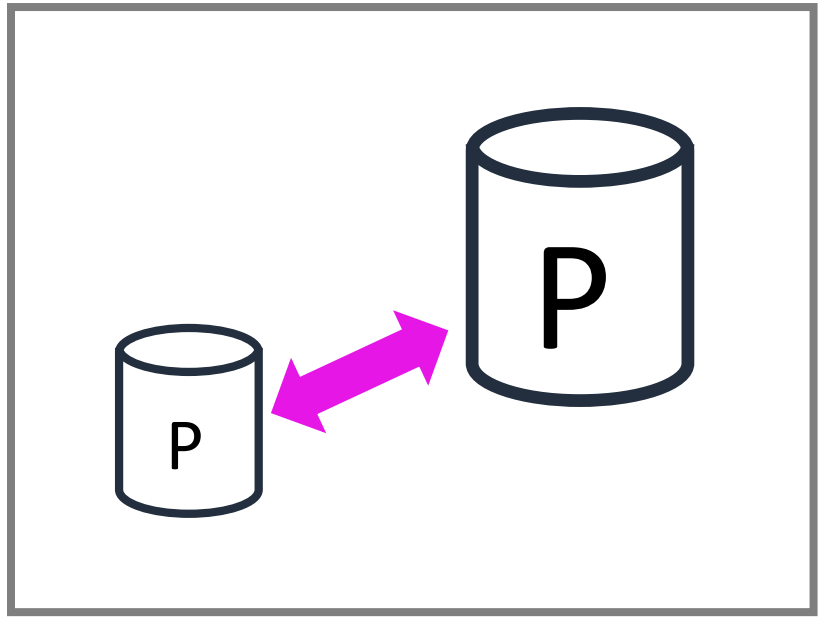
Consideraciones relativas al escalado automático

- Varios tipos de escalado automático
- Escalado sencillo, por pasos o de seguimiento de valores objetivo
- Varias métricas (no solo relativas a la CPU)
- Cuándo utilizar el escalado ascendente y cuándo utilizar el escalado descendente
- Uso de enlaces de ciclo de vida

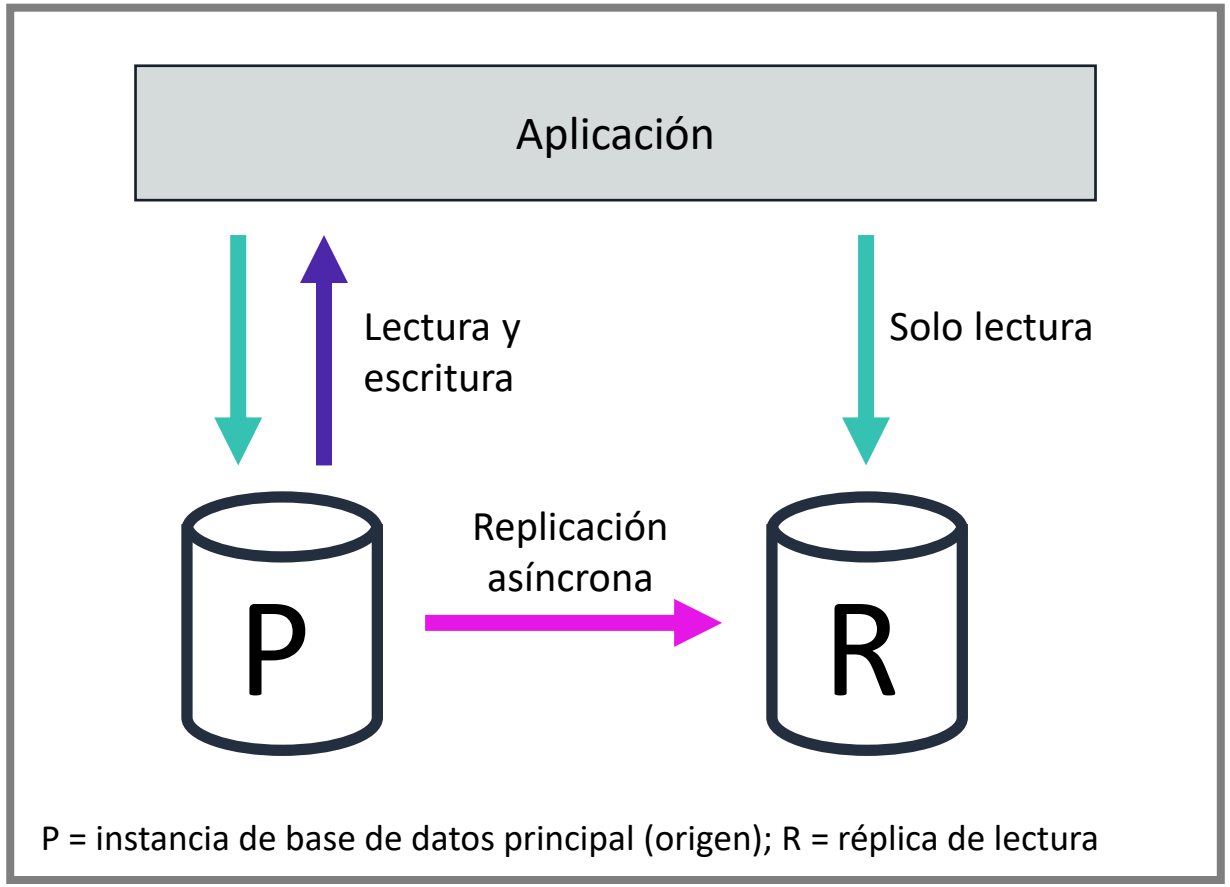
Escalado de las bases de datos

Escalado vertical con Amazon RDS: escalado con el botón de comando

- Escalar instancias de base de datos verticalmente (de forma ascendente o descendente)
- Desde micro hasta 24xlarge y todo lo que se encuentre en medio
- Escale verticalmente con un tiempo de inactividad mínimo



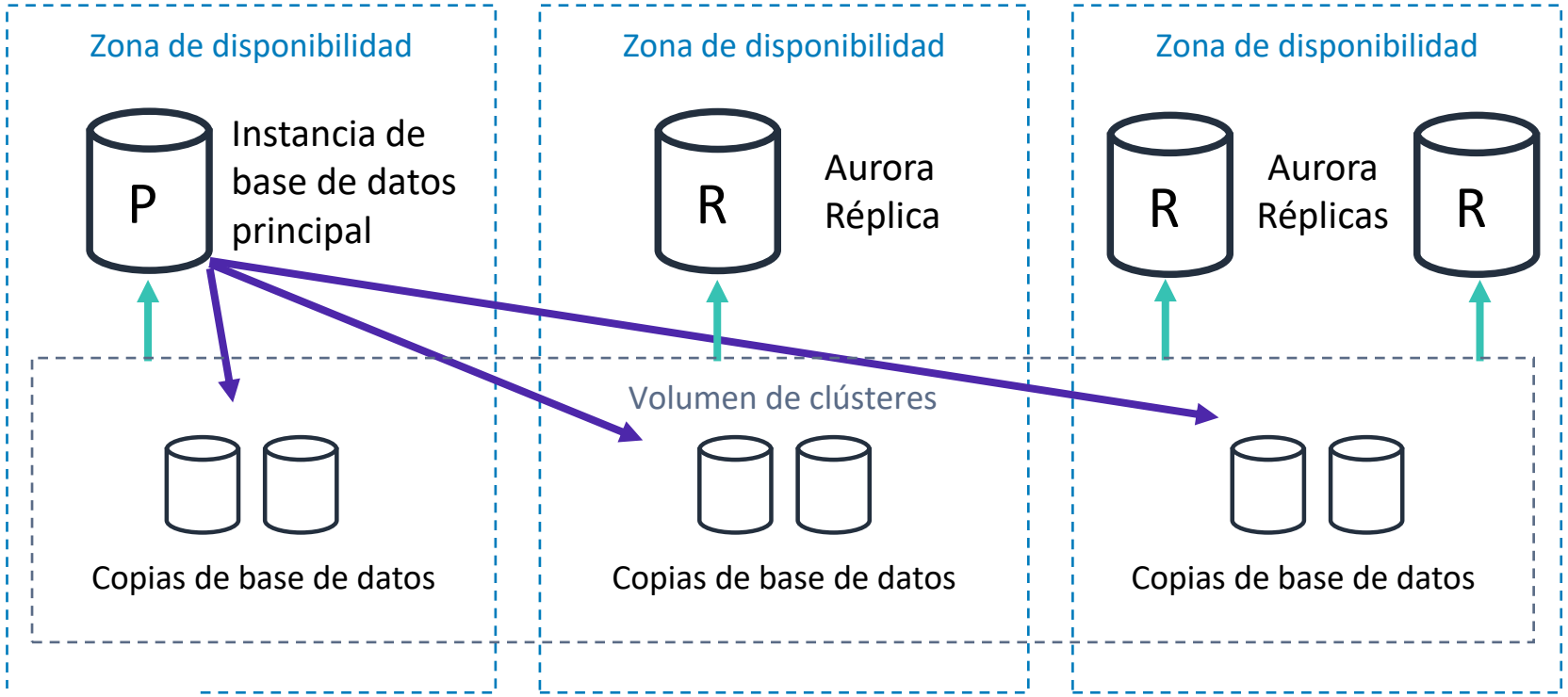
Escalado horizontal con Amazon RDS: réplicas de lectura



- Escalado horizontal para cargas de trabajo de lectura intensiva
- Hasta 5 réplicas de lectura y hasta 15 réplicas de Aurora
- La replicación es asíncrona
- Disponible para Amazon RDS for MySQL, MariaDB, PostgreSQL y Oracle

Escalado con Amazon Aurora

Cada clúster de la base de datos de Aurora puede tener hasta 15 réplicas de Aurora

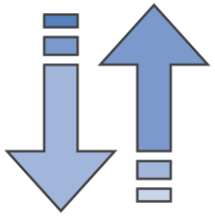


Amazon Aurora Serverless

Responde a su aplicación automáticamente:

- Escala la capacidad
- Comienza
- Termina

Pague por la cantidad de unidades de capacidad de Aurora (ACU) que se utilicen



Adecuado para cargas de trabajo intermitentes e impredecibles

Escalado horizontal: partición de base de datos

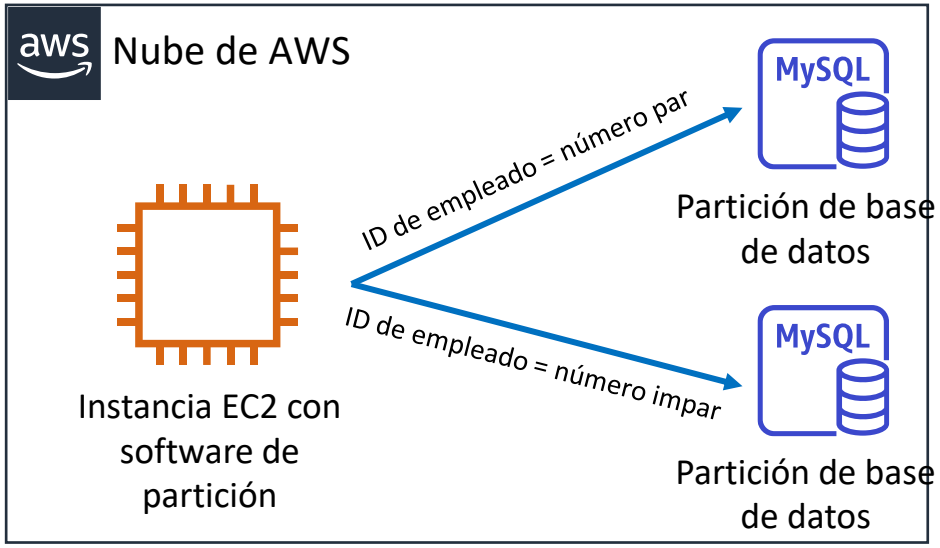
Si no se llevó a cabo una partición, todos los datos están en una única partición.

- Ejemplo: las ID de los empleados en una única base de datos

Si se llevó a cabo una partición, los datos se dividen en segmentos grandes (particiones).

- Ejemplo: las ID de los empleados con números pares en una base de datos y las ID de los empleados con números impares en otra base de datos

En muchas circunstancias, la partición mejora el rendimiento de la escritura.



Escalado con Amazon DynamoDB: bajo demanda

Bajo demanda

Pago por solicitud

Sin aprovisionamiento

Caso de uso: cargas de trabajo con picos de demanda impredecibles. Se adapta rápidamente a las necesidades.

Escalado con Amazon DynamoDB: escalado automático

Bajo demanda

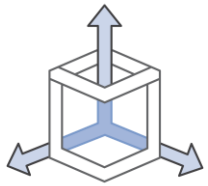
Pago por solicitud

Sin aprovisionamiento

Caso de uso: cargas de trabajo con picos de demanda impredecibles. Se adapta rápidamente a las necesidades.

Escalado automático

Predeterminado para todas las tablas nuevas



Se especifican los límites superiores e inferiores

Caso de uso: escalado general, solución adecuada para la mayoría de las aplicaciones.

Resumen del módulo

A modo de resumen, en este módulo, aprendió a hacer lo siguiente:

- Utilizar Amazon EC2 Auto Scaling dentro de una arquitectura para fomentar la elasticidad
- Explicar cómo escalar los recursos de las bases de datos
- Implementar un balanceador de carga de aplicaciones para crear un entorno de alta disponibilidad
- Utilizar Amazon Route 53 para la conmutación por error a nivel de DNS.
- Crear un entorno de alta disponibilidad

Recursos adicionales

- Set it and Forget it Auto Scaling Target Tracking Policies
- Introduction to Amazon Elastic Load Balancer – Application
- Configuring Auto Scaling Group with ELB Elastic Load Balancer
- ¿Qué es un balanceador de carga de aplicaciones?

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Gracias

aws academy