Arquitecturas de referencia de AWS

Amazon EC2

Amazon EBS

Elastic Load Balancing

Amazon S3

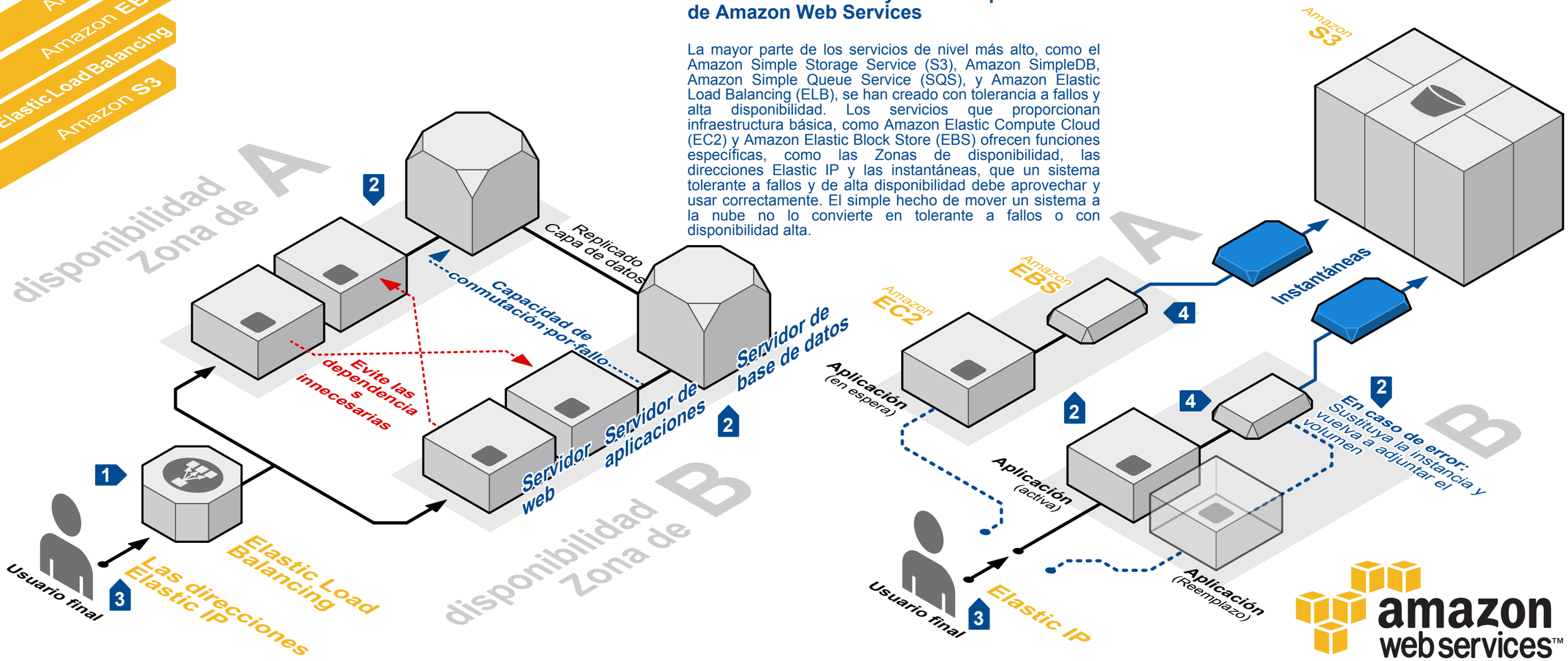
TOLERANCIA A FALLOS & ALTA DISPONIBILIDAD

Amazon Web Services proporciona servicios e infraestructura para generar sistemas fiables, tolerantes a fallos y alta disponibilidad en la nube. Estas cualidades se han incorporado a nuestros servicios tanto para gestionar los aspectos que no requieren ninguna acción especial por su parte como para proporcionar las características que se deben utilizar explícitamente y correctamente.

Amazon EC2 ofrece bloques de generación de infraestructura que, por ellos mismos, pueden no ser tolerantes a fallos. Los discos duros pueden fallar, el suministro eléctrico puede fallar y las estanterías pueden fallar. Es importante utilizar combinaciones de las funciones presentadas en este documento para lograr tolerancia a fallos y alta disponibilidad.

La tolerancia a fallos y la alta disponibilidad de Amazon Web Services

La mayor parte de los servicios de nivel más alto, como el Amazon Simple Storage Service (S3), Amazon SimpleDB, Amazon Simple Queue Service (SQS), y Amazon Elastic Load Balancing (ELB), se han creado con tolerancia a fallos y alta disponibilidad. Los servicios que proporcionan infraestructura básica, como Amazon Elastic Compute Cloud (EC2) y Amazon Elastic Block Store (EBS) ofrecen funciones específicas, como las Zonas de disponibilidad, las direcciones Elastic IP y las instantáneas, que un sistema tolerante a fallos y de alta disponibilidad debe aprovechar y usar correctamente. El simple hecho de mover un sistema a la nube no lo convierte en tolerante a fallos o con disponibilidad alta.



Descripción general del sistema

1 El equilibrio de carga es una manera efectiva de incrementar la disponibilidad de un sistema. Las instancias que fallan se pueden reemplazar perfectamente detrás del equilibrador de carga mientras otras instancias siguen funcionando **Elastic Load Balancing** se puede utilizar para equilibrar instancias en diversas zonas de disponibilidad de una región.

2 **Las Zonas de disponibilidad** son diferentes ubicaciones geográficas creadas para ser aisladas de los fallos que ocurran en otras Zonas de disponibilidad. Al poner instancias de **Amazon EC2** en diversas Zonas de disponibilidad, se pueden proteger las aplicaciones de los fallos en una única ubicación. Es importante ejecutar pilas de aplicación independiente en más de una Zona de disponibilidad, ya sea en la misma región o en otra, de modo que si una zona falla, la aplicación de la otra zona puede seguir ejecutándose. Cuando diseñe un sistema así, deberá conocer bien las dependencias de zona.

3 **Las direcciones Elastic IP** son direcciones IP públicas que se pueden asignar de forma programada entre instancias dentro de una región. Están asociadas con la cuenta AWS y no con una instancia específica ni con la vida de una instancia.
Las direcciones Elastic IP se pueden utilizar para evitar los errores de host o de la zona de disponibilidad mediante la rápida reasignación de la dirección a otra instancia en ejecución o a una instancia de repuesto que se acaba de iniciar. Las instancias reservadas pueden ayudar a garantizar que esa capacidad está disponible en otra zona.

4 Los datos valiosos nunca deben almacenarse solo en almacenamiento de instancias sin copias de seguridad, replicación o capacidad de recreación de datos. **Amazon Elastic Block Store (EBS)** ofrece volúmenes de almacenamiento fuera de instancia con un orden de magnitud más duradero que el almacenamiento en instancia. Los volúmenes de EBS se replican automáticamente en una única zona de disponibilidad. Para aumentar la durabilidad, se pueden crear instantáneas puntuales para almacenar datos en volúmenes en Amazon S3, que se replicarán en diversas Zonas de disponibilidad. Mientras que los volúmenes de EBS están vinculados a una Zona de disponibilidad específica, las instantáneas están vinculadas a la región. Con una instantánea puede crear nuevos volúmenes de EBS en cualquiera de las zonas de disponibilidad de la misma región. Esta es una manera efectiva de gestionar los errores de disco u otros problemas a nivel de host, así como los problemas que afecten la Zona de disponibilidad. Las instantáneas son incrementales, de modo que es recomendable limitarse a las instantáneas recientes.