AWS Reference Architectures

Amazon EC2

Amazon EBS

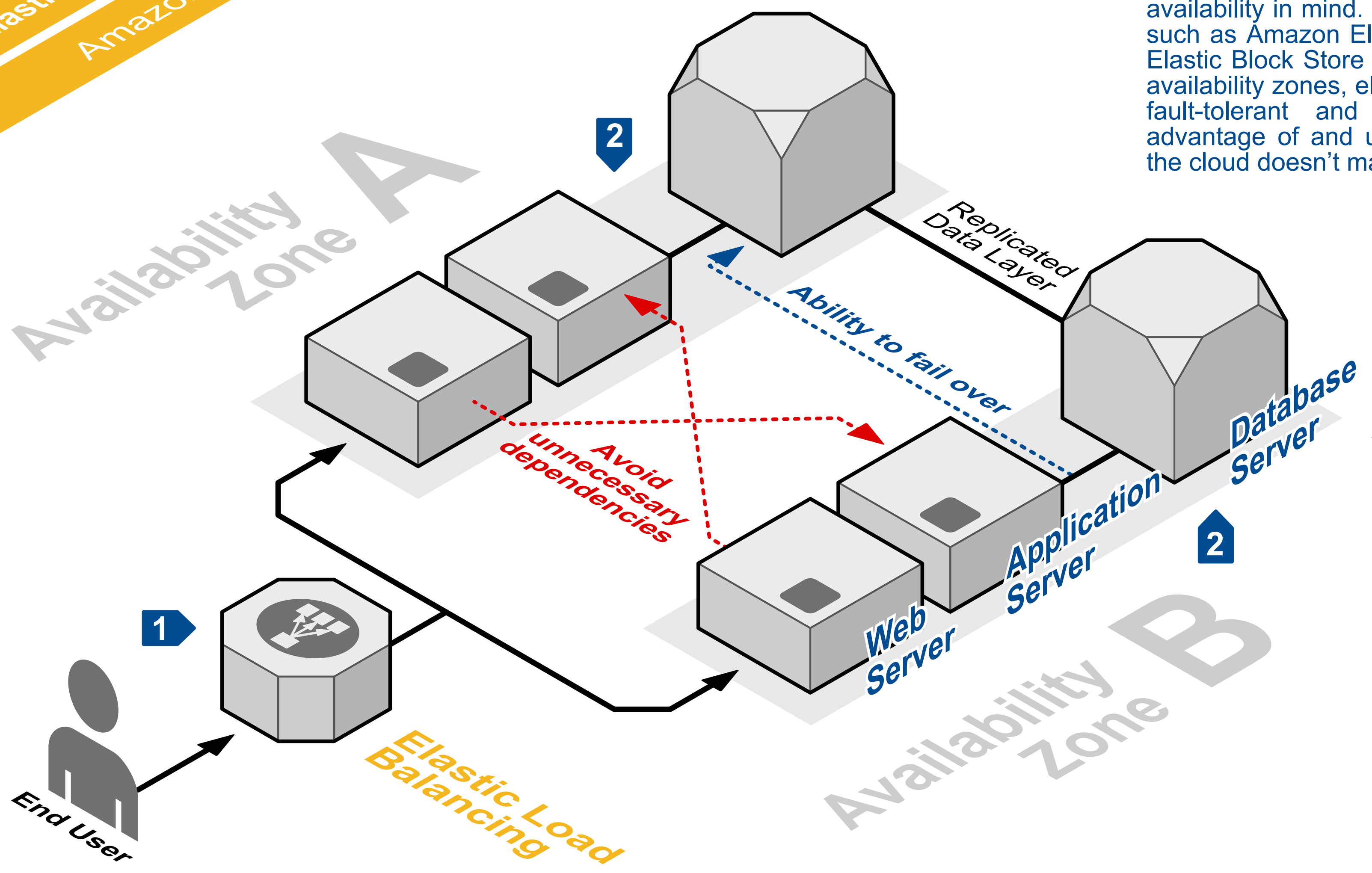
Elastic Load Balancing

Amazon S3

FAULT TOLERANCE & HIGH AVAILABILITY

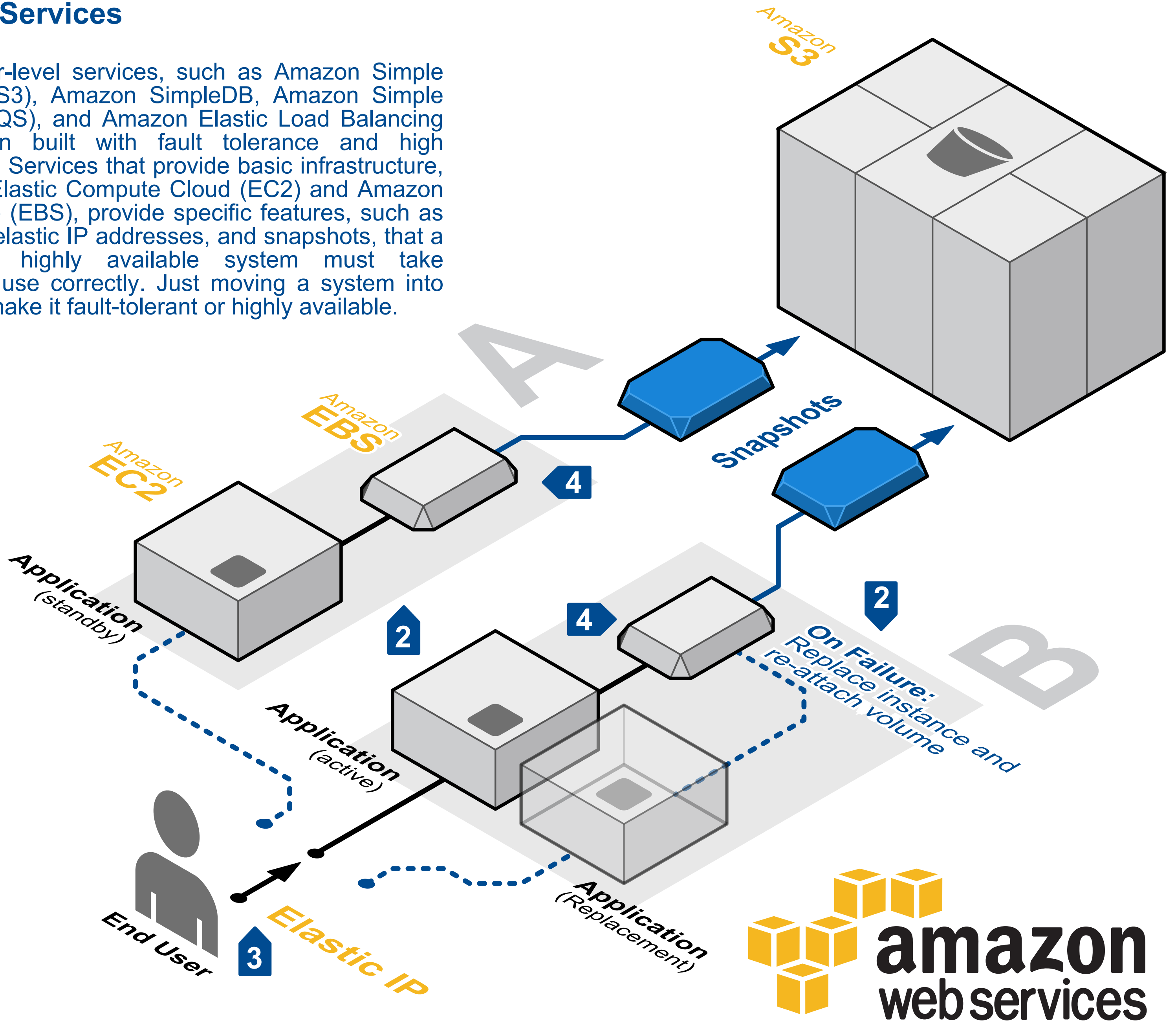
Amazon Web Services provides services and infrastructure to build reliable, fault-tolerant, and highly available systems in the cloud.These qualities have been designed into our services both by handling such aspects without any special action by you and by providing features that must be used explicitly and correctly.

Amazon EC2 provides infrastructure building blocks that, by themselves, may not be fault-tolerant. Hard drives may fail, power supplies may fail, and racks may fail. It is important to use combinations of the features presented in this document to achieve fault tolerance and high availability.



Fault Tolerance and High Availability of Amazon Web Services

Most of the higher-level services, such as Amazon Simple Storage Service (S3), Amazon SimpleDB, Amazon Simple Queue Service (SQS), and Amazon Elastic Load Balancing (ELB), have been built with fault tolerance and high availability in mind. Services that provide basic infrastructure, such as Amazon Elastic Compute Cloud (EC2) and Amazon Elastic Block Store (EBS), provide specific features, such as availability zones, elastic IP addresses, and snapshots, that a fault-tolerant and highly available system must take advantage of and use correctly. Just moving a system into the cloud doesn't make it fault-tolerant or highly available.



amazon web services

System Overview

1 Load balancing is an effective way to increase the availability of a system. Instances that fail can be replaced seamlessly behind the load balancer while other instances continue to operate. **Elastic Load Balancing** can be used to balance across instances in multiple availability zones of a region.

2 **Availability zones (AZs)** are distinct geographical locations that are engineered to be insulated from failures in other AZs. By placing **Amazon EC2** instances in multiple AZs, an application can be protected from failure at a single location. It is important to run independent application stacks in more than one AZ, either in the same region or in another region, so that if one zone fails, the application in the other zone can continue to run. When you design such a system, you will need a good understanding of zone dependencies.

3 **Elastic IP** addresses are public IP addresses that can be programmatically mapped between instances within a region. They are associated with the AWS account and not with a specific instance or lifetime of an instance.
Elastic IP addresses can be used to work around host or availability zone failures by quickly remapping the address to another running instance or a replacement instance that was just started. Reserved instances can help guarantee that such capacity is available in another zone.

4 Valuable data should never be stored only on instance storage without proper backups, replication, or the ability to re-create the data. **Amazon Elastic Block Store (EBS)** offers persistent off-instance storage volumes that are about an order of magnitude more durable than on-instance storage. EBS volumes are automatically replicated within a single availability zone. To increase durability further, point-in-time snapshots can be created to store data on volumes in **Amazon S3**, which is then replicated to multiple AZs. While EBS volumes are tied to a specific AZ, snapshots are tied to the region. Using a snapshot, you can create new EBS volumes in any of the AZs of the same region. This is an effective way to deal with disk failures or other host-level issues, as well as with problems affecting an AZ. Snapshots are incremental, so it is advisable to hold on to recent snapshots.