Regions, availability zones, data centers, racks, servers

CPU-bound (compute-bound)
needs a lot of CPU resources to process requests

general-purpose server

web application

database

memory-bound and disk **I/O-bound**
needs a lot of memory and large low latency disk storage

compute-optimized server

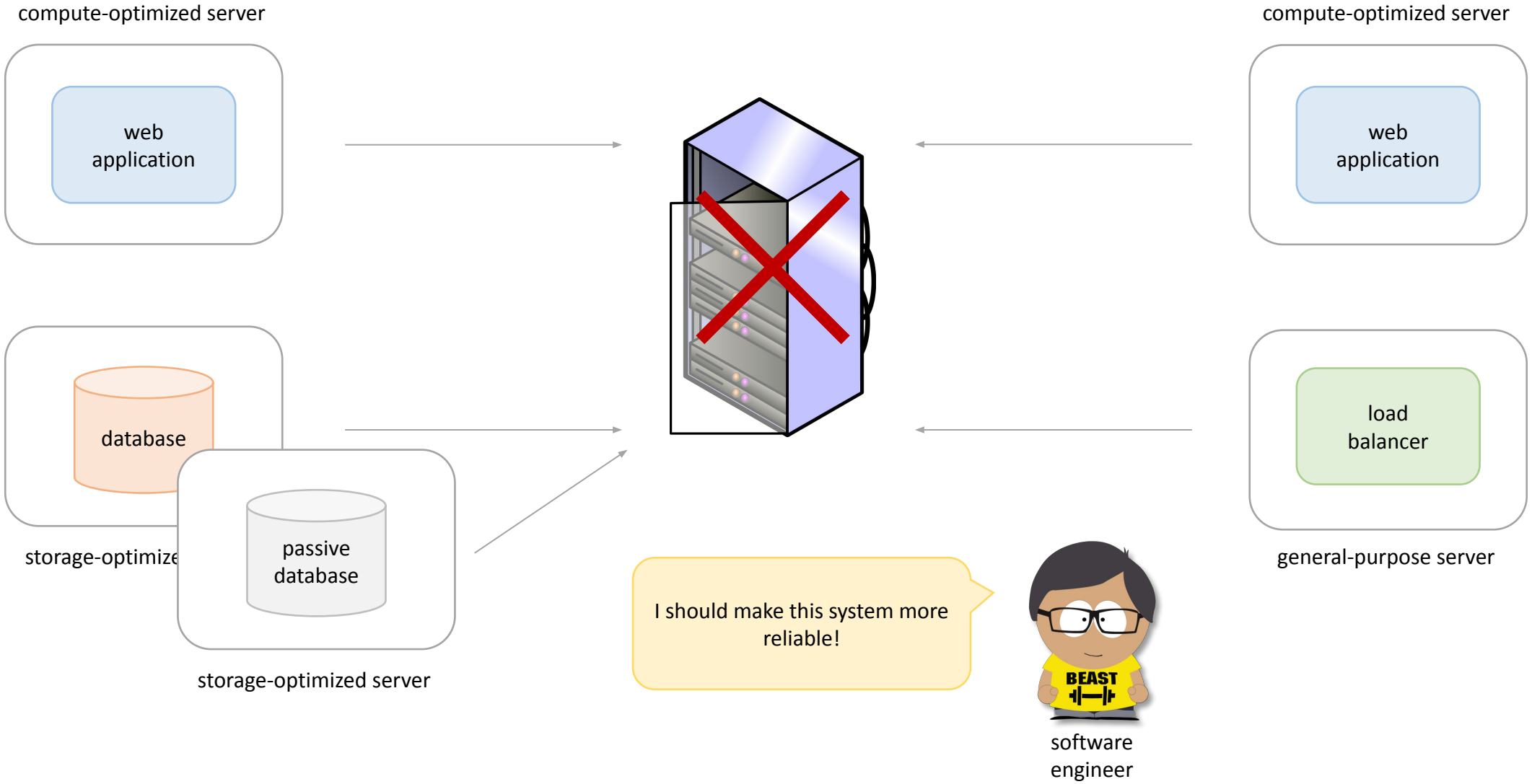
database

storage-optimized server

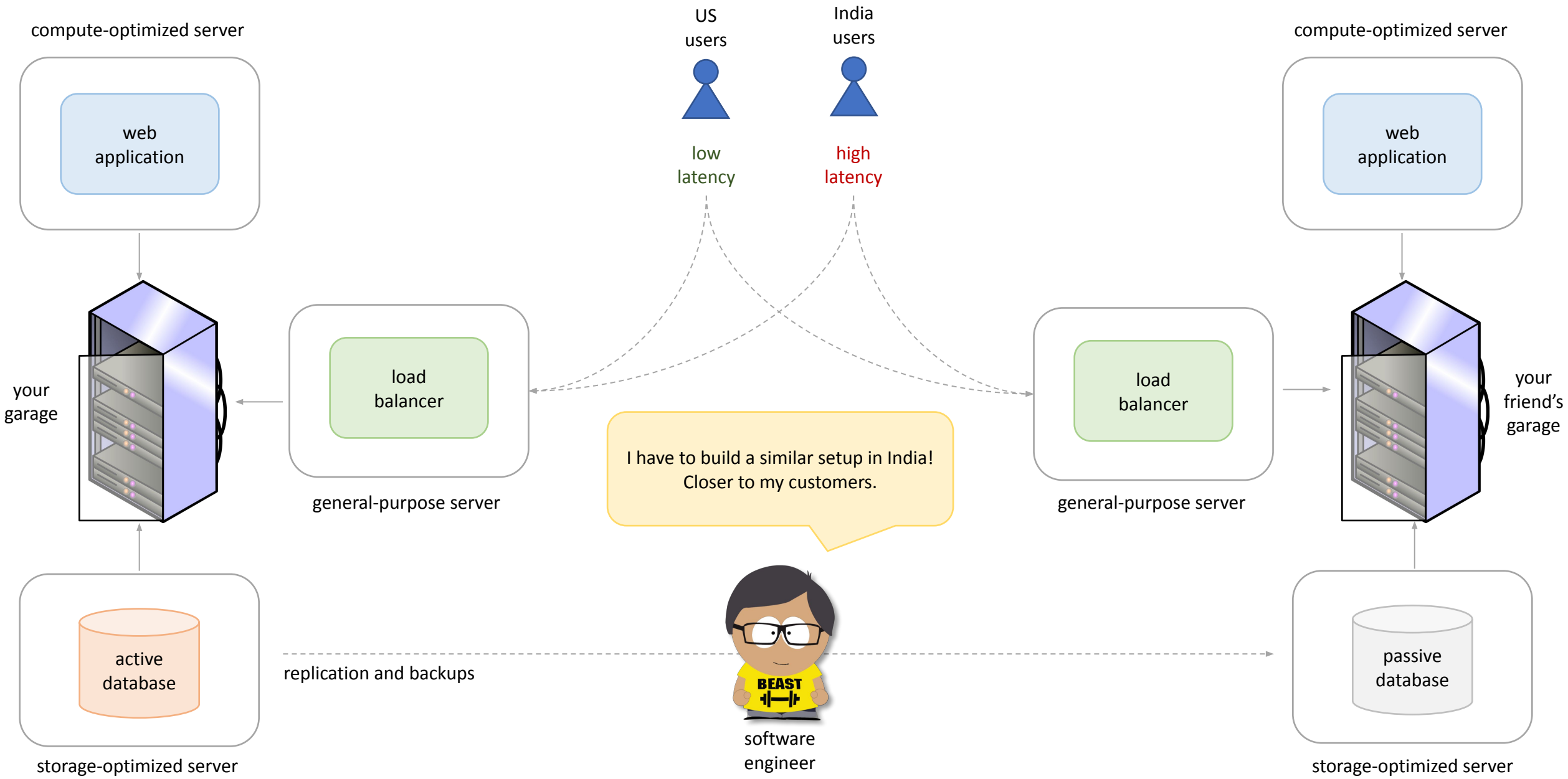
I should make this system more reliable!

software engineer

Regions, availability zones, data centers, racks, servers

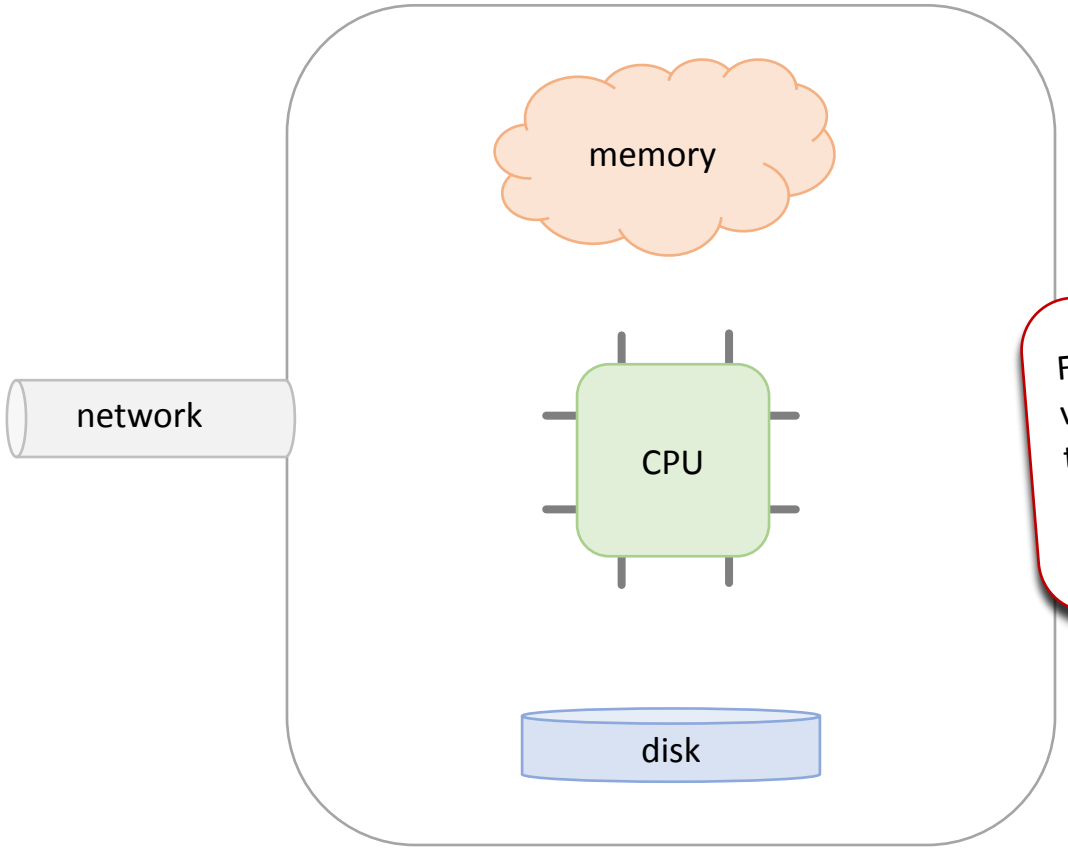


Regions, availability zones, data centers, racks, servers



Regions, availability zones, data centers, racks, servers

server



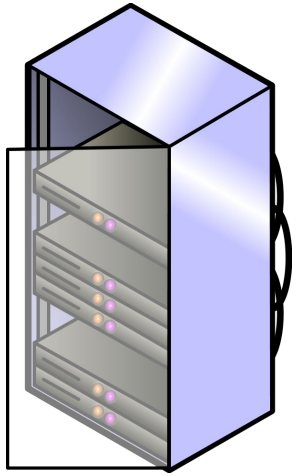
For every system component (e.g. microservice), we choose servers based on what resources the component needs the most.

This simplifies future scaling and reduces costs.

Regions, availability zones, data centers, racks, servers

rack

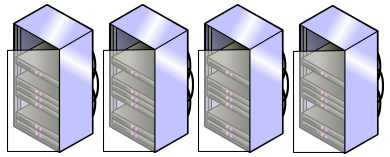
- Servers are physically easier to reach, examine, and manipulate.
- Simplifies cooling and increases security.
- Has its own network and power source.



- To increase availability, we can place servers in different racks.
- To reduce latency, we can place servers in the same rack.

Regions, availability zones, data centers, racks, servers

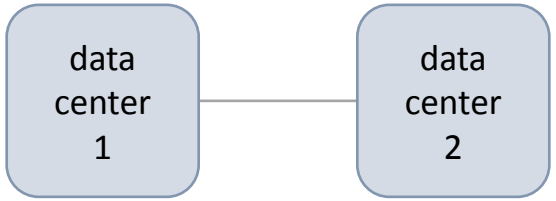
data center



- Has independent power, cooling, and physical security.
- May become unavailable due to power outage, earthquake.

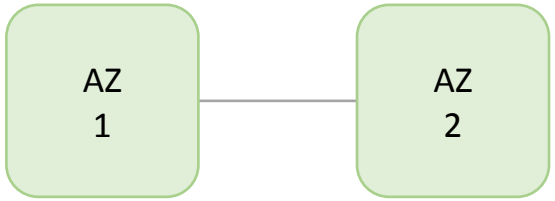
availability zone

(AZ)



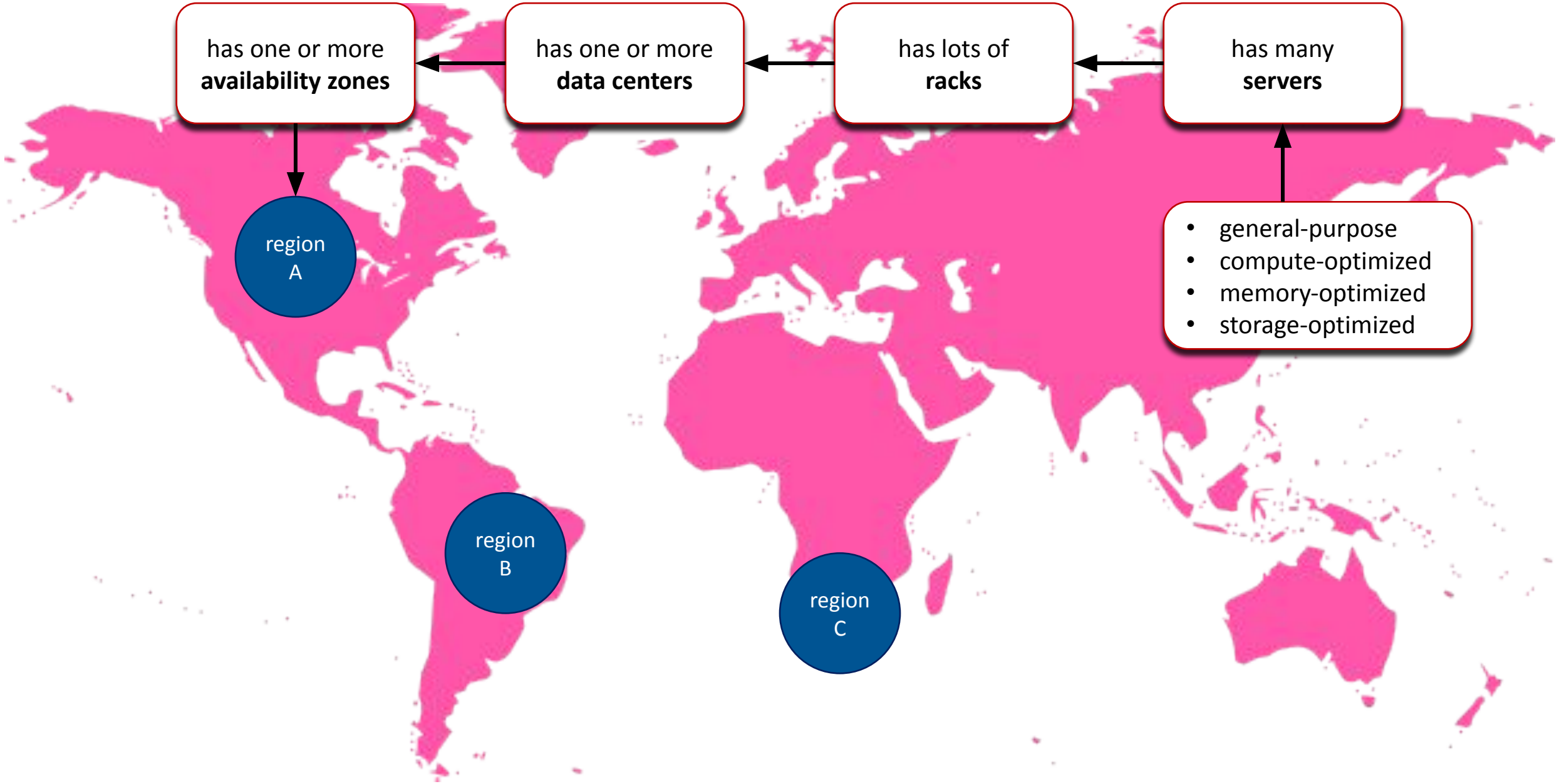
- Increases availability as hardware is distributed across multiple data centers.
- Increases scalability as there are multiple places to allocate hardware from.

region

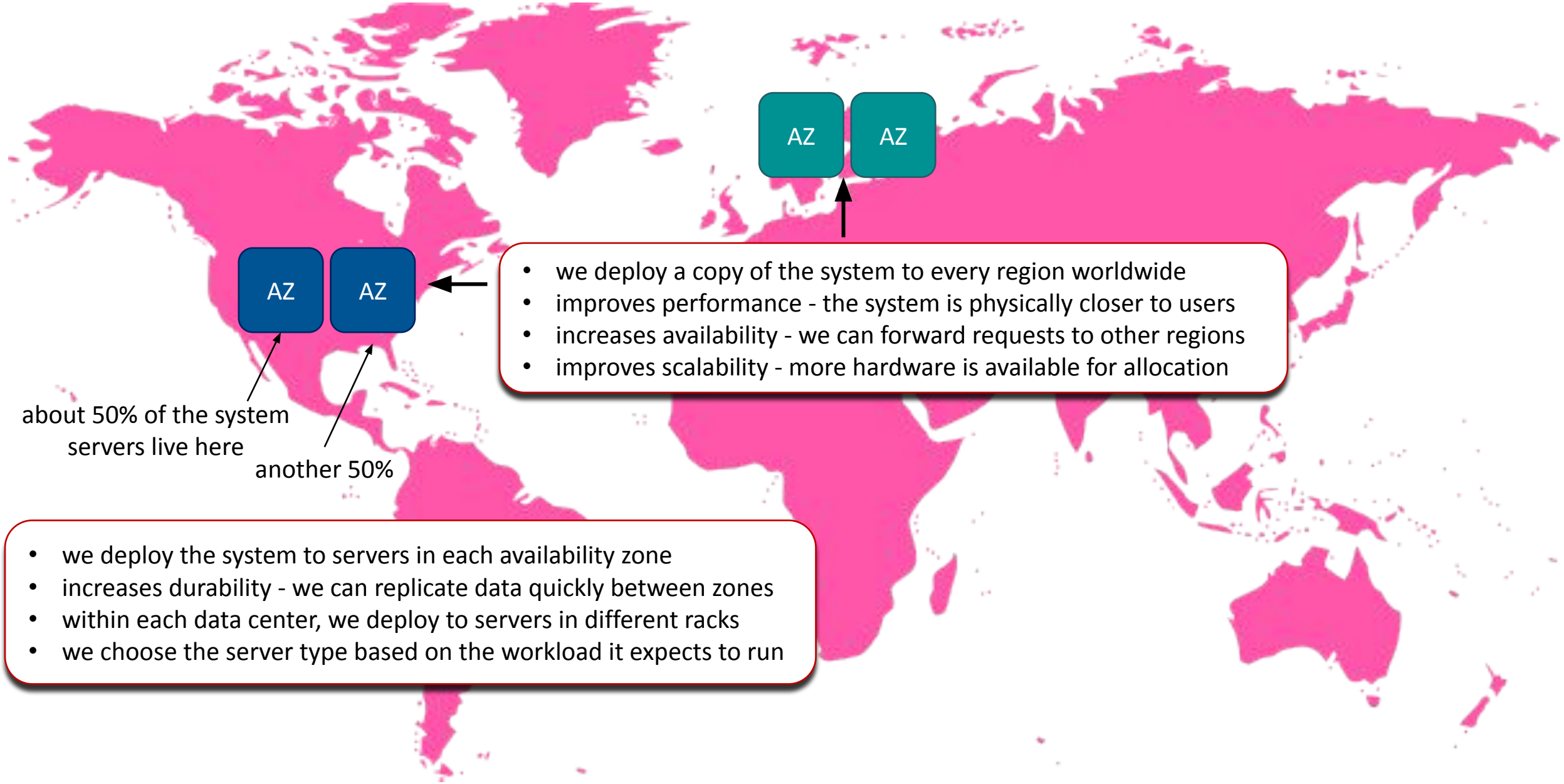


- Within a radius of 100 km.
- AZs in a region are interconnected with high-bandwidth and low-latency networking.
- Network latency between AZs is less than 2 ms.

Regions, availability zones, data centers, racks, servers



Regions, availability zones, data centers, racks, servers



- we deploy a copy of the system to every region worldwide
- improves performance - the system is physically closer to users
- increases availability - we can forward requests to other regions
- improves scalability - more hardware is available for allocation

- we deploy the system to servers in each availability zone
- increases durability - we can replicate data quickly between zones
- within each data center, we deploy to servers in different racks
- we choose the server type based on the workload it expects to run