Scalability is the property of the a system to handle a growing amount of load by adding more resources.

Ways for System growth - More users, more features, more data, more complexity and more geographies.

Ways to scale a system

1. Vertical scaling (scale up) - add more power (RAM, CPU, storage).
2. Horizontal scaling (scale out) - add more machines.
3. Load balancing - distributing traffic across multiple servers.
4. Caching - store frequently accessed data in-memory to reduce load on servers.
5. CDNs - store static assets closer to users.
6. Partitioning - split data or functionality across multiple nodes or servers.
7. Async communication - defer long running or non-critical tasks to background queues or message brokers.
8. Microservices - Break down application into smaller, independent services that can be scaled independently.
9. Auto-scaling - automatically adjust number of active servers based on load.
10. Multi region deployment - deploy app in multiple data centers or cloud regions.