**System Design**

**1. What is System Design?**

**System design is the process of creating a blueprint for a system, outlining its architecture, components, and interactions to achieve specific goals and meet defined requirements. It involves considering factors like scalability, reliability, performance, and maintainability to ensure a robust and efficient system.**

**2. Key Concepts in System Design**

* **Scalability: The ability of a system to handle increasing workloads.** 
  + **Vertical Scaling: Enhancing existing machines with more resources (e.g., CPU, RAM).**
  + **Horizontal Scaling: Adding more machines to distribute the workload.**

* **Reliability: Ensuring the system continues to function even in the face of failures.** 
  + **Achieved through techniques like redundancy, failover mechanisms, and replication.**

* **Performance: The speed and efficiency of the system.** 
  + **Improved through caching, indexing, and optimized algorithms.**

* **Consistency: Ensuring all data across the system is synchronized.** 
  + **Strong Consistency: All reads reflect the latest writes immediately.**
  + **Eventual Consistency: Data eventually converges to the same state across all nodes.**
* **Availability: The system's uptime and accessibility.** 
  + **High availability is achieved through load balancing and failover mechanisms.**

**3. System Design Process**

1. **Requirements Gathering: Define both functional (what the system should do) and non-functional requirements (scalability, performance, etc.).**
2. **High-Level Design: Create a high-level architecture diagram outlining major components and their interactions.**
3. **Low-Level Design: Detail the internal workings of each component, including data structures and algorithms.**
4. **Implementation: Develop the system based on the design specifications.**
5. **Testing: Validate the system against requirements to ensure it meets expectations.**
6. **Deployment: Release the system to production and monitor its performance.**

**4. Types of System Design**

* **High-Level Design (HLD): Focuses on overall system architecture, modules, and components.**
* **Low-Level Design (LLD): Focuses on the internal details of individual components, class diagrams, and specific logic.**

**5. Core Components of System Design**

* **Load Balancers: Distribute traffic across multiple servers to enhance performance and reliability.**
* **Caching: Store frequently accessed data to reduce latency.**
* **Databases:** 
  + **Relational: Structured data with a defined schema (e.g., MySQL, PostgreSQL).**
  + **NoSQL: Unstructured or semi-structured data for flexible scalability (e.g., MongoDB, Cassandra).**
* **Message Queues: Enable asynchronous communication and decouple system components.**
* **CDNs (Content Delivery Networks): Deliver content to users from servers closest to their location.**

**6. Steps to Design a System**

1. **Gather Requirements: Define both functional and non-functional requirements.**
2. **Define APIs: Design RESTful APIs for communication between components.**
3. **High-Level Architecture: Identify and define core system components.**
4. **Database Design: Choose the appropriate database type (SQL or NoSQL) and design the data schema.**
5. **Component Design: Detail the internal workings of each component.**
6. **Scaling and Optimization: Implement scaling strategies (sharding, replication, caching) to improve performance.**

**7. Example: Designing a URL Shortener**

1. **Requirements:** 
   * **Users can input a long URL and receive a shortened version.**
   * **Shortened URLs should redirect to the original URL.**
   * **Track the number of times a shortened URL is accessed.**
   * **Provide an API for programmatic access.**
2. **High-Level Design:** 
   * **Web server to handle user requests.**
   * **Database to store URL mappings and access counts.**
   * **Caching layer to improve performance for frequently accessed URLs.**
3. **Low-Level Design:** 
   * **Define data models, API endpoints, and error handling.**

**Common Architectural Patterns**

* **Microservices: Breaking down a large system into small, independent services.**
* **Monolithic: A single, self-contained application.**
* **Client-Server: A client interacts with a server to request and receive data.**
* **Three-Tier Architecture: Presentation layer, business logic layer, and data access layer.**

**Tools and Technologies**

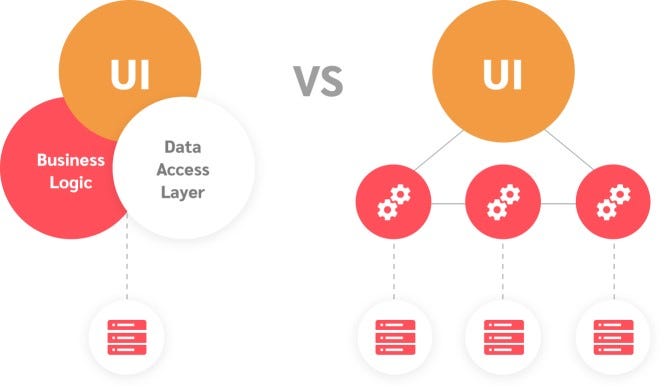
1. **Cloud Platforms: AWS, Azure, GCP**
2. **Containerization: Docker, Kubernetes**
3. **Databases: Relational (MySQL, PostgreSQL), NoSQL (MongoDB, Cassandra)**
4. **Caching: Redis, Memcached**
5. **Message Queues: Kafka, RabbitMQ**
6. **API Gateways: Kong, Apigee**

**8. Common Mistakes to Avoid**

* **Over-engineering: Designing a system that is overly complex for the given requirements.**
* **Ignoring non-functional requirements: Neglecting aspects like scalability, reliability, and performance.**
* **Poor planning for scaling: Failing to anticipate future growth and design for scalability from the outset.**
* **Poor API design: Creating APIs that are difficult to use, understand, or maintain.**

**This refined version aims to be more concise, while still covering the essential concepts of system design effectively.**

**Monolithic vs. Distributed Systems**



**1. Monolithic Systems**

A monolithic system is an architecture where all the components of the software are tightly integrated into a single unit. It works as one cohesive application, with all functionalities running in a single process or application.

**Key Characteristics:**

* Single codebase and application.\n
* All components (UI, business logic, and database) are tightly coupled.\n
* Deployed as a single unit.

**Advantages:**

1. **Simplicity:** Easier to develop and deploy initially.\n
2. **Performance:** Communication within the system is fast as it happens within the same process.\n
3. **Ease of Testing:** Testing is more straightforward due to fewer moving parts.

**Disadvantages:**

1. **Scalability Issues:** Scaling requires replicating the entire system.\n
2. **Tight Coupling:** Changes in one module often impact others.\n
3. **Maintenance Challenges:** As the application grows, it becomes harder to manage.

**Example:** A traditional e-commerce platform where the user interface, product catalog, order management, and payment processing all reside in one application.

**2. Distributed Systems**

A distributed system is an architecture where components are split across multiple, independent services or nodes that communicate over a network. Each component can run on separate machines.

**Key Characteristics:**

* Composed of loosely coupled services or nodes.\n
* Services communicate via APIs or messaging systems.\n
* Each service can have its own database.

**Advantages:**

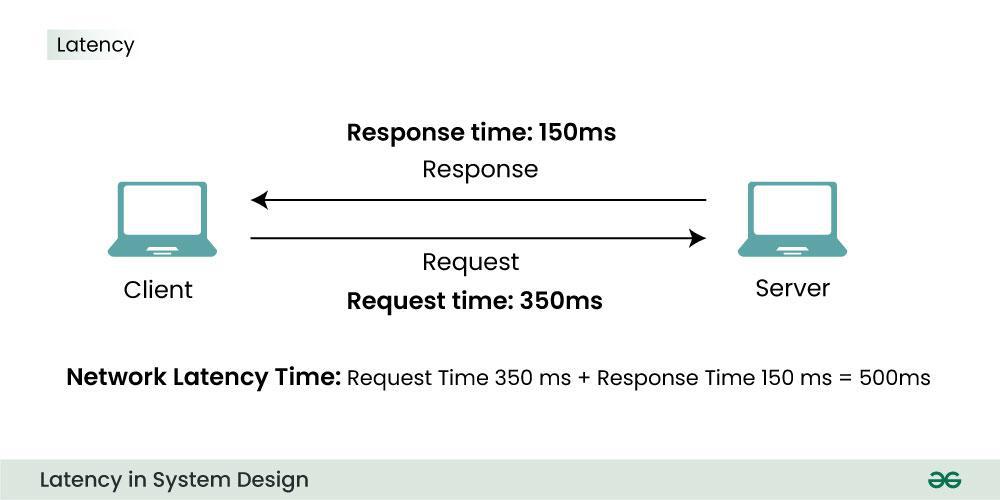
1. **Scalability:** Services can be scaled independently.\n
2. **Fault Tolerance:** A failure in one service doesn't necessarily bring down the entire system.\n
3. **Flexibility:** Different components can use different technologies.

**Disadvantages:**

1. **Complexity:** More challenging to develop, deploy, and maintain.\n
2. **Latency:** Communication between services introduces network delays.\n
3. **Consistency:** Ensuring data consistency across services can be complex.

**Example:** A modern microservices-based e-commerce platform where the user interface, product catalog, order management, and payment processing are separate services communicating via APIs.

**What is Latency in System Design?**



Latency refers to the time delay between a request being sent and the corresponding response being received in a system. It measures the time it takes for data to travel from the source to the destination and back, including any processing time involved. In system design, latency is a critical performance metric that directly impacts the user experience.

**Components of Latency**

1. **Network Latency**:
   * The time taken for data to travel across the network between systems.
   * Influenced by factors like bandwidth, distance, and network congestion.
2. **Processing Latency**:
   * The time spent processing a request at a server or application level.
   * Includes database queries, API handling, and other computational tasks.
3. **Disk Latency**:
   * The time required to read or write data to storage devices.
   * Depends on the type of storage (e.g., HDD, SSD).
4. **Queue Latency**:
   * The delay caused by requests waiting in a queue before being processed.
   * Common in systems with high traffic or limited resources.

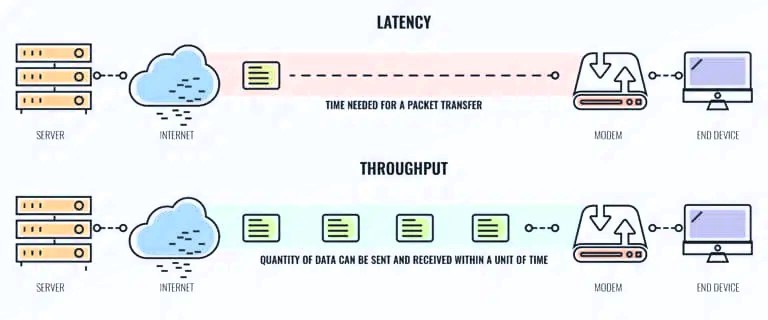
**Latency vs. Throughput**

* **Latency**: Measures the time to process a single request.
* **Throughput**: Measures the number of requests processed per unit of time.
* Optimizing for one can sometimes degrade the other, so balance is key.

**Key Strategies to Reduce Latency**

1. **Caching**: Store frequently accessed data closer to the user to minimize retrieval time.
2. **Load Balancing**: Distribute requests across multiple servers to prevent bottlenecks.
3. **Content Delivery Networks (CDNs)**: Use geographically distributed servers to reduce network latency.
4. **Asynchronous Processing**: Handle tasks that don’t require immediate user feedback separately.
5. **Database Optimization**: Optimize queries, use indexes, and adopt sharding or replication.
6. **Compression**: Reduce the size of data being transmitted over the network.
7. **Edge Computing**: Process data closer to where it's generated rather than relying on centralized servers.

**What is Throughput in System Design?**



Throughput refers to the number of tasks, operations, or transactions a system can process in a given period. It measures the system's capacity to handle workloads effectively and is typically expressed as "requests per second (RPS)" or "transactions per minute (TPM)."

**Key Characteristics of Throughput**

1. **Rate of Work**: Indicates the efficiency of the system in processing requests or data.
2. **Dependent on Latency**: While latency measures the time to process a single request, throughput is influenced by how quickly multiple requests can be handled concurrently.
3. **Influenced by Bottlenecks**: Any bottleneck in the system (e.g., network, CPU, database) can reduce throughput.

**Throughput vs. Latency**

* **Throughput** focuses on volume—how many tasks are completed in a period.
* **Latency** focuses on speed—how long a single task takes to complete.

For example:

* A system with low latency may handle individual tasks quickly but could have low throughput if it can only handle a few tasks at a time.
* A high-throughput system processes many tasks simultaneously but may not guarantee low latency for each task.

**Factors Affecting Throughput**

1. **Concurrency**:
   * The ability of a system to handle multiple tasks at the same time.
   * Achieved through multi-threading, parallel processing, or distributed systems.
2. **Resource Availability**:
   * Includes CPU, memory, disk I/O, and network bandwidth.
   * Insufficient resources lead to bottlenecks, reducing throughput.
3. **Workload Patterns**:
   * A system designed for consistent workloads may struggle with sudden spikes, affecting throughput.
4. **Scaling**:
   * Horizontal (adding more machines) or vertical (upgrading existing machines) scaling improves throughput.

**Strategies to Maximize Throughput**

1. **Load Balancing**:
   * Distribute incoming requests across multiple servers evenly.
2. **Efficient Resource Utilization**:
   * Optimize the use of CPU, memory, and other system resources.
3. **Asynchronous Processing**:
   * Offload non-critical tasks to background processes or queues.
4. **Batch Processing**:
   * Group multiple small tasks together to process them more efficiently.
5. **Database Optimization**:
   * Use indexing, partitioning, and efficient query design to reduce database response times.
6. **Horizontal Scaling**:
   * Add more servers or instances to handle increased traffic.

**Example**

* **High Throughput**: A streaming service like Netflix processes thousands of video streams simultaneously.
* **Low Throughput**: A simple file upload system that processes one file at a time.