Latency

We all have encountered multiple websites or web applications with high loading times like excessive buffering while streaming a video. These interruptions lead to a poor user experience. In short, this loading time is latency. In case of higher latency, it tends to slow the website load time to crawl and take up a lot of time before loading the complete web page. It is clear that a user would prefer a low latency website with faster responses.

It is very clear how latency affects the user experience and therefore largely impacts the system design. Let's understand the term better.

What is latency?

In short, latency means delay.

- Latency is the delay between a user's request and the web application's response to the corresponding action.
- It is the total time taken for a round trip of a data packet from one destined location to another.

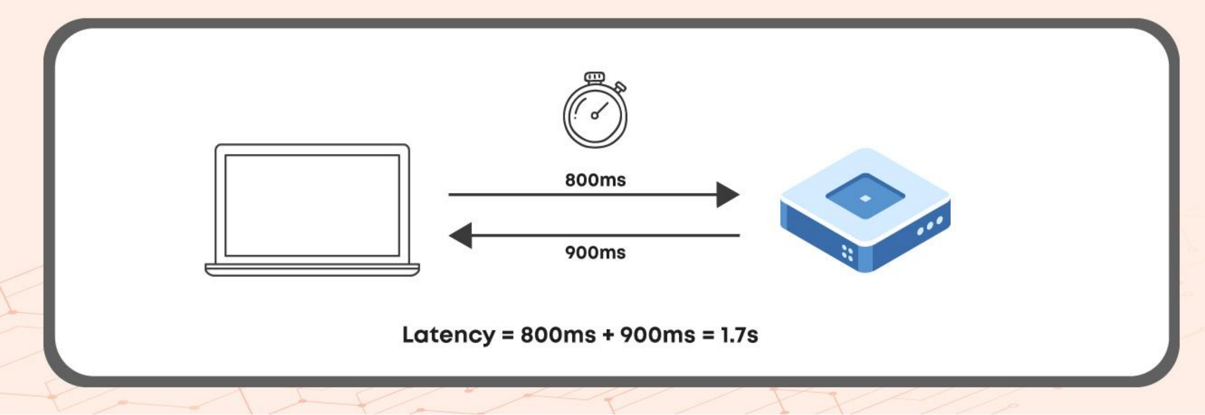

Fig: Example showing latency calculation (sum of time taken for round trip)

The unit of latency is milliseconds(ms) since it is a measure of time.

Main components that affect latency

Latency is majorly due to the Network Delays(for a distributed system) and the Computational Time.

These are the main components that affect latency:

1. Transmission media:
   Latency depends on the type of media that is being transmitted.
2. Packet size:
   Smaller the packet size, the faster the transmission. Packet size is directly proportional to latency.
3. Packet loss:
   Latency can be adversely affected by a huge loss of packets during transmission or different rates of transmission of each packet.
4. Signal strength:
   Poor signal increases latency. Latency is indirectly proportional to signal strength.
5. Propagation delays:
   Latency also depends on the distance between two communicating nodes. Higher the distance more the latency.
6. Other computer and storage delays:
   Retrieving and processing the stored information consumes time. More the time in accessing stored information greater the latency.

Latency and Architecture

Latency depends on two factors majorly:

1. Network Delays
2. Computational Delays

Monolithic architecture

Network latency is zero because there are no calls over the network.
All the calls in monolithic architecture are local.

Latency = Computational Delay + ~~Network Delay~~ (zero)

Distributed Systems

During a call in a distributed system, signals are sent over different networks and transmitted back. This further adds network latency.

Latency = Computational Delay + Network Delay

Therefore we can conclude that,

Latency (Distributed System) > Latency (Monolithic System)

Importance of Latency

1. Latency has a significant impact on User Experience and User Satisfaction. User satisfaction increases with a decrease in response time.
2. Latency is also a very important factor for latency-sensitive applications.

Some examples of latency-sensitive applications are:

1. Capital Market
   Low latency is most important in the capital or stock market where each second affects the decisions and increases the profitability of trades. High latency can adversely affect the complete market and make the trading process slow leading to huge losses.
2. Vehicles
   Vehicles are largely dependent on edge computing and low latency. A feature like an airbag must be activated without delay. Time-sensitive features in vehicles must have low latency.

Reducing Latency

1. Use of CDN(Content Delivery Network):
   CDN helps to reduce latency. CDN servers are located at different points in order to reduce the distance between users and data doesn't need to travel long distances thereby saving time.
2. Upgrading computer hardware/software:
   Upgrading or tuning the computer hardware, software or mechanical systems can help reduce the computational delays which help in reducing the latency.
3. Caching
   In computers, a cache is a high-speed data storage layer that caches a chunk of data that is typically transient so that subsequent requests for that data can be

delivered up faster than if the data were accessed directly from its primary storage location. This also reduces latency.