# 10 Key Performance Indicators Every Engineer Should Know

DesignGurus.io

# Time to First Byte (TTFB)

TTFB measures the time taken from the moment a client sends a request to a server until the client receives the first byte of data from the server.

**Example**: A website’s TTFB is 200 milliseconds, indicating a fast initial response from the server.

# Throughput

Throughput measures the number of operations or requests processed by a system per unit of time.

**Example**: An API handling 500 requests per second.

# Latency

Latency is the time taken for a request to travel from the sender to the receiver and for the response to travel back.

**Example**: The time taken for a user’s request to reach a server and receive a response is 100 milliseconds.

# Response Time

Response time is the total time taken for a system to process a request, including the time spent waiting in queues and the actual processing time.

**Example**: A database query takes 250 milliseconds to complete, including 50 milliseconds spent in the queue.

# Error Rate

Error rate is the percentage of requests that result in errors, such as timeouts or failures, compared to the total number of requests.

**Example**: Out of 10,000 requests, 100 fail, resulting in an error rate of 1%.

# Mean Time Between Failures (MTBF)

MTBF is the average time between system failures or disruptions.

**Example**: A server has an MTBF of 30,000 hours, which means it is expected to operate without failure for an average of 30,000 hours.

# Mean Time to Repair (MTTR)

MTTR measures the average time taken to repair or recover from a system failure.

**Example**:  A system has an MTTR of 2 hours, indicating that it takes an average of 2 hours to restore operations after a failure.

# Network Bandwidth

Network bandwidth is the maximum rate of data transfer across a network connection.

**Example**:  A network connection with a bandwidth of 100 Mbps can transfer 100 megabits of data per second.

# Request Rate

Request rate is the number of requests received by a system per unit of time.

**Example**:  A web server receives 300 requests per minute during peak hours.

# Concurrent Connections

Concurrent connections represent the number of active connections to a system at a given moment.

**Example**:  A database server can handle 5,000 concurrent connections without performance degradation.

➡ Learn system design basics: **Grokking System Design Fundamentals**

➡ Learn about system design interview questions: **Grokking the System Design Interview**

**@ DesignGurus.io**