SYSTEM DESIGN

সিস্টেম পারফরম্যান্স মেট্রিক্স:

Throughput && Latency

Throughput কী?

Throughput হচ্ছে কোনো সিস্টেম নির্দিষ্ট সময়ে কতগুলো কাজ সম্পন্ন করতে পারে, তার পরিমাপ। তার মানে সহজ ভাষায় বলতে গেলে আপনার সিস্টেম "Per second এ কতটা request, data, বা transaction handle করতে পারছে সেটাকে বুঝায়

উদাহরণ:

- আপনার server প্রতি second-এ 500 টা request serve করতে পারে

➡️ Throughput = 500 req/sec

- কোনো payment gateway এক মিনিটে 1000 টা payment প্রসেস করতে পারে

➡️ Throughput = 1000 tx/min

- Data transfer: 200 Mbps

➡️ Throughput = 200 megabits per second

THROUGHPUT আমাদের কেন দরকার হয়

Throughput জানলে আপনি বুঝতে পারবেন:

1. আপনার system কতটা efficient
2. High traffic আসলে system crash করবে নাকি চালু থাকবে
3. Load balance ঠিকঠাক কাজ করছে কিনা
4. Scaling করা লাগবে কি না

Basic Formula:

$$\text{Throughput} = \frac{\text{Total number of successful operations}}{\text{Total time taken}}$$

LEARN HUNTER

কি কি টুলস আছে থ্রুপুট মাপার জন্য চলুন দেখি

1. Apache Benchmark (ab)
2. JMeter
3. Locust
4. Gatling
5. Prometheus + Grafana (live monitoring)

Throughput না থাকলে বা মেপে না দেখলে কী হতে পারে?

- User বেশি হলে system slow হয়ে যাবে
- Sudden traffic এ server hang/crash করতে পারে
- কোথায় system আটকে যাচ্ছে বুঝতে পারবোনা

THROUGHPUT বাড়ানোর উপায়

Load ব্যালেন্সিং এর মাধ্যমে Traffic ভাগ করে multiple server এ দিয়ে দেয়া

Caching এর মাধ্যমে বারবার একই data নিয়ে আশা হলে দ্রুত রেসপন্স হবে

Queue System এর মাধ্যমে Async process করে throughput বাড়ানো

DB Optimization && Code Optimization

Latency কী?

Latency হলো কোনো request পাঠানোর পর সেইটার response আসতে যত সময় লাগে, সেটা।
যদি আরো সহজ ভাষায় আপনাকে বলি
"Latency মানে – কাজটা শুরু হয়ে শেষ হতে যে delay বা response time লাগে।"

উদাহরণ:

- আপনি একটা website খুললেন, আর সেটা 3 সেকেন্ডে load হলো → Latency = 3s
- আপনি mobile এ SMS পাঠালেন, আর সেটা 1.2 সেকেন্ডে পৌঁছালো → Latency = 1.2s
- Database query দিলে 200ms পরে result আসে → Latency = 200ms

Latency কেন দরকার?

User Experience → Response ধীর হলে user বিরক্ত হয়

System Responsiveness → কাজের গতি পরিমাপ করা যায়

Problem Diagnosis → কোথায় delay হচ্ছে, সেটা ধরতে latency জরুরি

Optimization Planning → কোন অংশ slow সেটি detect করে improve করা যায়

Latency কিভাবে মাপা হয়?

- Ping → Network latency দেখায়
- Postman / cURL → API response time
- Chrome Dev Tools → Page load latency
- JMeter / K6 / Locust → Backend latency
- Prometheus + Grafana → Realtime latency monitoring

মাপার পদ্ধতি (basic):

$$\text{Latency} = \text{Response Received Time} - \text{Request Sent Time}$$

LEARN HUNTER

LATENCY কোথায় দরকার হয়?

সিস্টেম	Latency গুরুত্বপূর্ণ কেন?
Web API	User যত দ্রুত response পায়, তত ভালো
Video Streaming	Buffer কমাতে latency কম হতে হবে
Online Gaming	Real-time interaction দরকার (e.g., 30ms latency)
E-commerce Checkout	Order delay মানেই revenue loss
Messaging App	Message send delay হলে ভুল communication হয়

Latency কমানোর উপায়:

উপায়	ব্যাখ্যা
✅ CDN ব্যবহার	Static content দ্রুত পৌঁছে যায়
✅ Caching	আগের ডেটা cache করে রাখা যায়
✅ Database Indexing	Query execution দ্রুত হয়
✅ Load Balancing	Proper routing দিয়ে delay কমানো যায়
✅ Code Optimization	Logic ছোট ও clean রাখলে delay কমে

LEARN HUNTER

শেষ কথা:

সুতরাং এই ছিল আজকের ছোট আলোচনা আমরা System Performance Metrics খুব ভালো করে জানলাম এগুলো আমাদের কি? কেন দরকার উপকারিতা গুলো কি কি। পরবর্তীতে আমরা আমাদের প্রয়োজনীয় আরো বিষয় বস্তু গুলো নিয়ে জন্য।

LEARN HUNTER