Core Terms

🧠 **System Design Core Terms**

1. Latency (⏱️ Response Time)

📌 **Latency কী?**

👉 **একটা request পাঠানোর পর response আসতে যত সময় লাগে।**

Example:

- তুমি WhatsApp এ message পাঠালে
- "Delivered" দেখাতে যদি 150ms লাগে → latency = 150ms

🔍 **কোথা থেকে latency আসে?**

- Network delay
- Server processing
- Database query
- Cache miss

🎯 **FAANG expects:**

"We aim for p95 latency < 200ms"

📌 **p95 মানে**:
95% request এই সময়ের ভেতর শেষ হয়।

2. Throughput (🚀 Capacity)

📌 **Throughput কী?**

👉 **এক সেকেন্ডে system কয়টা request handle করতে পারে।**

Example:

- System handle করে
    - 5,000 requests/sec
      → Throughput = 5k RPS

🎯 **Real-life analogy:**

- Latency = গাড়ি কত দ্রুত
- Throughput = একসাথে কয়টা গাড়ি চলতে পারে

📌 FAANG tip:

- High throughput ≠ low latency
- দুটো আলাদা জিনিস

3. SLA (📜 Promise)

📌 **SLA কী?**

👉 **Service Level Agreement**
মানে company যে promise দেয় user-কে।

Example:

- “99.9% uptime monthly”
- “Response time < 300ms”

📌 SLA ভাঙলে:

- Penalty
- Trust loss

🎯 **Interview sentence:**

“Our SLA requires 99.9% availability, so we design with redundancy.”

4. Availability (✅ Uptime)

📌 Availability কী?

👉 System কত সময় usable থাকে।

Formula:

Availability = Uptime / (Uptime + Downtime)

🔢 Common numbers:

- 99% → ~3.65 days downtime/year
- 99.9% → ~8.7 hours/year
- 99.99% → ~52 minutes/year

📌 FAANG insight:

More availability = more cost

🔁 Relation between them (IMPORTANT)

Term	Focus
Latency	কত দ্রুত
Throughput	কত বেশি
Availability	কত সময় চালু
SLA	কতটা promise

📌 Example sentence (interview GOLD):

“To meet our SLA of 99.9% availability, we replicate services across regions, which slightly increases latency.”

🎯 One-line Memory Trick 🧠

- Latency = **speed**
- Throughput = **capacity**
- Availability = **uptime**
- SLA = **promise**

🚀 FAANG-level Tip

Interviewer খুশি হয় যখন তুমি বলো:

- “We trade latency for availability here”
- “Throughput increases via horizontal scaling”