Service Level Agreement (SLA) – Revision Notes

- **Definition**
  - **SLA**: A **formal contract** between a **Service Provider (SP)** and **Service Consumer (SC)** that defines **guaranteed service performance**.
  - Basis of **consumer trust** in the provider.
- **Purpose**
  - Define **formal basis** for performance, availability, and QoS.
  - Contains **Service Level Objectives (SLOs)**: **measurable conditions**.
- **SLA Contents**
  - List of services to be delivered.
  - Specific definitions of each service.
  - Roles/responsibilities of SP & SC.
  - Metrics for performance measurement.
  - Auditing mechanism.
  - Remedies for SLA violations.
  - SLA versioning/change mechanism.
- **Service Level Objectives (SLOs)**
  - **Measurable QoS parameters** like:
    - Availability (e.g., 99.9%)
    - Response time (e.g., 3-5 sec)
    - Throughput, penalties, billing

Web Service SLA vs Cloud SLA

Feature	Web Service SLA	Cloud SLA
QoS Parameters	Response time, reliability, cost	Security, privacy, trust, automation
Automation	Manual SLA negotiation	Automated provisioning and monitoring
Resource Allocation	Via UDDI (central registry)	Globally distributed without central registry

- **SLA Types**
  1. **Off-the-shelf / Non-negotiable SLA**:
     - Predefined by SP, not customizable.
     - Common in state-of-the-art clouds.
  2. **Negotiable SLA**:
     - Negotiated via agents.
     - Suitable for mission-critical services.

- **Web Service SLA Standards**
  - **WS-Agreement**:
    - XML-based for SLA negotiation & management.
    - Runtime monitoring & violation detection.
  - **WSLA (Web Service Level Agreement Framework)**:
    - XML schema for expressing & interpreting SLAs.
    - Measures QoS, reports violations.
    - Lacks formal semantic definitions.

- **Service Level Management (SLM)**
  - Tracks performance using SLOs.
  - **Provider view**: Business + tech decision-making.
  - **Consumer view**: Efficient cloud usage decisions.

- **Key Performance Indicators (KPIs)**
  - Low-level metrics: **uptime, downtime, inbytes, outbytes**.
  - Used to calculate SLOs.
    - Example: Availability = 1 – (downtime/uptime)

- **SLA Requirements**
  - Security
  - Data encryption
  - Privacy
  - Data retention & deletion
  - Hardware erasure
  - Transparency & certification
  - Monitoring & auditability

- **Limitations of SLA**
  - Service measurement complexity

- Vendor bias
- Lack of active customer-side monitoring

◇ **Expected SLA Parameters**
- **IaaS, SaaS, PaaS**, **Storage as a Service**

▧ **Cloud Computing: Economics**

◇ **Economic Properties**
- **Common Infrastructure**:
  - Shared & standardized → Statistical multiplexing.
- **Location-Independence**:
  - Reduces latency, improves UX.
- **Online Connectivity**:
  - Enables service access, quantifiable by net performance.
- **Utility Pricing**:
  - Pay-per-use model.
- **On-Demand Resources**:
  - Scalable & elastic provisioning.

◇ **Value of Common Infrastructure**
- **Economies of Scale**:
  - Lower cost via bulk purchasing.
- **Statistics of Scale**:
  - Higher utilization → Lower per-resource cost.
  - Multiplexing demand reduces SLA violation chances.

◇ **Coefficient of Variation (CV)**
- **CV = σ / |μ|** (Standard deviation / Mean)
- Measures **smoothness of demand**.
  - Lower CV → Better utilization.
- Aggregated demand:
  - Mean = $n \times \mu$, Variance = $n \times \sigma^2$

◇ **Value of On-Demand Services**
- **Penalty** = $\int |D(t) - R(t)|\, dt$
  (Difference between demand and resources)
- Flat demand → Penalty = 0
- Smooth demand → Better resource matching