

UNIT 1

Introduction to Cloud Computing

(6 Hours) — Cloud Computing (CACS402)

Disclaimer: These notes were prepared as a study aid and may contain errors or omissions. Please cross-check important definitions, figures and product details (especially vendor services such as Google App Engine, Microsoft Azure and Aneka, which change often) against your prescribed textbook, class notes and official documentation before relying on them in an exam.

1.1 Overview of Cloud Computing

Cloud computing is a model for delivering computing resources — servers, storage, databases, networking, software and analytics — over the Internet (“the cloud”) on an on-demand, pay-as-you-go basis. Instead of buying and maintaining physical hardware, users rent only what they need from a provider and release it when done.

The **NIST definition** (SP 800-145) describes cloud computing as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service-provider interaction.

Example 1.1 — Everyday analogy

Electricity: a household does not run its own power plant; it plugs into the grid and pays for the units consumed.

Cloud computing works the same way — you “plug in” to a provider's data centre and pay for the CPU hours, GB of storage or requests you actually use.

Key ideas

- **Resources as a service** — compute, storage and software are consumed, not owned.
- **Internet delivery** — reachable from any device with a network connection.
- **Elasticity** — capacity grows or shrinks automatically with demand.
- **Metered billing** — you are charged for measured usage.

1.2 Evolution of Cloud Computing

Cloud computing is the result of several decades of progress in hardware, networking and distributed systems. Each stage tried to make computing power easier to share and cheaper to use.

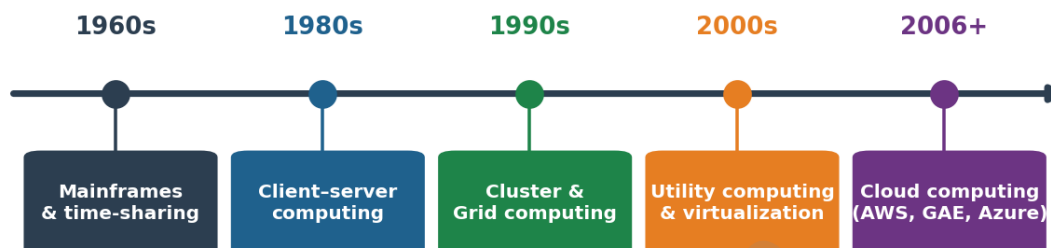


Figure 1.1 — Evolution towards cloud computing

Stage	Main idea	Limitation that led to next stage
Mainframes / time-sharing (1960s)	Many users share one powerful central computer through terminals.	Very expensive, centralised, poor flexibility.
Client-server (1980s)	Work is split between PCs (clients) and servers on a LAN.	Servers under-utilised; costly to maintain.
Cluster computing	Many similar computers linked to act as one system.	Tightly coupled, single organisation.
Grid computing (1990s)	Geographically distributed, heterogeneous resources shared for large scientific tasks.	Complex to use; no strong commercial billing model.
Utility computing	Computing sold as a metered utility (like water/electricity).	Needed better resource isolation and elasticity.
Virtualization + Web 2.0 + SOA	Many virtual machines on one host; web services; broadband Internet.	Combined to enable modern clouds.
Cloud computing (2006–)	Amazon EC2/S3 (2006), Google App Engine (2008), Microsoft Azure (2010).	—

Note: Grid computing vs cloud computing: grids focus on pooling resources across organisations for big computations; clouds focus on on-demand, virtualised, commercially billed services.

1.3 Characteristics of Cloud Computing

NIST lists five essential characteristics that a system must have to be called a cloud:

Characteristic	Meaning
On-demand self-service	Users provision resources (e.g., a VM) by themselves through a portal or API without human help from the provider.
Broad network access	Services are available over the network from thin/thick clients — laptops, phones, tablets.
Resource pooling	Provider's resources serve many customers (multi-tenancy) and are dynamically assigned; the customer usually does not know the exact physical location.
Rapid elasticity	Capacity can be scaled out/in quickly, often automatically; to the user it appears unlimited.
Measured service	Usage is monitored, controlled and reported, giving a pay-per-use billing basis.

Other commonly cited features: **multi-tenancy**, **scalability**, **reliability** and **fault tolerance**, **cost efficiency**, **location independence** and **automation**.

1.4 Types of Cloud and Cloud Services

Service models

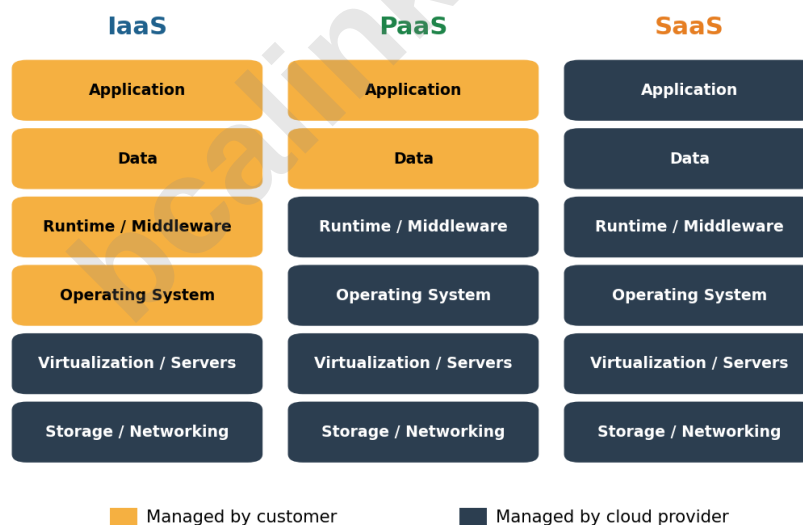


Figure 1.2 — IaaS, PaaS and SaaS: who manages what

Model	What is delivered	Example
IaaS — Infrastructure as a Service	Virtual machines, storage, networks; user controls OS and software.	Amazon EC2, Google Compute Engine, Azure VMs
PaaS — Platform as a Service	Runtime + tools to build and deploy apps; provider manages OS and servers.	Google App Engine, Azure App Service, AWS Elastic Beanstalk

Model	What is delivered	Example
SaaS — Software as a Service	Ready-to-use application accessed via browser.	Gmail, Google Docs, Microsoft 365, Salesforce

Deployment models (overview)

- **Public cloud** — owned by a provider and offered to the general public.
- **Private cloud** — operated solely for one organisation.
- **Community cloud** — shared by organisations with common concerns.
- **Hybrid cloud** — combination of two or more of the above, bound by technology allowing data/application portability.

Note: Each model is discussed in detail in Unit 2.

1.5 Benefits and Challenges of Cloud Computing

Benefits	Challenges
Cost saving — no upfront capital expenditure (CapEx → OpEx).	Security and privacy of data held by a third party.
Scalability and elasticity to handle peaks.	Vendor lock-in — hard to move between providers.
Fast deployment — servers in minutes.	Dependence on Internet connectivity and bandwidth.
High availability and disaster recovery.	Downtime / outages of the provider.
Automatic updates and maintenance.	Compliance and legal issues (data residency, GDPR).
Access from anywhere; supports collaboration.	Limited control and customisation; performance variability.
Environmentally efficient through resource sharing.	Hidden or unpredictable costs (data egress, over-provisioning).

1.6 Applications of Cloud Computing

- **Storage and backup** — Google Drive, Dropbox, OneDrive, Amazon S3.
- **Web and mobile app hosting** — start-ups host apps without owning servers.
- **Big-data analytics and AI/ML** — rent GPUs and clusters when needed.
- **Software development and testing** — quick test environments, CI/CD pipelines.
- **E-learning and collaboration** — Google Classroom, MS Teams, Zoom.

- **Healthcare, banking, e-commerce, entertainment** — patient records, mobile banking, Netflix streaming.
- **Disaster recovery and business continuity** — replicate data in another region.

1.7 Cloud Storage

Cloud storage lets data be saved on remote servers managed by a provider and accessed via the Internet. Data is replicated across multiple disks and often multiple locations for durability.

Type	Description	Typical use / example
Object storage	Data stored as objects (data + metadata + ID) in flat buckets; accessed by REST APIs.	Backups, media, data lakes — Amazon S3, Azure Blob
Block storage	Raw fixed-size blocks attached to a VM like a hard disk.	Databases, OS disks — Amazon EBS
File storage	Hierarchical shared file system (NFS/SMB).	Shared drives — Amazon EFS, Azure Files

- **Advantages:** pay for what you use, virtually unlimited capacity, durability through replication, easy sharing.
- **Concerns:** data privacy, latency, transfer costs and provider dependence.

1.8 Cloud Services Requirements

For a cloud service to be useful to customers, it must satisfy several requirements:

- **Scalability and elasticity** — handle growth without redesign.
- **Availability and reliability** — commonly expressed in an SLA (e.g., 99.9% uptime).
- **Performance** — acceptable response time and throughput.
- **Security and privacy** — authentication, encryption, isolation between tenants.
- **Interoperability and portability** — standard APIs to avoid lock-in.
- **Manageability and monitoring** — dashboards, logging, alerts.
- **Cost transparency** — clear metering and pricing.
- **Compliance** — meeting legal and regulatory rules.

Example 1.2 — SLA availability

A provider promises 99.9% monthly uptime. How much downtime is allowed in a 30-day month?

Solution

Total minutes = $30 \times 24 \times 60 = 43,200$.

Allowed downtime = $0.1\% \times 43,200 = 43.2$ minutes per month (about 43 min).

For 99.99% it would be only ≈ 4.3 minutes.

1.9 Cloud and Dynamic Infrastructure

A **dynamic infrastructure** is an IT environment that automatically adapts to changing business needs. Physical and virtual resources are pooled, and software policies decide how they are allocated, moved or retired — with little manual effort.

- **Virtualization** decouples workloads from hardware so they can move freely.
- **Automation and orchestration** provision, configure and scale resources automatically (e.g., auto-scaling groups).
- **Service management** monitors performance and enforces SLAs.
- **Resource pooling** maximises utilisation and reduces idle capacity.
- **Self-service portals** let users request resources instantly.

Example 1.3 — Auto-scaling

An e-commerce site normally needs 4 web servers. During a festival sale, CPU load exceeds 70%, so the auto-scaling policy launches 6 more servers; after the sale, extra servers are terminated and billing stops.

1.10 Cloud Adoption

Cloud adoption is the process by which an organisation moves its applications, data and processes to the cloud. A common approach:

Phase	Activities
1. Assess	Identify business goals, review applications, estimate costs (TCO) and risks.
2. Plan / Strategy	Choose service and deployment models, provider, security and compliance controls, migration approach.
3. Pilot	Migrate a low-risk workload to test performance and cost.
4. Migrate	Move workloads using rehost, replatform, refactor, repurchase, retire or retain (the “6 Rs”).

Phase	Activities
5. Operate & optimise	Monitor, secure, control cost and continuously improve.

- **Drivers:** cost reduction, agility, innovation, business continuity.
- **Barriers:** security concerns, skills shortage, legacy systems, regulatory limits, resistance to change.

Key Takeaways — Unit 1

- ✓ Cloud computing delivers on-demand, metered computing resources over the Internet.
- ✓ It evolved from mainframes → client-server → grid → utility computing → virtualised clouds.
- ✓ NIST's five characteristics: on-demand self-service, broad network access, resource pooling, rapid elasticity, measured service.
- ✓ Service models: IaaS, PaaS, SaaS; deployment models: public, private, community, hybrid.
- ✓ Main benefits: cost, scalability, agility; main challenges: security, lock-in, connectivity.
- ✓ Cloud storage types: object, block, file. Adoption follows assess → plan → pilot → migrate → optimise.

— End of Unit 1 —