Modern Cloud Computing

Lecture 1: Introduction to Cloud Computing

Chongzhou Fang
Rochester Institute of Technology

Today's Topics

- **Logistics**
- **Overview of Cloud Computing**
- **Service delivery models:** SaaS, PaaS, IaaS, DBaaS
- **Deployment models:** public, private, community, hybrid
- **Cloud Responsibilities and Challenges**

Logistics

Course Info

Item	Information
Instructor	Chongzhou Fang (cxfeec@rit.edu)
Teaching Assistant	Vedant Khairnar (vk6145@rit.edu)
Lecture	Tue Thu 5:00 PM – 6:15 PM
Office Hours	Tue Thu 4:00 PM – 5:00 PM (@ GLE-3435)
Classroom	GLE-1129
Website	MyCourses, Course Website
Course Materials	Cloud Computing: Theory and Practice, 2nd Edition
	External Documentations; Research Papers (Links provided)

Course Work

Homework

- HW1-HW4: Concept understanding + programming assignment
- HW5: Literature survey paper

Quiz

- In class quizzes

Semester Project

- Research project; or
- Sophiscated product
- Team Formation + Proposal (w/ Presentation) + Checkpoint Report + Final Report (w/ Presentation)

All reports: USENIX LaTeX template

Grading Policy

Component	Percentage
Homework	30% 6% each; lowest HW1-HW4 dropped
Quiz	10% 2% each; lowest dropped
Project Proposal	10%
Project Checkpoint Report	10%
Project Presentation: Proposal and Preliminary Results	5%
Paper Presentation	5%
Final Project Presentation	10%
Final Project Report	20%
Bonus: Class Attendance	5%

Letter Grade	Total Points
A	> 92
A-	90–91
B+	88–89
B	82–87
B-	80–81
C+	78–79
C	72–77
C-	70–71
D	65–69
F	< 64

Topics Covered

Topic	Week
Concurrency & Parallelism	1
Cloud Network & Storage	2
Cloud HW&SW	3
Delivery Model	3-4
Resource Management	4-5
Virtualization	5-6
Cloud Security	6
Cloud Application Development	7
Advanced Topic: Heterogeneous Computing, Modern Resource Scheduling, ML in Cloud, Cloud and IoT, Cloud Security, Future Cloud Infrastructure, Cloud in the Agentic-AI Era	9-14

Lecture Expectation

- Read reference materials before class
    - Lots of information to digest, but don't worry, simply having an idea of what will be covered is enough.
- Actively paritipation in discussion
- Presentation: Present materials in a clear & logical way

Introduction to Cloud Computing

Overview of Cloud Computing

Computing as a Utility

Cloud computing extends the idea that computing can be consumed like a utility:

- Electricity, water, and gas are provided by utility infrastructures.
- Computing resources can also be provided by large-scale infrastructures.
- Users consume computing, storage, and networking resources on demand.
- Users pay for what they consume rather than owning all infrastructure.

Historical Vision

John McCarthy predicted that

> ...computing could someday become a public utility, just as the telephone system in a public utility...

That vision is now realized through cloud computing:

- Large data centers host hardware and software resources.
- Users access these resources over networks.
- Computing becomes an industry based on service delivery.

The era started in 2006 when Amazon offered Elastic Cloud Computing (EC2) and Simple Storage Service (S3).

Why Cloud Computing Matters

Cloud computing is a disruptive computing paradigm because it changes many areas of computing:

- Data storage
- Computer architecture
- Networking
- Resource management
- Scheduling
- Security
- Software deployment
- Application design

The Internet Made the Cloud Possible

Cloud computing depends on fast, reliable, and low-cost communication.

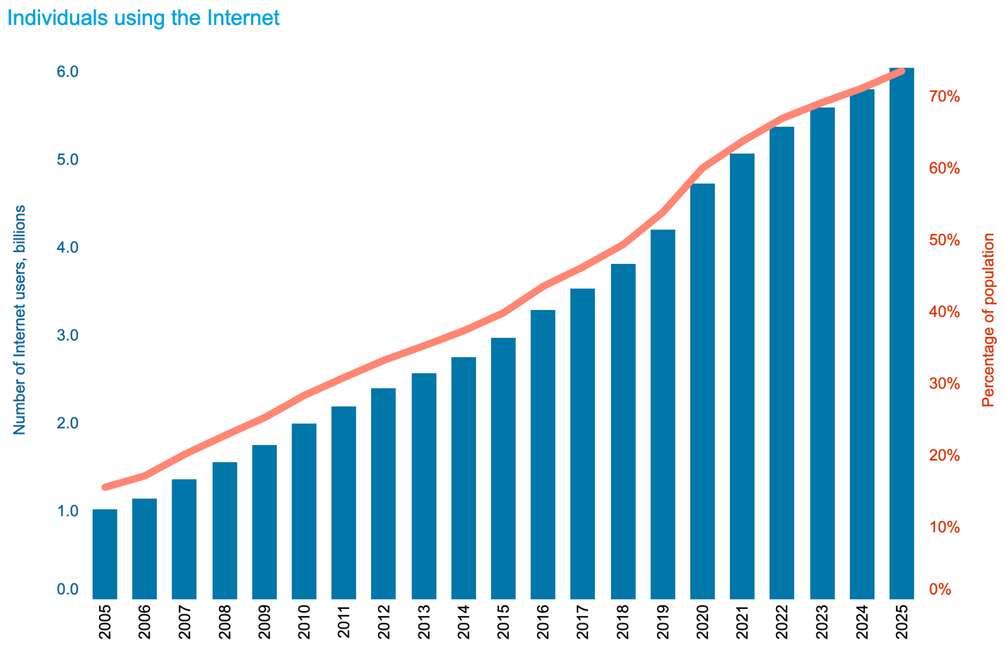
Without the Internet, users could not realistically access:

- Remote storage
- Remote computation
- Distributed services
- Large-scale applications hosted in distant data centers

Growth of Internet Users

The rapid growth of Internet users created the environment for cloud adoption.

- First billion Internet users: around 2005
- Second billion: around 2010
- Third billion: around 2014
- 2025: 6 Billion
  - Continued global growth increased demand for scalable online services



(Source: https://www.itu.int/itu-d/reports/statistics/2025/10/15/ff25-internet-use/)

NIST Definition of Cloud Computing

NIST defines cloud computing as a model that enables:

> Ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Motivation

- Data processing and storage can be done more efficiently on large farms of computing and storage systems accessible via Internet.

Five Essential Characteristics

According to the NIST-style definition, cloud computing is characterized by:

1. On-demand self-service
2. Broad network access
3. Resource pooling
4. Rapid elasticity
5. Measured service

On-Demand Self-Service

Users can provision resources without direct human interaction with the provider.

Examples:

- Launch a virtual machine
- Create a database instance
- Allocate object storage
- Deploy a web application

Broad Network Access

Cloud services are accessed over networks using standard mechanisms.

This enables access from:

- Laptops and desktops
- Mobile phones
- Tablets
- Embedded and IoT devices
- Enterprise systems

Location-independent from the user's perspective.

Resource Pooling

Cloud providers pool physical and virtual resources to serve many users.

Resources include:

- CPU cycles
- Memory
- Storage
- Network bandwidth
- Databases

This creates **multi-tenancy**, where multiple customers share provider infrastructure.

Rapid Elasticity

Cloud resources can scale up or down quickly according to workload demand (e.g. in FaaS).

Examples:

- Add more web servers during traffic spikes
- Remove resources during low demand
- Automatically scale based on CPU utilization

Measured Service

Cloud usage can be metered.

Users are charged based on consumption, such as:

- VM running time
- Storage capacity
- Network transfer
- Requests per second
- Database operations

Business model: pay-as-you-go.

Cloud Computing as a Business Model

Cloud computing is also a business model for delivering computing services:

- Provider owns and operates infrastructure.
- Customers rent resources or services.
- Costs move from capital expenditure to operational expenditure.
- Providers benefit from economy of scale.

Cost-Effective: How

Cloud computing reduces cost through:

- Resource multiplexing
- Higher average utilization
- Commodity hardware
- Centralized management
- Specialized operations teams
- Economy of scale

Network-Centric Computing

Cloud computing shifts computing from local machines to network-accessed resources.

In network-centric computing:

- Data processing happens on remote systems.
- Data storage happens in remote data centers.
- Users interact through networks and thin clients.
- Applications often compose multiple services.

Network-Centric Content

Network-centric content means content is stored, accessed, and processed through the network.

Examples:

- Photos and videos stored in cloud services
- Music and movies streamed on demand
- Shared documents edited collaboratively
- Application data synchronized across devices

Content becomes accessible from any connected device.

Characteristics of Network-Centric Applications

Many cloud applications are both:

- **Data-intensive:** analytics, simulation, multimedia, sensor data
- **Network-intensive:** data streaming, collaboration, distributed computation

They often require:

- Large-scale storage
- High-bandwidth networks
- Low-latency communication
- Workflow management
- Shared data access

Benefits of Network-Centric Computing

Key benefits include:

- Easier data sharing
- Elastic capacity for peak workloads
- Lower upfront investment
- Reduced local infrastructure maintenance
- Access through thin clients and mobile devices

Concerns of Network-Centric Computing

Moving computation and storage to the cloud introduces concerns:

- Security and privacy risks
- System state concern
- Quality-of-service (QoS) challenges

Important Factors to Make Cloud Computing Practical

Cloud computing became practical because several trends converged:

- Faster processors and multicore systems
- Cheaper memory and storage
- Improved networking
- Virtualization
- Large-scale data center engineering
- Better distributed software systems
- Service-oriented software architectures

Hardware Trends

Modern clouds rely on commodity hardware at massive scale.

Important hardware trends:

- Multicore processors
- High-density servers
- Large memory capacity
- Fast storage devices
- High-speed data center networks

Clouds use many low-cost systems rather than a few one-of-a-kind supercomputers.

Software Architecture Trends

Cloud applications often use multi-tier architecture.

Common three-tier model:

1. **Presentation tier:** user interface, browser, mobile app
2. **Application/logic tier:** business logic and service logic
3. **Data tier:** databases and persistent storage

This separation improves scalability and maintainability.

Economy of Scale

Large data centers can reduce costs compared with medium-size centers.

Major cost categories:

- Networking
- Storage
- Administration
- Energy
- Cooling
- Facility operations

Cloud providers spread these costs across many customers.

Energy and Data Centers

Data centers consume significant electricity.

Energy costs affect:

- Where providers build data centers
- How providers cool facilities
- How workloads are scheduled
- How providers optimize infrastructure

Reasons of Commercial Success

- Mostly homogeneous infrastructure within one administrative domain
- Strong financial incentives from large IT companies
- Enterprise demand for scalable computing
- Elastic resource model
- Pay-as-you-go pricing
- Mature networking and virtualization technologies

Major Obstacles

Cloud computing still faces important obstacles:

- Availability of Service: What if service provider cannot deliver
    - Solution: Service Level Agreement (SLA)
- Vendor lock-in: customer hooked to one cloud service provider
- Data confidentiality and accountability
- Data transfer bottlenecks
    - > Don't underestimate the bandwidth of a semi-truck loaded with disks
- Performance unpredictability
- Elasticity

Data Transfer Bottleneck Example

Transferring 1 TB of data can be surprisingly slow.

Approximate transfer time:

- At 1 Mbps: about 10 days
- At 1 Gbps: slightly more than 2 hours

For data-intensive applications, network transfer can dominate total cost and runtime.

Major Cloud Providers

The cloud ecosystem includes major providers such as:

- Amazon Web Services
- Google Cloud
- Microsoft Azure
- IBM cloud offerings
- Oracle cloud offerings

Amazon was an early pioneer in IaaS with EC2 and S3.

AWS as an Example

AWS introduced key cloud services in 2006:

- **EC2:** Elastic Compute Cloud for virtual servers
- **S3:** Simple Storage Service for object storage

AWS demonstrated that cloud infrastructure could support enterprise computing and computing for the masses.

EC2: Elastic Compute Cloud

EC2 provides virtual server instances.

Users choose:

- Region
- Availability zone
- Instance type
- Operating system image
- Storage and networking options

An instance behaves like a virtual private server.

S3: Simple Storage Service

S3 is an object storage service.

It supports basic object operations:

- Write
- Read
- Delete

Objects are stored in buckets and retrieved with keys.

S3 is designed for large-scale storage rather than local file-system semantics.

Cloud Delivery Models

Cloud Delivery Models

Cloud services are commonly grouped into delivery models:

- **IaaS:** Infrastructure as a Service
- **PaaS:** Platform as a Service
- **SaaS:** Software as a Service
- **DBaaS:** Database as a Service
- **FaaS:** Function as a Service

IaaS: Infrastructure as a Service

In IaaS, users provision fundamental computing resources.

Users can control:

- Virtual machines
- Operating systems
- Storage volumes
- Deployed applications
- Some networking components

Provider controls the underlying physical infrastructure.

When IaaS Is Useful

IaaS is useful when:

- A new business needs infrastructure quickly
- Users need control over operating systems and software stacks
- Workloads require flexible VM and storage allocation

PaaS: Platform as a Service

In PaaS, users deploy applications using provider-supported tools and platforms.

Users control:

- Their deployed applications
- Some application configuration

Users do **not** manage:

- Physical servers
- Storage infrastructure
- Operating system details
- Underlying network

When PaaS Is Useful

PaaS is useful for:

- Software development
- Automated deployment
- Collaboration among developers
- Managed runtime environments

SaaS: Software as a Service

In SaaS, users access provider-hosted applications.

Users typically do **not** control:

- Servers
- Operating systems
- Storage infrastructure
- Application internals

Examples:

- Web email
- Collaboration tools
- Customer relationship management
- Web-based office applications

When SaaS Is Useful

SaaS is well suited when:

- Many users need the same application
- Demand has periodic peaks
- Web or mobile access is important
- The need is short-term
- Users do not want to manage infrastructure

SaaS is less suitable when data cannot be externally hosted or strict real-time control is required.

FaaS: Function as a Service

Function as a Service (FaaS) is a cloud model where developers deploy small functions instead of managing servers.

The cloud provider handles:

- Server provisioning
- Scaling
- Load balancing
- Runtime management
- Fault tolerance

The user focuses on writing code that responds to events.

Key idea: You upload a function; the cloud runs it when triggered.

When FaaS Is Useful

FaaS is especially useful for **event-driven** and **bursty** workloads.

Less suitable for:

- Long-running applications
- Workloads requiring persistent local state
- Applications needing fine control over the runtime or infrastructure

Key idea: FaaS is best when computation is small, stateless, and triggered by events.

Comparing IaaS, PaaS, SaaS, and FaaS

Category	IaaS	PaaS	SaaS	FaaS
Definition	Infrastructure as a Service	Platform as a Service	Software as a Service	Functions as a Service
User Control	Full control over infrastructure	Control over applications and data; not infrastructure	Minimal control; mostly configuration	No control over infrastructure or runtime
Vendor Responsibility	Hardware, networking	Infrastructure, runtime, and middleware	Everything: infrastructure, app, updates, and security	Everything except user-defined functions
Customization	Very high; full flexibility	High; app-level customization supported	Low; limited to available options	Low; limited to function logic
Scalability	User-managed scalability	Built-in auto scaling	Vendor-managed	Event-driven auto scaling
Deployment Speed	Slow; complex setup	Fast; prebuilt services	Instant; out-of-the-box	Very fast for isolated tasks
Typical Use Case	Full custom ecommerce platform	Modular ecommerce with custom logic	SMB-focused ecommerce, such as Shopify	Triggered actions, such as resizing images
Example Providers	AWS EC2, Google Compute Engine	Microsoft Azure App Services, Virto Commerce	Shopify, Salesforce Commerce, Gmail	AWS Lambda, Azure Functions
eCommerce Fit	Custom enterprise ecommerce stacks	Scalable B2B ecommerce with complex workflows	Fast go-to-market with standard needs	Microservices inside composable platforms
Pricing Model	Pay-as-you-go hardware usage	Pay-as-you-go plus development tools	Subscription-based	Pay-per-execution

(Source: IaaS, PaaS, SaaS, and FaaS: Cloud Models Explained)

Cloud Deployment Models

Cloud Deployment Models

Clouds can be deployed in different organizational forms:

- **Public cloud**
- **Private cloud**
- **Community cloud**
- **Hybrid cloud**

The right model depends on cost, security, compliance, control, and scalability requirements.

Public Cloud

A public cloud is available to the general public or a large industry group.

Characteristics:

- Owned by a provider selling cloud services
- Multi-tenant infrastructure
- Large scale
- Pay-as-you-go pricing
- Rapid provisioning

Examples include large commercial cloud providers.

Private Cloud

A private cloud is operated for a single organization.

It may be:

- Managed by the organization or a third party
- Located on-premises or off-premises
- Designed for specific security, compliance, or performance needs

Private clouds offer more control but may reduce some utility-computing benefits.

Community Cloud

A community cloud is shared by organizations with common concerns.

Examples of shared concerns:

- Mission
- Security requirements
- Policy requirements
- Compliance requirements

This model sits between private and public cloud in terms of sharing and control.

Hybrid Cloud

A hybrid cloud combines two or more cloud infrastructures.

The clouds remain distinct but are connected through technology enabling:

- Data portability
- Application portability
- Load balancing
- Cloud bursting during demand spikes

Hybrid cloud is common when organizations combine private infrastructure with public cloud capacity.

Cloud Responsibilities and Challenges

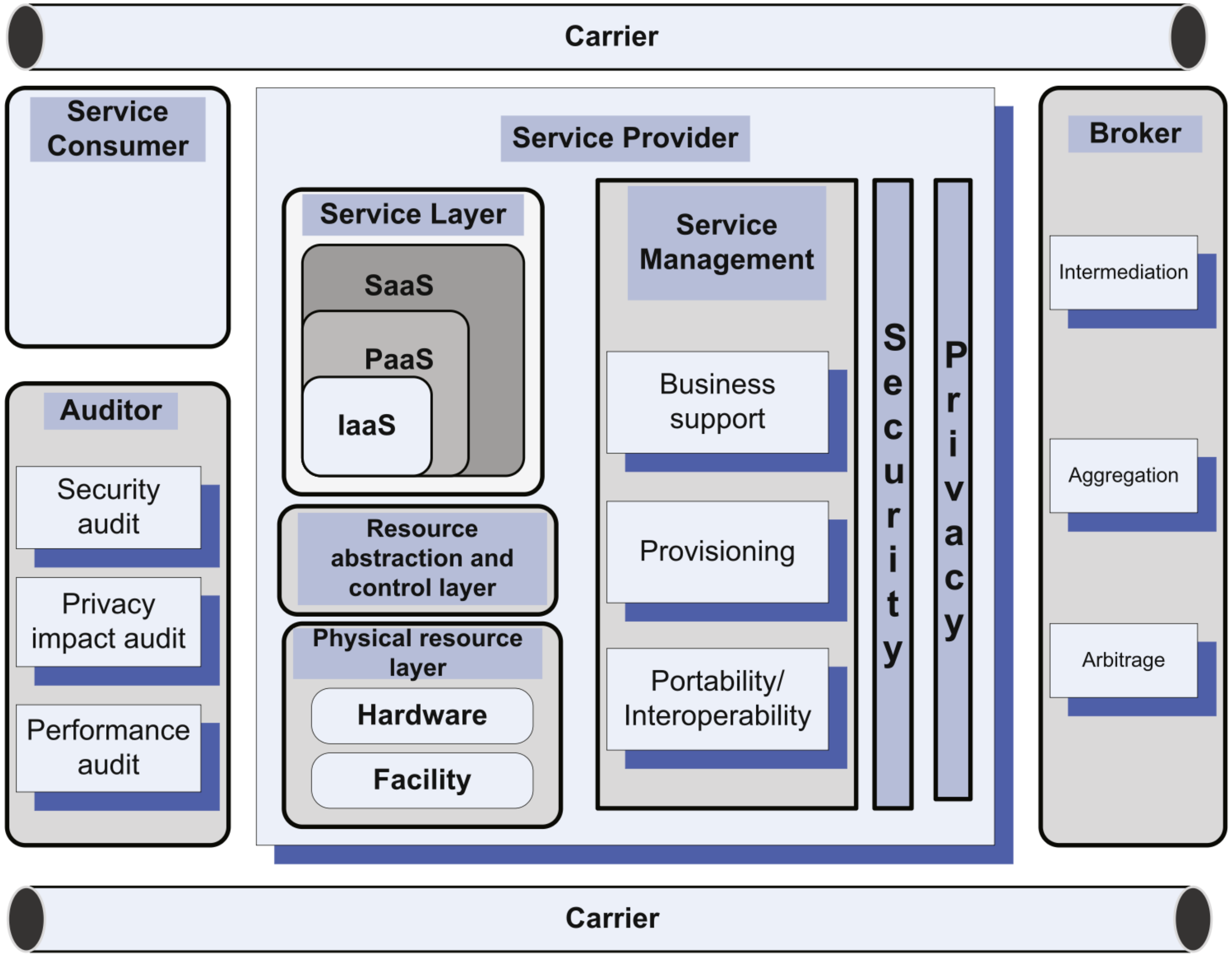
Cloud Ecosystem

The cloud ecosystem includes more than providers and users.

Roles include:

- Service consumer
- Service provider
- Carrier
- Broker
- Auditor

(Source: Cloud Computing: Theory and Practice, 2nd Edition, Fig. 2.3)



Cloud Service Provider Responsibilities

Cloud providers support many operational functions:

- Service provisioning
- Virtualization
- Monitoring
- Billing and accounting
- Backup
- SLA management
- Security management
- Technical support

Cloud platforms are complex service-oriented systems.

Service-Level Agreements (SLAs)

A **Service-Level Agreement (SLA)** records a common understanding between a cloud user and a cloud provider.

An SLA typically defines:

- Services to be delivered
- Priorities
- Responsibilities
- Guarantees
- Warranties
- Security requirements
- Handling of confidential information
- Problem management
- Termination conditions

In cloud computing, SLAs often specify a **target level of service** or a **minimum level of service**.

SLA Example: Call Center

Common metrics are service-specific

For this application, SLA includes

1. *The abandonment rate*: percentage of calls abandoned while waiting to be answered.
2. *The average speed to answer*: average time before the service desk answers a call.
3. *The time service factor*: percentage of calls answered within a definite time frame.
4. *The first-call resolution*: percentage of incoming calls that can be resolved without a callback.
5. *The turnaround time*: time to complete a certain task.

SLA Management

Two components

1. Negotiation
   - Object of negotiation
   - Negotiation protocol
   - Decision models: Proposals, counter proposals
2. Monitoring of Fulfillment in real time

Ethical Issues in Cloud Computing

Cloud computing raises ethical issues because:

1. Control is relinquished to third-party services.
2. Data may be stored across multiple sites and organizations.
3. Multiple services interoperate across networks.

Privacy and Accountability

Cloud providers may store large amounts of sensitive personal and enterprise data.

Important questions:

- Who can access the data?
- Where is the data stored?
- Who is accountable when something goes wrong?
- What logs are recorded?
- How do we balance accountability with privacy?

Cloud Vulnerabilities

Cloud systems can be affected by:

- Malicious attacks
- Power failures
- Software bugs
- Control-plane failures
- Load balancer failures
- Network outages
- Interactions among independent services

Large-scale failures can affect many users simultaneously.

Summary

Cloud computing is a model for delivering computing as a service through large-scale networked infrastructure.

Key points:

- Cloud computing is both a technology model and a business model.
- It depends on Internet-scale networking, data centers, virtualization, and service automation.
- Essential characteristics include self-service, broad access, pooling, elasticity, and measured service.
- Delivery models differ in how much control users retain.
- Cloud adoption creates major opportunities and major security, ethical, and reliability challenges.

Questions?