DSC204A: Scaleable Data Systems Spring 2025

15: Cloud Computing

Instructor: Umesh Bellur *Scribe: Ananay Gupta, Kedar Desai, Keisha Mehta*

This class primarily focused on the concept and real-world applications of cloud computing. We began by understanding parallelism, which refers to performing multiple tasks simultaneously. Parallelism can occur on a single machine with multiple cores (CPU), on specialized hardware like GPUs, or across distributed systems in a network. Cloud computing supports all these forms of parallelism.

The core idea behind cloud computing is that computing resources—like processing power, storage, and applications—can be accessed remotely over the Internet, similar to how we use electricity or water as utilities. This model allows organizations and individuals to use powerful resources without the overhead of managing physical infrastructure.

What is the Cloud?

Cloud computing allows us to deliver and consume computing services on demand. It includes infrastructure (e.g., a database engine like Snowflake), platforms (e.g., Google App Engine), and software (e.g., Gmail). This utility-like model helps scale up or down depending on demand.

Key Characteristics of Cloud Computing

- Always available, like utilities—electricity, water, internet.
- Scalable and elastic—resources can be increased or reduced instantly.
- Pay-as-you-go—only pay for what you use.
- Reduces the need for physical hardware and in-house IT support.

Cloud Computing Delivery Models

- **Software as a Service (SaaS)**: Ready-to-use apps over the internet.
  - SaaS allows users to access applications over the internet without worrying about installation, maintenance, or hardware requirements. These applications run on cloud servers, and users typically interact with them through a web browser. The provider handles everything behind the scenes—updates, backups, infrastructure, and scalability.
  - Ideal for end users who want direct functionality.
  - Example: Gmail, Google Docs, Zoom, and Salesforce are widely-used SaaS applications.
  - Use case: A small business can use Google Workspace for email, documents, and meetings without needing in-house IT support.
- **Platform as a Service (PaaS)**: Platforms for developers to deploy apps.
  - PaaS offers a ready-made development environment in the cloud, where developers can build, test, and deploy applications without setting up hardware or software manually. It includes tools, libraries, and services like databases and runtime environments.
  - Ideal for developers.
  - Examples: Heroku, Google App Engine, Azure App Services.

  - Use case: A startup developing a mobile app can use PaaS to quickly deploy it with integrated support for analytics and storage.

- **Infrastructure as a Service (IaaS)**: Virtualized computing infrastructure.
  - IaaS gives users access to basic computing infrastructure—virtual machines, storage, and networking—over the internet. Users have full control over the operating system and applications but don't need to manage the physical hardware.
  - Ideal for IT administrators needing control.
  - Examples: Amazon EC2, Azure VMs, Google Compute Engine.
  - Use case: A company can rent powerful servers for data analysis for a few weeks, instead of buying expensive machines.

Cloud Deployment Models

- **Public Cloud:** Services are offered over the internet by third-party providers like Amazon Web Services (AWS), Microsoft Azure, or Google Cloud Platform (GCP). These providers manage all the infrastructure, and users simply access resources on demand. It's ideal for startups, students, or businesses that want quick access to scalable computing without owning hardware. Example: Hosting a website on AWS using free tier credits.
- **Private Cloud:** Built and maintained by a single organization for its exclusive use. It provides more control over security, compliance, and infrastructure. Though more expensive than public cloud, it suits industries like finance or healthcare where sensitive data needs strict protection. Example: A bank like Bank of America may operate a private cloud to ensure its customer data stays internal and compliant with regulations.
- **Hybrid Cloud:** Combines public and private cloud environments to get the best of both worlds. Organizations can keep critical or sensitive data on their private cloud while using public cloud services for less-sensitive operations or handling peak traffic. Example: A hospital may store patient records in a private cloud but use the public cloud to run data analytics on anonymous health trends.
- **Community Cloud:** Shared infrastructure between multiple organizations with similar goals, requirements, or industries. It allows for collaborative efforts while maintaining some level of privacy and customization. Example: Aerospace firms or research institutions might pool computing resources to run simulations or share specialized software tools.

What Does the Cloud Offer?

- **Applications:** These are ready-to-use software programs accessible through the internet. Tools like Zoom, Google Docs, and Microsoft Office 365 are hosted entirely in the cloud, so users don't have to install or maintain them. They can be used from any device with a browser and internet connection, making collaboration and remote access easy and seamless.
- **Services:** Cloud platforms offer various on-demand services such as machine learning toolkits (like AWS SageMaker or Google AutoML), speech-to-text engines, and translation APIs. Developers and businesses can integrate these services directly into their apps without building them from scratch, saving time and resources.
- **Networks and Security:** The cloud includes robust networking capabilities—such as virtual private networks (VPNs), firewalls, and secure access controls—that connect resources while protecting them from unauthorized access. Providers invest heavily in security infrastructure and compliance standards, so even small businesses benefit from enterprise-grade protection.

- **Infrastructure:** This includes the foundational components like storage, computing power, and virtualization. Cloud infrastructure is built on distributed systems that automatically manage resources, balance loads, and recover from failures. Features like autoscaling and backup recovery are handled without user intervention.
- **Cloud Data Centers:** Large facilities with thousands of servers connected by fast networks. Their scale enables cost efficiency, around-the-clock operation, and seamless access to computing resources for millions of users worldwide.

Historical Models

- **Grid Computing:** A model that connects multiple, often geographically distant computers to work on a common task. Each computer contributes its unused resources—like processing power or storage—toward solving a large-scale problem. A famous example is *SETI@home*, which invited users around the world to donate idle computing power to analyze radio signals from space. Grid computing was especially useful in scientific research where massive datasets had to be processed, such as climate modeling or cancer simulations.
- **Utility Computing:** This approach treats computing resources like utilities such as electricity or water—you pay based on usage. Instead of buying and maintaining dedicated servers, businesses could rent infrastructure from a provider and be billed for the exact resources they consumed. It introduced the idea of eliminating large upfront capital expenses in favor of more flexible, operational costs. This model laid the foundation for today's cloud pricing strategies and on-demand resource allocation.

Shared Resources and Metering

- **Users share servers via VMs or containers:** In cloud environments, a single physical server can host multiple users at once by isolating their workloads using virtual machines (VMs) or containers. This allows for efficient use of hardware and keeps costs low. Each user's environment runs independently, so even though resources are shared, operations remain secure and separate.
- **Metering tracks usage:** Cloud platforms monitor exactly how much compute power, memory, storage, and network bandwidth each user consumes. This usage is tracked in real time and used for billing—similar to how a utility company measures electricity usage. It ensures transparency and fairness, as users only pay for what they actually use.
- **Providers manage patching, security, and updates:** Cloud providers take care of regular maintenance tasks like software updates, security patches, and performance monitoring. This reduces the burden on individual users or businesses, allowing them to focus on their core applications without worrying about underlying system upkeep.

Advantages of Cloud Computing

- **No upfront hardware cost:** You don't need to buy expensive laptops or servers—just rent what you need.
- **Flexible scaling:** Easily scale your usage up or down based on demand—great for startups or sudden spikes in traffic.
- **Access from anywhere:** Use your services from any internet-connected device, whether you're at home, work, or traveling.
- **Pay-as-you-go:** You only pay for what you use—no need for permanent investments in unused resources.

- **Saves on licensing:** Software like Salesforce or Adobe can be rented monthly, avoiding huge one-time costs.
- **Environmentally efficient:** Providers consolidate workloads to fewer machines, which saves energy and costs.
- **Virtualization via VMMs:** Multiple operating systems run efficiently on shared physical machines using Virtual Machine Monitors, often coded in C for speed.

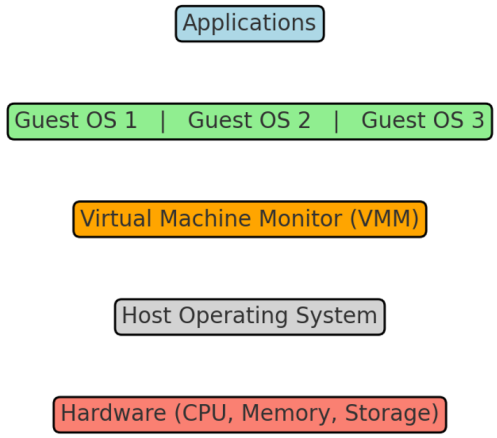

Figure: Virtual Machine Monitor (VMM) allowing multiple OSs on shared hardware

Challenges and Limitations

- **Long-term cost:** Using the cloud continuously can add up—like taking a taxi every day instead of owning a car.
- **Downtime risk:** If the cloud provider has issues, your service could go down—affecting users and business.
- **Confidentiality concerns:** Since data resides on external servers, there's a need to trust the provider with sensitive information.
- **Vendor lock-in:** Moving your services and data from one provider to another can be complex and costly.
- **Latency issues:** Cloud services may experience delays, which is problematic for real-time systems like autonomous vehicles.
- **Regulatory barriers:** Laws like GDPR restrict where data can be stored, especially for global companies.
- **Lack of transparency:** Users don't know where or how their jobs are being executed—this can affect performance or raise concerns.