DSC516: Cloud Computing

Marios D. Dikaiakos

http://www.cs.ucy.ac.cy/mdd

University of Cyprus
Department of Computer Science

Objective

- Provide an introduction to and understanding of advanced **concepts** in the field of Cloud Computing
- Enable students to **design**, **develop**, **deploy**, **monitor** and **analyze** applications on state-of-the-art Cloud computing platforms.
- Covers **key elements** and **technologies** of Cloud Computing Infrastructures, Services, and Applications.
- Students who attend the course will gain an understanding of the Cloud Computing paradigm and the technical underpinnings of Cloud services; they will be able to describe and analyze key middleware components of Cloud services, to understand the main Cloud application development paradigms, and to use state-of-the-art Cloud service offerings for Data Science-related projects.

DSC516: Cloud Computing

Part I: Basic Concepts and Models

Module1: Distributed Computing Concepts and Models

L1: From Mainframes to the Cloud

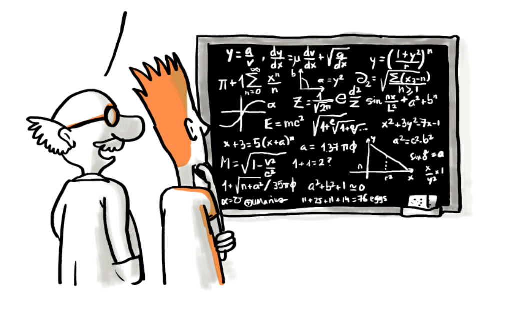
Historical Overview

- Centralization and Mainframes
- Personal Computing
- Client-Server Computing
- Web Computing
- Grid Computing
- Utility Computing
- The emergence of Cloud Computing

Learning Objectives

- Understand and explain the evolution of ICT that lead to the introduction of Cloud Computing.
- Understand and explain the role of Moore's Law, Location, Exponential Phenomena, Network Effects, GPT, Value Proposition.
- Understand, explain and apply the concepts of Web Computing, eScience, SaaS, Utilities and Utility Computing, Possession of Computation and its implications, Ownership of Data, the Value Proposition of the Cloud, Ex post regulation and Ex ante agreements

Topics Discussed



Discussed the key technological and economic developments that led to Cloud Computing Infrastructures.

Explored and explained the concepts and role of:

- Moore’s Law, Mainframes, PCs and Client-Server
- Cluster Computing, Web Computing, Internet-scale Services, Exponential Phenomena, Network Effects
- Portability and preservation of data, Data Ownership
- Possession of Computation and its implications
- General Purpose Technologies and Public Utilities
- Grid Computing, Utility Computing, Software-as-a-Service

Knowledge Check

- Can the Grid be considered as a GPT? Explain.
- Can the Cloud be considered to be a GPT? Explain.
- Explain what Network Effects are and give an example of a Cloud service that has benefited from Network Effects.
- Describe the core value proposition of the Cloud.
- Why is location important in modern computing? Name three key reasons.

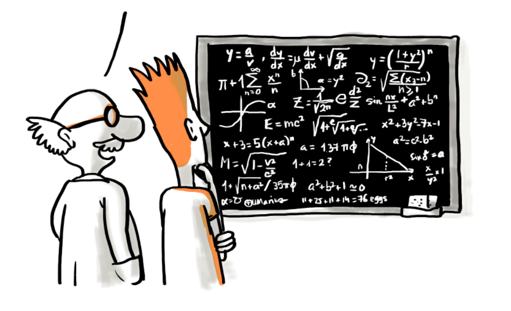
L2: Distributed Computing: Concepts, Models, Middleware

- Key concepts and Abstractions
- Architecture Models
- Middleware
- End-to-end Arguments in System Design

Learning Objectives

- Review and explain key concepts: Architecture, System Architecture, Resource, Middleware, Client, Server, COD, REV, Middleware, End-to-End Arguments in Systems Design.
- Review, explain and apply distributed computing models, client-server computing, etc.
- Understand and explain functional decomposition concerns in the design of distributed systems.

Topics Discussed



- Abstraction
- Architecture
- System Architecture
- Resource, Physical and Logical
- Process and State
- Distributed Computing Models: Client-Server, REV, COD, MA
- Middleware
- Middleware services and categories

- Middleware services and categories

- End-to-end arguments in system design and functional decomposition, performance tradeoffs, applying e-2-e arguments in various application scenarios

Knowledge Check

- What information constitutes the state of a resource component?
- What is an execution environment? Give an example.
- What is a socket?
- Explain the difference between Remote Evaluation and Client Server models. Explain one problem that REV has for the security of systems.
- Describe and explain a scenario where a Mobile Agent model may be preferable over Client Server.
- Explain the differences between synchronous and asynchronous communication.
- Give a definition of middleware and 3 examples of middleware services.

Knowledge Check

- Why is location important in modern computing? Name three key reasons.
- Describe the difference between and API and an SDK.
- Describe the key abstractions comprising the client-server, remote evaluation, code on demand and mobile agent distributed computing models.
- What is the difference between the traditional client-server model found in early systems, like X Windows, and the client-server model that emerged from the World-Wide Web development?
- Give the definition of a Web service.
- Analyze the engineering tradeoffs in implementing secure data transmission in packet-switched networks.
- Is e-mail spam detection a functionality that needs to be dealt in the core or at the end of the network?

DSC516: Cloud Computing

Part I: Basic Concepts and Models

Module 2: Cloud Computing Definitions and Models

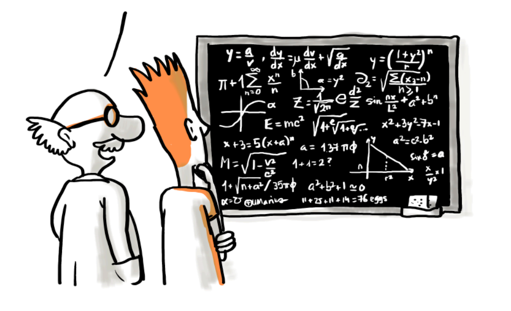
L3: Cloud Computing: Introduction, Definitions, Taxonomy

- Key Features of the Cloud
- Elasticity & the Economics of Cloud Provision
- The benefits of moving to the Cloud
- Profile of a Cloud Provider
- Cloud Data Centers
- Data Center Location
- Cloud Service Models
- Cloud Deployment Models
- Concerns in Adopting Cloud Computing
- Cloud Computing Application Models

Learning Objectives

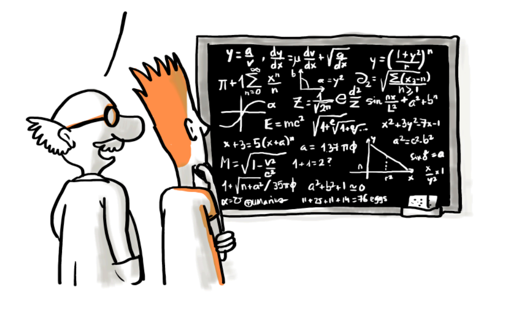
- Understand, explain and apply Cloud Computing Definitions
- Understand and explain the key factors of the Cloud evolution.
- Identify, Explain and Compare the concepts of Cloud, IAAS, PAAS, SAAS, Elasticity, Economics.

Topics discussed



- Definition and essential characteristics of Cloud Computing
- The concept of elasticity and types thereof
- Economics of Cloud computing: over provisioning, under-provisioning, financial cost thereof, cloud or cluster - cost-benefit analysis, summary of AWS pricing models
- Moving a business to the Cloud: factors to consider
- Cloud provider profile: key requirements and characteristics
- Cloud service models: IaaS, PaaS, SaaS; pros and cons.
- Cloud deployment models

Topics discussed



- Obstacles and opportunities to Cloud adoption
- Cloud computing models: IaaS, PaaS, SaaS
- On-demand service provision, Utility computing
- Pay per-use, Elasticity, CapEx, OpEx
- Warehouse Scale Computers
- Economies of Scale
- Cost Associativity
- Shipping photons or electrons?
- Compute locally or ship data remotely?
- Mobile Apps and Analytics
- Elasticity: What and Why?
- Resource Abstraction

Knowledge Check

- What is Cloud Computing, and how is it different from previous paradigm shifts such as Software as a Service (SaaS)?
- Why Cloud Computing took off in 2009, whereas previous attempts have foundered?
- What does it take to become a Cloud Computing provider, and why would a company consider becoming one?
- What new opportunities are either enabled by or are potential drivers of Cloud Computing?
- How might we classify current Cloud Computing offerings across a spectrum, and how do the technical and business challenges differ depending on where in the spectrum a particular offering lies?
- What, if any, are the new economic models enabled by Cloud Computing, and how can a service operator decide whether to move to the cloud or stay in a private datacenter?
- What are the top 10 obstacles to the success of Cloud Computing—and the corresponding top 10 opportunities available for overcoming the obstacles?
- What changes should be made to the design of future applications software, infrastructure software, and hardware to match the needs and opportunities of Cloud Computing?
- Analyze 4 key factors that determine where to establish a datacenter for a large cloud provider.
- Is horizontal or vertical elasticity available over AWS EC IaaS service? Explain.
- Give the definition vertical and horizontal elasticity, provide representative examples and discuss the tradeoffs between them.

University of Cyprus
Department of Computer Science

Knowledge Check

- Explain what is horizontal and vertical elasticity, how they differ and which one is more difficult to achieve.
- Explain what overprovisioning means and discuss if it is a problem and why.
- Suppose a biology lab creates 1 TB of new data for every wet lab experiment. A computer the speed of one EC2 instance takes 1 hour per GB to process the new data. The lab has the equivalent 5 instances locally. Explore the tradeoffs between computing the experiments in house, on Amazon, on Google Cloud, on Azure or on a local (Cypriot) Cloud provider.
- Suppose a biology lab has a dataset with 1000 files with 1 TB of data in each file. To process each file, they need to run one EC2- server for 1 hour. Explain what cost-associativity is and how it should be taken into account when deciding how to process these files in the shortest possible time while keeping expenses under control.
- Give a diagram of the full hardware/software stack for a typical cloud application and show which parts of the stack are controlled by the cloud service customer in the case of a SaaS, an IaaS and a PasS. Explain.

Knowledge Check

- Explain what is customer lock-in and why is it a problem for cloud users. What design steps can be taken to mitigate this problem in cloud offerings and/or cloud application development and at what cost?
- Does the use of the cloud instead of private data center simplifies or makes more difficult the management of external security threats for a company? Discuss the pros and cons.
- What is the main mechanism offering internal security in modern cloud offerings and why?

Knowledge Check

- Name and describe AWS solutions for providing external security to its customers.
- Are there any measures that Amazon provides to protect its customers from itself? Name and discuss.

L4: Modern Computing Landscape

- From Embedded Devices to the Internet of Things (IOT)
  - Cyber-physical Systems and IoT
  - A Tiered Model of Internet Computing
  - Sensing on the IoT: Data-driven Applications
  - Case Study: City Buses Application
  - Energy Concerns in the Modern Computing Landscape
  - Case Study: ENEDI Platform

Learning Objectives

- Understand and explain the current computing landscape.
- Understand and explain concepts, key constraints and key challenges in Edge and Fog computing.
- Memorize and use values of key properties of distributed systems' components and their evolution: CPU speed, network latencies, power consumption.

Learning objectives

- Understand and explain the basic characteristics of **Warehouse Scale Computers**.
- Understand and explain the differences between **WSCs** and **Datacenters**.
- Understand and explain the **software infrastructure building blocks** of WSC.
- Understand and describe main characteristics of **WSC buildings** and their **power provision**.
- Understand and explain the **basic power, cooling components** of a modern datacenter.
- Understand and explain the concept of **energy efficiency** in datacenters, and associated mechanisms.
- Understand, and explain the cost structure of running a modern datacenter, and the concept of TCO (**total cost of ownership**).
- Explore, understand and explain concepts and techniques for energy-efficiency in datacenters.

DSC516: Cloud Computing

Part II: Cloud Building Blocks

Module 3: Cloud Computing Infrastructure

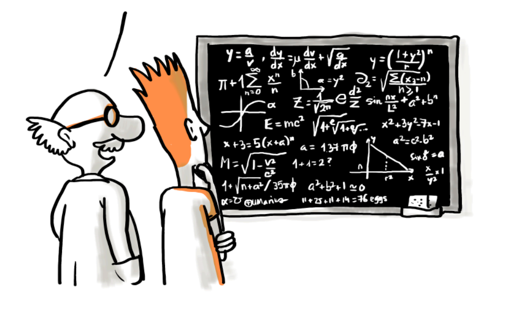
L5: Data Centers & Warehouse Scale Computers

- Data Center Basics
- Data Center Hardware
  - Server Hardware
  - Data Center Networking
  - Storage
- Building Infrastructure
  - Tier Classification System
  - Power Systems
  - Cooling Systems
- Energy and Power Efficiency
  - Power Efficiency beyond the Facility
  - Energy Efficiency of Computing
  - Energy-proportional computing
- Cost Modeling

Learning objectives

- Understand and explain the basic characteristics of **Warehouse Scale Computers**.
- Understand and explain the differences between **WSCs** and **Datacenters**.
- Understand and explain the **software infrastructure building blocks** of WSC.
- Understand and describe main characteristics of **WSC buildings** and their **power provision**.
- Understand and explain the **basic power, cooling components** of a modern datacenter.
- Understand and explain the concept of **energy efficiency** in datacenters, and associated mechanisms.
- Understand, and explain the cost structure of running a modern datacenter, and the concept of TCO (**total cost of ownership**).
- Explore, understand and explain concepts and techniques for energy-efficiency in datacenters.

Summary



- Explained why public cloud infrastructures rely on Warehouse scale computers for deploying and running their services.
- Reviewed the key requirements for large-scale public cloud infrastructures (WSC): high availability, cost-efficiency, fault-free operation, and the implications thereof.
- Examined the IT architecture of WSC, and the key characteristics of WSC clusters.
- Introduced terms like *Form Factor, Thermal Design Power, Total Cost of Ownership, TOR Switch, Bisection Bandwidth, Network Fabric, Network Attached Storage (NAS), Oversubscription Ratio of Intra to Inter-rack networking*
- Analyzed the key design considerations that determine the Form Factor of WSC servers
- Reviewed key percentages regarding energy usage of IT and other components of a WSC, and how these can drive decisions on workload management.
- Examined the structure of WSC network fabric and their main components.
- Discussed the concept of storage hierarchy, the components of the storage hierarchy of a WSC, and their key performance characteristics.
- Reviewed the basic functionality of distributed file systems, unstructured and structured storage of WSC.
- Discussed the pros and cons of Flash Storage vs Hard Drives and the implications of flash storage on networking performance requirements.
- Explored the main parts of a WSC's building infrastructure and their role and requirements.

Sample Questions

- What is the typical requirement for availability of Warehouse Scale Computers running cloud services?
- What are the key factors that determine cost-efficiency of WSCs?
- Draw an architectural diagram with the main IT components of a WSC and give indicative values for the key performance characteristics of these components.
- Give the definition of the Thermal Design Power (TDP) of a CPU and the Form Factor of a server, and explain how they affect the running cost of a Cloud infrastructure.
- Name three types of accelerators found in modern data centers and explain why they have been incorporated in modern DCs.
- What is a TOR? Describe its key characteristics.
- What is the bisection bandwidth of a rack and why it is an important consideration?
- Give a definition of the oversubscription ratio of a TOR switch. Why is it an important metric?
- What percentage (approximately) of the power consumption of a cloud server is spent by the CPU?
- What is a NAS and what are its pros and cons for storing cloud data?
- How is fault-tolerance and high-availability achieved in a distributed file system?
- Draw an architectural diagram with the memory hierarchy of a WSC and give indicative values for the key performance characteristics of these components.
- Explain what is redundancy in WSC, where it is used and what is the aim of adopting redundancy measures?

Sample Questions

- Suppose you plan for the development of a Tier-3 data center with 20000 servers, adopting best practices in the design/implementation. Can you provide an estimate of how much power will be required for cooling?
- Give the equation that defines the energy efficiency of a data center and explain its components
- What is the PUE? For a PUE of 1.4, what is the percentage of power going to IT components of a data center?
- What is SPUE and which power losses contribute to its value?
- What is the "True PUE" and what does it measure?
- Suppose the True PUE of a data center is 1.23. By how much is the total energy efficiency going to be improved if you eliminate all electromechanical overheads?
- How can we calculate energy efficiency of computing by running a benchmark?
- Name and explain the three main factors affecting the energy efficiency of computing.

L6: Cloud Infrastructure Software, Workloads, and Metrics

- Cloud Infrastructure Software
  - Platform-level Software
  - Cluster-level Infrastructure Software
  - Monitoring Infrastructure
  - Key Cloud Programming Concepts
  - Cloud Software
- Cloud Application Workloads
  - Giant-Scale Web Services
  - Web Search
  - Video Streaming
  - Machine Learning
- Availability and Performance
  - Faults and Failures
  - Availability in Giant-scale Services
  - CAP Theorem
  - Tail Latency Concerns

Learning objectives

- Understand and explain the concepts of **resource management**, **monitoring systems**, **performance debugging**, **blackbox** monitoring, **instrumentation, site reliability engineering.**
- Understand and explain the **software infrastructure building blocks** of WSC offering cloud services.
- Understand and explain the characteristics of **typical workloads running on WSC**.
- Review, understand and explain common techniques for **improving the performance and availability of WSC**.
- Be familiar and explain concepts like **cloud native, load balancing**, **sharding** (partitioning), **replication**, **integrity-checking**, **eventual consistency**, **redundant execution**, **tail-tolerance** in the context of cloud infrastructures.
- Understand, explain and apply the concepts of **Availability**, **Mean Time Between Failure** (MTBF) and **Mean Time to Repair**
- Understand and explain the **CAP theorem** and the concept of **tail-latency**.
- Understand and explain the concept of "**Cloud native**" software.

Sample Questions

- Describe the main functionalities of Resource Management software in cloud computing infrastructures. Provide some examples of the inputs that a RMS needs and the objectives it is trying to reach.
- Describe some core functionalities offered by Cluster Infrastructure Management softwares.
- Be familiar and explain concepts like load balancing, sharding (partitioning), replication, integrity-checking, eventual consistency, redundant execution, tail-tolerance in the context of cloud infrastructures.
- Explain what cloud native development means and what are the key requirements that it tries to meet.
- Failures in you data centre happen once a week. It takes 7 minutes to recover. What is your uptime and how can you realistically increase your uptime by an extra 9?
- Explain which concepts the letters of the acronym CAP correspond to and what the CAP theorem says.
- Describe a key difference in the profile of a cloud workload corresponding to Google's Web search and AdSense services. Explain why this difference can be very important for the design, implementation, and configuration of the underlying cloud infrastructure.