Chapter 1
Cloud Computing: A Comprehensive Introduction

Anita Lee-Post
University of Kentucky, USA

Ram Pakath
University of Kentucky, USA

ABSTRACT

Cloud Computing refers to providing computing and communications-related services with the aid of remotely located, network-based resources without a user of such resources having to own these resources. The network in question typically, though not necessarily, is the Internet. The resources provisioned encompass a range of services including data, software, storage, security, and so on. For example, when we use a mail service such as Gmail, watch a movie on YouTube, shop at Amazon.com, or store files using Dropbox, we are using cloud-based resources (The Google Chrome Team, 2010). In this chapter, the authors examine the evolution of Cloud Computing from its early roots in mainframe-based computing to the present day and also explain the different services rendered by Cloud Computing in today’s business and personal computing contexts. This chapter provides a comprehensive view of the rapidly flourishing field of Cloud Computing and sets the stage for more in-depth discussions on its security, trust, and regulatory aspects elsewhere in this compendium.

1. THE EVOLUTION OF CLOUD COMPUTING

The adjective “Cloud” in Cloud Computing refers to the network used for service provisioning. In diagrams describing cloud-based services, the cloud is often literally depicted as the outline of a hand-drawn cloud on paper. The use of cloud-like shapes in diagrams depicting networks such as the Internet dates back many years and is a staple of mainstream text books and articles on data communication networks. The term “Cloud Computing,” though, is relatively new. To better comprehend this relatively nascent phenomenon, let us go back in computing history and examine earlier models of provisioning services over a communications network, i.e., the precursors of present-day Cloud Computing.

DOI: 10.4018/978-1-4666-5788-5.ch001
1.1 Time-Sharing on Mainframe Computers

The early 1950s saw the advent of commercial “mainframe” computers such as the IBM 701. These computers were single-user, non-shareable, one-job-at-a-time systems and were rented by companies for about $25,000 a month. Several programmers signed up, on a first-come-first-served basis, for “sessions” on a mainframe where each session was a block of time dedicated to processing a single “job” (i.e., a program). Each programmer took about 5 minutes to set-up his/her job including punching in at a mechanical clock, hanging a magnetic tape, loading a punched card deck, and pressing a “load” button to begin job processing (Chunawala, n.d.). Inefficiencies in the process due to excessive manual intervention resulted in much wasted processing time even as jobs were queued and often delayed.

To improve process efficiency, General Motors (GM) and North American Aviation (NAA) (today, part of Boeing) developed an operating system, the GM NAA I/O (Input/Output) system and put it into production in 1956 (Chunawala, n.d.). This heralded the advent of “batch processing” where multiple jobs could be set up at once and each run to completion without manual intervention (“Batch Processing”, n.d.). Further improvements were realized with the advent of the IBM System 360 mainframe in 1964 which separated I/O tasks from the CPU (Central Processing Unit) and farmed these out to an I/O sub-system, thus freeing up the CPU to perform computations required by a second job when another job was interrupted for I/O operations. Batch processing offered several benefits: Individual jobs in a batch could be processed at different times based on resource availability, system idle time was reduced, system utilization rates were improved and, as a consequence, per-job processing costs were reduced.

With batch processing, a computer’s time is considered considerably more valuable than a human’s and human work is scheduled around the machine’s availability. In contrast, “interactive computing,” considers a human’s time as being the more valuable and views a computer only as a capable “assistant.” Early implementations of interactive computing include the IBM 601 that allowed a single user interactive use at a time. However, allowing one user to monopolize a scarce resource also resulted in considerable inefficiency in resource utilization. On the other hand, offering several interactive users seemingly concurrent usage would result in better use of the electronic assistant (“Interactive Computing”, n.d.). In 1961 MIT introduced the world’s first Time Sharing Operating System, the Compatible Time Sharing System (CTSS). In due course, IBM introduced a Time Sharing Option (TSO) in the OS 360 operating system used in the IBM System 360. Time Sharing introduced further processing efficiencies over batch processing. Rather than process a job in its entirety, time sharing would devote a short duration of time called a “time slice” to processing a job and then turns to devote similar attention to another job. The CPU so rapidly switches from job to job that it appears to each user that his/her job has the full and complete attention of the CPU -- a user experiences no noticeable delays.

A natural outgrowth of interactive computing was remote access to a computer via terminals. Several terminals were “multiplexed” over telephone lines using individual modems to connect users to a single mainframe. Shared mainframe interactive access and use via multiplexed terminals and the telephone network may be regarded the earliest Cloud Computing model although it was then referred to as Time-Sharing. In the 1960s, several vendors offered Time-Sharing “services” to businesses. These included Tymshare, National CSS, Dial Data, and BBN using equipment (mainframe and minicomputers) from IBM, DEC, HP, CDC, Univac, Burroughs, and others.