Chapter 7
Infrastructure as a Service

Shamim Hossain
IBM Corporation, Australia

ABSTRACT

Infrastructure as a Service is the pillar on which a cloud computing architecture is built. With the advancement of technologies in communications, computing, and storage devices, IaaS has emerged as a highly efficient platform to construct SaaS and PaaS layer on top of it. IaaS solutions vary from an organization to another. One single solution does not fit all. This chapter looks at the general constituents of IaaS.

INTRODUCTION

Infrastructure as a Service is one of the building blocks of cloud computing. It is the linchpin which revamps the cloud solution by offering underlying IT infrastructure. In this chapter we will take a closer look at this important service which is considered as the backbone of cloud computing. Massive data centers and servers farms comprise the cloud's underlying infrastructure on which SaaS and PaaS run. According to Matthieu (Hug, 2008) from his online article:

An emerging computing paradigm where data and services reside in massively scalable data centers and can be ubiquitously accessed from any connected devices over the internet.

From a very high level these massive data centers are cloud IaaS. We will see cloud data center architecture later in the chapter. These data centers provide large scale of economies and dynamic scalability among others.

In many different literatures, several different versions of definition for IaaS exist. The definition from NIST has been widely accepted. Let us revisit the definition of IaaS as defined by NIST (NIST, 2010).

The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of selected networking components (e.g., host firewalls).

DOI: 10.4018/978-1-4666-6539-2.ch007

IaaS itself is comprised of many different components. A list of components is provided here. Detailed discussion will follow throughout the chapter. Components for IaaS include, but are not limited to, the following (Rittinghouse & Ransome, 2010) (Reese, 2009):

- Servers (both physical and virtual)
- Storage systems by means of NAS and SAN
- Network segmentation using different network blocks and VLANs
- Communication network (including routers, switches, firewalls, load balancer, etc.)
- High Speed Internet connectivity (often on OC 192 backbones)
- Platform virtualization environment
- Service-level agreements
- Utility computing billing
- Security by means of hardware or VM based firewall and intrusion detection & prevention system
- Hardware load balancer
- DNS, DHCP and other management and support services
- Power, cooling and disaster recovery system

Many of terms and jargons above will be discussed throughout this chapter. In this chapter we first present a background of cloud data centers so that users can get an overview of them before looking at technical details. It is followed by a detailed description of cloud components and some energy efficiency metrics to measure data center energy usage. Cloud Components section discusses about access devices, high-speed broadband access, virtualization and functional areas of a data center (network, computing and storage infrastructure and security services). Later in this section we provide a complete picture of a generic cloud data center. Then we discuss different additional attributes of IaaS in the Section IaaS Characteristics. Before conclusion we present a section on Cloud Standard Bodies. These standard bodies have been working for the emergence of open standards and interoperability of technologies around cloud computing.

BACKGROUND

Before we delve into the details, let us look at the massive scale and size of some of the cloud data centers by Internet Powerhouses like Google, Microsoft, Amazon and Yahoo. Technology provider IBM, Google and others are investing huge amount of money to spur innovation of cloud computing. Cloud's scalability and elasticity emanates from its massive scale of economies. Gigantic data centers from the tech titans are leveraging the economies of scale in computing power, energy consumption, cooling, site operations and administration (Erdogmus, 2009). Cloud service providers are acquiring more and more computing powers and expanding their data centers (Economist, 2008). According to this report from Economist, Google has more than 30 data centers comprised of over 1 million servers. Organizations tend to build large-scale data centers in areas where affordable land, good communications (e.g. optical fiber connections), water for cooling and cheap electricity are readily available. It is not surprising that today's large-scale cloud data center hosts even million servers (Katz, 2009). Following is an excerpt from this paper.

These new data centers are the physical manifestation of what Internet companies are calling cloud computing. The idea is that sprawling collections of servers, storage systems, and network equipment will form a seamless infrastructure capable of running applications and storing data remotely, while the computers people own will provide little more than the interface to connect to the increasingly capable Internet.