Moving to the Cloud:
Estimating the Internet Connection Bandwidth

Luís Ferreira da Silva[1], Fernando Brito e Abreu[1,2]

1) QUASAR, CITI, FCT/UNL, 2829-516 Caparica, Portugal

luis.silva@di.fct.unl.pt

2) DCTI, ISCTE-IUL, 1649-026 Lisboa, Portugal

fba@iscte.pt

Abstract

IT Infrastructures (ITIs) have long been understood in terms of people and resources such as servers, routers, firewalls and operating systems, among other components, running and providing services inside the organization. The need to reduce the cost of ITI ownership, by offloading its capacity to third parties, has motivated organizations to consider the *Cloud Computing* alternative. The main drawback they face when opting for the cloud is the dependency on and requirements of the internet connection, since it must be fast and reliable. This paper addresses this concern by providing guidelines for estimating internet connection bandwidth requirements for a prototypical cloud-based organization represented by means of an IT infrastructure pattern. ITI patterns are reusable and proven solutions to support the ITI design process and to facilitate the communication among stakeholders.

Keywords: *Information Technology, IT Infrastructure, Design Patterns*

1 Introduction

Cloud Computing is an attractive offer because it can provide several capabilities to organizations. Among those capabilities are (i) the unlimited computing power, (ii) broad network access from anywhere and from multiple devices (e.g. tablets, computers, mobile phones, etc.) and (iii) reduction of IT costs due to automation and elasticity that allow organizations to pay only for what they consume. There is little consensus on how to define the term *Cloud Computing* [Geelan 2008]. The US National Institute of Standards and Technology (NIST) defines *Cloud Computing* as *"The model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."* [Mell, Grance et al. 2009]. Cloud-based solutions existing today, include communications solutions (e.g. email, online meetings, telephony), collaboration solutions (intranet and extranets, document storage, workflows), business applications, storage, ITI management software and other ITI solutions. There are three main delivery models in *Cloud Computing*, which are *Software as a Service* (SaaS), *Platform as a Service* (PaaS) or *Infrastructure as a service* (IaaS). In the *SaaS* delivery model an application is offered to a user as a service. The user consumes the application through an interface such a web browser or plug-ins and not a "locally-installed" application. Examples of *SaaS* applications include E-mail, Customer Relationship Management (CRM) and Web and Video conferencing software. The *PaaS* delivery model consists in an application platform to build

applications and services completely from the internet, without the need to download and install software. Like any application platform, a *PaaS* environment includes design, development, testing, deployment and hosting. The *PaaS* provides also the supporting infrastructure capabilities, such as authentication, authorization, session management, transaction integrity, reliability, availability, and scalability. The *IaaS* is a delivery model in which an organization outsources the equipment used to support operations, including storage, hardware, servers and networking components. These three cloud delivery models provide several cloud-based solutions that virtually replace all services currently provided by installed ITIs within the organization.

Cloud Computing is more a computing model than a technology. This computing model assumes that services are delivered over the internet (with the exception of private clouds where services are provided inside the ITI). *Cloud Computing* uses technologies like virtualization (farms of physical servers with multiple virtual machines), automatically provisioning (servers installed unattended and automated) and internet connections.

The use *Cloud Computing* solutions is becoming simple, since all that is really needed is an ITI LAN with an internet connection with enough bandwidth and a credit card or other payment method to subscribe the service. Large organizations often have more than one internet connection (for redundancy) with enough bandwidth, they use several appliances to provide traffic optimization, traffic shaping and in most cases connection are symmetrical (sending and receiving data at the same rate). Small and medium organization however frequently have a single internet provider through a single port on a router and all systems connecting through the LAN share the Internet bandwidth equally. Independently of the organization size moving from services provided by local ITIs to online services provided by cloud vendors presents several networking challenges [Zhang, Cheng et al. 2010].

The main focus of the paper is to present an ITI cloud pattern, to simplify the challenge task of estimating the amount of bandwidth required when organizations decide to provide services in public clouds.

2 Related Work

2.1 Network Performance

The network performance effectiveness often depends directly on the efficiency with which a network delivers data [Peterson and Davie 2000]. The most common terms used to refer the overall network performance are the *speed*, *bandwidth*, *throughput* and *latency*. These terms are often presented as synonyms but they are different. The term *speed* is generic and often refers to the nominal speed of a networking technology (e.g. Fast Ethernet has a nominal speed of 100 megabits per second). *Bandwidth* represents the maximum amount of data that can be moved over a given link or connection in a unit of time. The *throughput* is the actual speed over a link or connection (e.g. an organization may have a network with 100 Mbps Ethernet and a throughput of 75 Mbps for instance). The *throughput* of a network is measured in bits per second and is the average data rate over a specific communications link. *Latency* or delay is the time (usually measured in *ms*) required to transfer a single empty message from one source to the destination (the higher the *latency* the longer it takes to transfer data) and can be measured as round-trip time (also known as the ping time). Research has been conducted to minimize the impact of *latency* and maximize network capacity through effectively manage network communications to deliver data using the most efficient path with the highest available bandwidth to provide higher quality. The quality is measured by the data flow consistency in terms of performance and is considered good when is above 90%. If there is congestion (too much traffic) or regulation (intentional ISP delays) and the data flow at different speeds the percentage will drop. Notice that even a slow connection can have 99% quality. A fast

connection speed with inconsistent *throughput* can present more application problems (e.g. VoIP), than a slower connection with consistent *throughput*. Other common used metrics that should be assessed when evaluating a change to a public cloud are *Transfer Time*, *Transfer Rate*, *Quality* and *TCP Window Size*. The *Transfer Time* is the time that individual messages take to be transferred between two interconnected computers. *Transfer Rate* represents the speed at which data can be transferred (usually measured in *bps*) from one place to another. With the *Latency* and *Transfer Rate* it is possible to calculate the time to transfer a message with a specific number of bits (message size). The *TCP Window Size* is the maximum amount of received data, in bytes, which specifies the number of bytes that can be buffered at one time on the receiving side of a connection. The sending host can send only that amount of data before waiting for an acknowledgment. The window size can be defined in servers, however special care must be taken since changing this setting will affect the amount of memory (for buffering) needed. Another option is to use WAN accelerators at each end, which will use a larger TCP window without requiring tuning on servers.

2.2 Bandwidth estimation

Bandwidth estimation has been an active area of research for several years. There are three main bandwidth related metrics which are (i) *capacity* that represents the maximum link transmission rate, (ii) *available bandwidth* which is the unused or spare capacity during a certain time period, and (iii) *bulk transfer capacity* that can be defined as the maximum throughput achievable by a single TCP connection. Obtaining bandwidth indications like those, provides valuable information to assess existing network ITIs and to support the definition of a cloud strategy plan. Several techniques and tools for predicting bandwidth across network paths have been proposed to obtain more efficient end-to-end communications, such as Available Bandwidth Estimation Techniques and Tools (ABETTs) [Prasad, Dovrolis et al. 2003]. These techniques and tools can be useful to improve the speed of services provided by performing actions such as monitoring and verifying SLAs, determine the best network topology to maximize efficiency, among other aspects. The main limitation with ABETTs is that not all tools are suited to estimate bandwidth across internet access technologies [Guerrero and Labrador 2010].

Most of these tools work by injecting specially designed streams of probe packets and then observe the end-to-end delays to estimate the available bandwidth. The techniques used with these tools vary and can be classified in direct probing and interactive probing. In direct probing it is assumed that the tight-link[1] capacity is known and each probing stream results in a sample of the available bandwidth. Examples of these techniques are Delphi [Ribeiro, Coates et al. 2000], IGI [Ningning and Steenkiste 2003] and Spruce [Strauss, Katabi et al. 2003]. Interactive probing does not assume any knowledge about the tight-link capacity and is based in self-induced congestion which consists of sending streams of packets whose input rate iteratively increases. The available bandwidth is the lowest input rate overloading the network. Examples of these techniques are Train of Packet Pairs (TOPP) that uses trains of packet pairs in each probe stream [Melander, Bjorkman et al. 2000], Self-Loading Periodic Streams (SLoPS) [Suman Banerjee 2000; Jain and Dovrolis 2003] and ”chirps” which are streams of exponentially spaced packets [Ribeiro, Riedi et al. 2003].

There are numerous network measurement tools, mainly focusing on performance evaluation. There are two network measurement techniques which are *passive* and *active* measurement. The passive measurement relies on monitoring existing traffic between end-hosts to extract estimates. Tools for passive measurement do not generate extra traffic. Examples of these types of tools include *Nettimer* (for bandwidth estimation), *Viznet* (for throughput tests), *Sting* (for latency tests) among others. These measurement tools are less applicable due to the fact that existing traffic is not always suitable to produce an indicative estimate [Saroiu,

[1] - link with the minimum available bandwidth of a path

Gummadi et al. 2002]. Active measurement encompasses sending streams of probing packets to explore the entire network. Example of active measurement tools include *bing, b|c probe, clink, iperf, netperf, pathload, pathrate, PathView, pchar, SProbe, TReno, ttcp, nttcp, Nettimer* and *pathchar,* among others, as described in the performance measurement tools taxonomy maintained by CAIDA [CAIDA 2011].

3 CloudTraffic Estimator Pattern

The concept of ITI conveys the use of various components of information technology (hardware, software and network infrastructure) upon which IT services are provided [Sirkemaa 2002]. To be aligned with the business ITIs must be quickly adapted to support new technologies or paradigms (e.g. Cloud computing, Grid services, Web services, internet applications, and application integration) and new types of services (e.g. wireless, broadband media, and voice services), while enforcing stronger access control and auditing policies and keeping high degrees of flexibility and agility.

In such a scenario, one of the major problems faced by ITIs is their increasing size and complexity, that may jeopardize the delivery of real business value [Sessions 2008]. The size and complexity are often the result of ITIs created, designed or adapted by non ITI experts such as business decision makers, consultants, administrators, developers, software engineers, solution architects and other individuals (sometimes conflicting due to their own point of view) without ITI design guidelines, the migration of a simple application to the cloud may affect all the services provided. Designing or changing ITIs is a challenge task mainly because it requires knowledge of existing organization processes, the views of different players, and the coordination of technical expertise in three ITI domains (hardware, networking and infrastructure software) that rarely reside in a single individual.

The design of solutions is achieved in most engineering fields by using appropriate abstractions. Although often the devil is on the details, raising the level of abstraction allows practitioners to find, share and apply standardized solutions to recurrent phenomena, by only retaining the information which is relevant for a particular purpose.

In the area of IT infrastructures the application solutions to recurrent problems was caught as a business opportunity by several companies to standardize typical ITI building blocks based on their commercial components. Some of those companies developed methodological approaches to ITI pattern-based design, by proposing design "blueprints" embodying vendor-specific components [Trowbridge, Mancini et al. 2003; Lofstrand and Carolan 2005].

The use of ITI design patterns can be seen as a process to simplify the ITI design process, while reducing its risk and cost through the use of well-known solutions for recurrent problems. The solutions addressed by design patterns are not intended to be static and final. In fact, they are templates that can be customized and extended. Design patterns help breaking ITI complexity into smaller modules, thus allowing architectural decisions to be taken at a higher abstraction level. Design of infrastructures using this approach makes them more robust, scalable, reliable, and maintainable.

This following section presents one of several ITI design pattern for cloud computing with the name "*CLOUDTRAFIC ESTIMATOR".* From the several ways to organize patterns [Gamma, Helm et al. 1995; Buschmann, Meunier et al. 1996; Fowler 2006] we decided to use a structure similar to GoF, since it is one of the most structured and well-known forms.

CLOUDTRAFFIC ESTIMATOR

3.1 Context

The level of connectivity to access a public cloud is crucial to make the cloud deliver the best services. Different organizations have different sizes, require different services and have different needs. Depending on the bandwidth of the internet connection and the amount of data exchanged with the cloud provider, the experience for end users may vary. Most organizations simple assume they need a broadband IP VPN, whether or not the latter delivers the performance, reliability, availability and security required to access the public cloud.

3.2 Example

An organization has several applications hosted in internal ITI and an internet connection to provide employees access to the internet and to some internal applications such as corporate e-mail that needs to exchange messages with other corporations via internet and "customer facing" applications. There is also some applications developed in-house which are business critical and heavy utilized during some periods. Mainly to reduce computing costs, achieve a more flexible computing environment and ensure capacity is there when needed, this organization decided to embrace *Cloud Computing* and evaluate the impact on internet connection bandwidth of moving some of these applications to a public cloud.

3.3 Problem

How to estimate the required internet bandwidth when moving applications to a public cloud?

3.4 Forces

The following forces influence the solution:

- *Traffic Pattern:* The internet traffic produced by users, applications or network devices, may vary according to a period of the day, week, and month, what influences the required bandwidth.
- *Multiple Purposes:* The internet connection is shared between users and applications and network devices. The available bandwidth has impact in the user's experience.
- *Capacity:* The number of users, applications and network devices influence the internet bandwidth requirements. More applications in cloud tend to require more internet connection bandwidth.
- *Operations:* The internet bandwidth requirements are influenced by the type of operations performed. Different users, applications or network devices perform different operations and may require different bandwidth requirements.

3.5 Solution

Evaluate network performance by using traffic generators, network analyzers and active measurement tools and create a network capacity plan.

The network capacity plan is an important instrument to define what services and applications can be migrated to the cloud and what will be the impact on existing infrastructure. The network capacity plan should be integrated with cloud adoption strategy, which has detailed information regarding business objectives, effort, business impact and cost analysis, risks, among other aspects. The general objectives of the network capacity plan are:

- Understand current services and applications network capacity requirements.
- Document assumptions regarding requirements and workload forecasts.
- Define the required network capacity forecast for services and applications.

- Provide network recommendations to ensure that there is sufficient network capacity to support the forecasted workload.

It is important that the network capacity plan includes the (i) *Definition of Service Level Requirements* that should categorize services and applications, quantify user's expectations, define workloads and identify service levels for each workload, the (ii) *Analysis of Current Capacity* to understand services and applications requirements on internet connection bandwidth, and the (iii) *Planning for Future Capacity* to forecasts future needs and system requirements by determining future processing requirements to maintain the service levels.

To evaluate network performance there are two types of tools that can help to analyze current capacity and plan future capacity which are (i) Traffic Generators and (ii) Network Analyzers. The *Traffic Generators* also called load generators are used to generate dummy packets and keep track of the packet delivery in the network and useful to view and analyze the performance and capacity of existing devices, network topologies and internet connections. *Network Analyzers* are important to provide more information regarding ITI internal applications and determine connection requirements needed to support users through a network traffic analysis. This analysis must capture incoming and outgoing traffic data to and from each application. A typical system to perform these analyses has three components: (i) sensor (or sensors), (ii) collector and (iii) reporting system. The sensor is also known as a probe, and is an agent that listens to the network and captures traffic data. The sensor may capture traffic from switches, routers and firewalls, among other devices. The collector is a server that receives and stores data from sensors. The reporting system is responsible for analyzing stored data and producing network traffic reports (Figure 1).

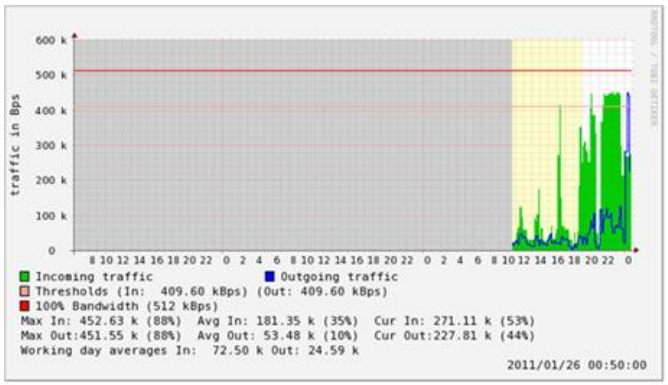

Figure 1 – Incoming and Outgoing traffic captured using network analyzers.

Multiple networking factors can have impact and affect the delivery of data between the organization and the cloud. The network capacity plan should include details regarding latency, packet loss, retransmission and throughput, network devices performance (e.g. endpoints, routers, switches and hubs), bandwidth usage, type of circuit(s) used, upstream and downstream transfer rate, number of remote locations and their access to the cloud.

To ascertain which workloads are the major, which help to narrow the attention to the workloads that are making the greatest demand on internet connection, the network capacity plan should identify for each service and application the following aspects:

- How many people are using the service or application.
- Current response time (e.g. excellent, fair, poor) from multiple ITI locations.
- Application or service peak usage (time of day, number of users).
- Type of service or application (e.g. e-mail, office, voice, video conferencing).

- Application or service usage patterns (some applications have seasonal utilization patterns such as end of week, end of month.).
- Impact on application or service in the internet connection bandwidth.
- Devices used to connect to services or applications.

Due to the demand on internet connection, the internet access to address cloud requirements is frequently dedicated, high-end with low latency and high bandwidth.

It is normal recommended to establish a service level agreement between the organization and the internet service provider to define acceptable services. The service level agreement should be defined in terms of networking performance metrics such as response time and throughput.

3.6 Consequences

The **CLOUDTRAFFIC ESTIMATOR** provides the following **benefits**:

- *Decrease Costs:* Organizations will have the opportunity to have an internet connection bandwidth according to their needs.
- *Cloud Strategy:* Identify services and applications that consume most network resources. Understand how much bandwidth each service and application is using before moving to the cloud helps in the cloud strategy definition.
- *Improve Quality:* Identify what will be the services and application requirements bandwidth to provide the expected service quality. These requirements are important to help in the definition of service level agreements and to have a reliable and optimized connectivity.

On the other hand, the pattern carries several **liabilities**:

- *Time Consuming:* Depending on the number of services and applications the network performance analysis may take some time.
- *Increase Costs:* Based upon the network capacity plan analysis it may be necessary to have a different type of internet connection or a bandwidth increase.
- *Point in time analysis:* New services and application are implemented decommissioned, so the network analysis is a "current point in time".

3.7 Example Resolved

Based upon the capacity planning analysis, the organization was able to understand the impact that existing services and applications have in the internet connection and the amount of traffic associated with each application deployed in ITI (that can benefit from being in a public cloud). The amount of traffic associated with each application, was very important to classify applications according to their internet connection requirements and better evaluate where the public cloud delivered strategy, provides de most benefits and gives the highest return value.

3.8 Related Patterns

The **CLOUDTRAFFIC ESTIMATOR** is related with the IT Infrastructure Patterns for *Cloud Computing* presented in Figure 2.

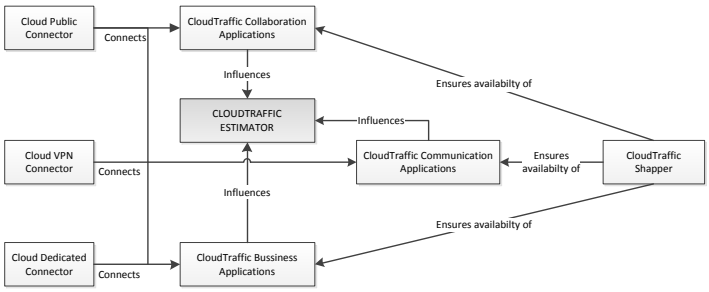

Figure 2 – Related ITI Design Patterns for cloud computing.

4 Conclusion

Different organizations may have different cloud strategies and will need to send different services and application with different requirements to the cloud. Before starting the migration process is important to figure out which services and applications are possible candidates to migrate, based upon a general cloud strategy and the number of network resources each consume. There are multiple networking factors such as latency, packet loss, retransmission and throughput, bandwidth usage that can have impact and affect the delivery of data between the organization and the cloud. This paper presents an IT Infrastructure Pattern for Cloud Computing to solve the problem of estimating the internet connection bandwidth requirements when moving to the cloud. The pattern defines a methodology that is based upon extensive research conducted to estimate bandwidth requirements in the networking area. Building a network capacity plan to assess applications and determine their impact in existing internet connection bandwidth is a crucial step to build a successful cloud adoption strategy.

5 References

Buschmann, F., R. Meunier, et al. (1996). A system of patterns: Pattern-oriented software architecture. New York, Wiley.

CAIDA. (2011). "Performance Measurement Tools Taxonomy." Retrieved June 2011, from http://www.caida.org/tools/taxonomy/performance.xml#bw.

Fowler, M. (2006). "Writing Software Patterns." from http://martinfowler.com/articles/writingPatterns.html.

Gamma, E., R. Helm, et al. (1995). Design Patterns: Elements of Reusable Object-Oriented Software. New York, Addison-Wesley Professional.

Geelan, J. (2008). "Twenty one experts define cloud computing." Cloud Computing Journal.

Guerrero, C. D. and M. A. Labrador (2010). "On the applicability of available bandwidth estimation techniques and tools." Comput. Commun. **33**(1): 11-22.

Jain, M. and C. Dovrolis (2003). "End-to-end available bandwidth: measurement methodology, dynamics, and relation with TCP throughput." IEEE/ACM Trans. Netw. **11**(4): 537-549.

Lofstrand, M. and J. Carolan (2005). Sun's Pattern-Based Design Framework: The Service Delivery Network. Santa Clara, CA, USA, Sun Microsystems.

Melander, B., M. Bjorkman, et al. (2000). A new end-to-end probing and analysis method for estimating bandwidth bottlenecks. Global Telecommunications Conference, 2000. GLOBECOM '00. IEEE.

Mell, P. a. Grance, et al. (2009). "The NIST Definition of Cloud Computing." National Institute of Standards and Technology(6): 50.

Ningning, H. and P. Steenkiste (2003). "Evaluation and characterization of available bandwidth probing techniques." Selected Areas in Communications, IEEE Journal on **21**(6): 879-894.

Peterson, L. L. and B. S. Davie (2000). Computer networks: a systems approach San Francisco, CA, USA, Morgan Kaufmann Publishers Inc.

Prasad, R., C. Dovrolis, et al. (2003). "Bandwidth estimation: metrics, measurement techniques, and tools." Network, IEEE **17**(6): 27-35.

Ribeiro, V., R. Riedi, et al. (2003). "pathChirp: Efficient available bandwidth estimation for network paths." Passive and active ….

Ribeiro, V. J., M. J. Coates, et al. (2000). Multifractal Cross-Traffic Estimation. ITC Specialist Seminar on IP Traffic Measurement.

Saroiu, S., P. Gummadi, et al. (2002). "Sprobe: A fast technique for measuring bottleneck bandwidth in uncooperative environments." IEEE INFOCOM.

Sessions, R. (2008). Simple Architectures for Complex Enterprises. Redmond, Microsoft Press.

Sirkemaa, S. (2002). IT infrastructure management and standards. Proceedings of the International Conference on Information Technology: Coding and Computing. Las Vegas, IEEE Computer Society**:** 201.

Strauss, J., D. Katabi, et al. (2003). A measurement study of available bandwidth estimation tools. Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement. Miami Beach, FL, USA, ACM**:** 39-44.

Suman Banerjee, A. K. A. (2000). Estimating Available Capacity of a Network Connection. Proceedings of the 8th IEEE International Conference on Networks, IEEE Computer Society**:** 131.

Trowbridge, D., D. Mancini, et al. (2003). Enterprise solution patterns using Microsoft. NET version 2.0. Patterns & Practices**:** 342.

Zhang, Q., L. Cheng, et al. (2010). "Cloud computing: state-of-the-art and research challenges." Journal of Internet Services and Applications **1**(1): 7-18.