Impact of RAS for Cloud Computing

Tejash Panda [*,1]

*Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616, USA

Abstract

Cloud computing is a technology that allows for the delivery of computing resources, such as data storage and processing power, over the internet. It has gained popularity due to its ability to reduce costs and increase flexibility and scalability for businesses. However, there are also concerns about the security, reliability, and availability of cloud computing, which have led some organizations to be cautious about adopting it. These issues have been widely discussed in the field of cloud computing, and are considered critical to its widespread implementation.

This article follows RAS and its impact on cloud computing. Firstly, we will cover what is cloud computing to show why RAS is needed, a key element of the cloud infrastructure. Then we will dive into the impact RAS has on the cloud infrastructure.

I. Cloud Computing: an Introduction to the Cloud

Cloud computing refers to the delivery and access of computing resources, including data storage and processing power, over a network, typically the internet. It aims to reduce complexity for clients by allowing them to virtually store and access data, applications, and technologies remotely, rather than on personal computers or local servers. This is achieved through the use of virtualization technologies and self-service access to computing resources via network infrastructure. In cloud environments, multiple virtual machines are hosted on the same physical server [1]. Customers pay for only the resources they consume and do not need to invest in local storage or infrastructure. Furthermore, Cloud computing encompasses both applications delivered as services over the internet along with the hardware and systems software in data centers that provide those services [1]. Currently, there are three types of cloud environments: public, private, and hybrid [3].

A public cloud is a model in which providers make a variety of resources, such as applications and storage, available to the public [3]. These services may be free or come with a fee. In a public cloud environment, applications are run by large service providers and offer some advantages over private cloud environments.

A private cloud is a set of internal services that are not available to the public [3]. It is essentially an architecture that provides hosted services for a specific group of people behind a firewall.

A hybrid cloud is a combination of a private and public cloud, where a company controls some resources internally and provides others for public use [3]. In this model, the cloud provider offers a service that allows the creation of a private cloud, which is only accessible by internal staff and protected by a firewall from external access, as well as a public cloud environment for external user.

Cloud Services

Cloud computing is a style of computing where highly scalable and flexible IT capabilities are provided as services to external customers over the internet [1]. Cloud providers offer a variety of services, known as "Everything as a Service" (XaaS), which can include the following main services[2] [7]:

A. SaaS

Software as a Service (SaaS) is a type of cloud service that offers network-hosted applications. SaaS is a delivery model in which cloud providers develop and host web-based software applications and make them available to customers over the internet [2] [7]. As a result, customers do not need to purchase software licenses or additional hardware and typically only pay fees, also known as annuity payments, on a periodic basis to use the cloud provider's web-based software. There are two main types of SaaS: business applications, which provide software to help businesses perform tasks efficiently, and development tools, which are used primarily for product development and management.

B. PaaS

Platform as a Service (PaaS) is a type of cloud service that provides users with a platform to build, test, and deploy applications without the need to install and maintain the platform on their own local machine [2] [7]. This allows users to focus on developing and improving their applications without having to worry about the underlying infrastructure.

C. IaaS/HaaS

Infrastructure as a Service (IaaS) is a type of cloud service that involves outsourcing the infrastructure used to support an organization's operations, such as storage, hardware, servers, and networking components, to a service provider [2] [7]. The provider owns and is responsible for maintaining the equipment and the client pays for its use on a per-use basis. IaaS is sometimes referred to as Hardware as a Service (HaaS).

There are several groups of clients that use cloud computing for various purposes.

These include regular customers, academics, and enterprises.

Regular customers are primarily concerned with the service and privacy of their data on the cloud, and often use Software as a Service (SaaS) [1] [2] [7].

Academics often have good networks and prefer to use the infrastructure they already have to improve the performance of computations and overcome grid limits. Cloud computing allows them convenient access to high-performance clusters or grid-based computation infrastructure without the need to purchase new hardware [1].

Enterprises, particularly in the IT industry, often use cloud computing to reduce costs and improve performance in their businesses. Some companies have even entered the cloud-related industry or use cloud services to achieve these goals [1].

II. Cloud RAS

Reliability, Availability and Serviceability (RAS) are three important attributes that must be taken into consideration when designing, purchasing, or using a computer product or component. RAS is relevant to both hardware and software, and can be applied to a wide range of computer-related systems, such as networks, applications, operating systems, personal computers, servers, and supercomputers [4].

Cloud service providers are expected to ensure high levels of reliability and availability in order to provide the best quality of service to users. They may use solutions such as partitioning to optimize performance. The management and control of these performance parameters, including RAS, may vary depending on the type of cloud environment, whether it is public, private, or hybrid [3].

Reliability refers to a product's ability to consistently perform according to its specifications [4]. To an user a system is reliable when an active system; can be accessed from any device from any location, provides no downtime or interruptions throughout the day, and has the system complete a task as intended.

Availability is the proportion of time a system or component is functional to the total time it is expected or required to function. In any given network, achieving high availability is the goal for all systems [4]. For cloud vendors and customers, it is important that the system remains available at all times to ensure that there are no interruptions in a workflow.

Serviceability refers to the ease with which a system or component can be maintained and repaired, and early detection of potential

problems is important in this regard [4]. Some systems have built-in features that allow them to automatically correct problems before they cause serious issues. Ideally, maintenance and repair operations should cause as little downtime or disruption as possible.

Although RAS may seem simple on the surface, it has a huge impact on the network built on these cloud platforms. Ensuring high RAS levels is crucial for the smooth operation of any system and can mean the difference between success and failure. For example, in August of 2013, NASDAQ experienced a three-hour power outage, this resulted in the inability of the system to fully revert to backup mode [5]. NASDAQ's temporary crash highlighted the importance of reliability, availability, and serviceability in the design and operation of their computer systems [5]. This RAS failure had significant consequences for stock investors, moreover, in the current era of online global marketplaces such as NASDAQ, the impact of this type of failure can be particularly severe.

However, it is important to acknowledge that despite the best efforts in planning and preparation, incidents and failures can still occur. In such instances, it is crucial to have effective strategies in place for addressing and mitigating the impact of these events. This may include measures such as implementing robust disaster recovery plans, conducting regular testing and simulations, and implementing monitoring and alert systems to detect potential issues early on. By anticipating and addressing potential failures, cloud service providers can minimize the impact of such events and maintain the reliability of their cloud computing systems.

Furthermore, having cloud service providers offer servers to a diverse range of clients comes at the cost of millions of servers which can lead to complexity in error detection in the short term. To mitigate this complexity, we determine failures through the unit of "FITs", or "Failures in Time" [6]. Failures can occur at different stages of a device's lifespan and often due to various circumstances, high altitude, extreme temperature differences and even cosmic rays can cause errors in the machine. These errors are typically categorized into three phases: the early life failure rate, the normal life phase, and the wearout phase [6].

The early life failure rate, also known as infant mortality, is characterized by a higher initial failure rate that decreases rapidly [6].

The normal life phase is characterized by a relatively constant failure rate, which is described in units of FITs or as a "Mean Time Between Failures" (MTBF) in hours [6].

The wearout phase represents the point at which intrinsic wear-out

mechanisms begin to dominate and the failure rate begins increasing exponentially [6]. The product lifetime is typically defined as the time from initial production until the onset of wear-out.

In our case, we aim to maintain a low FIT during the normal life phase of a server and we do so by Error detection, correction and logging.

III. Error Detection, Correction and Logging

Errors occurring on cloud server platforms are inevitable. However, they can be predicted and detected using various methods. A common method of error detection is Hamming Parity Error Detection however, like many other methods of error detection, it has its limits. Furthermore, errors can occur on components that aren't being actively used. Even though errors come in many different forms, Platform Architects and Engineers have grouped errors into two discrete categories: Correctable and Uncorrectable errors[8].

To ensure RAS maintains operation of a system, it is important to implement mechanisms for detecting and addressing hardware errors in real-time. In addition, monitoring for signs of hardware degradation can allow for proactive maintenance, replacing potentially faulty components before they lead to data loss or system downtime[8]. By tracking the frequency of error detections, it is possible to identify when the likelihood of hardware errors may be increasing, and implement preventive measures accordingly.

Effective hardware error detection is not only about identifying flaws, but also determining the minimal replaceable unit (MRU) that needs to be replaced in order to restore reliability [8]. To help ensure that the correct component is replaced and to minimize downtime, the system must accurately diagnose issues, detect errors and determine the corresponding MRU.[8]

Conclusion

In conclusion, cloud computing is a technology that allows for the delivery and access of computing resources, including data storage and processing power, over a network, typically the internet. It aims to reduce complexity for clients by allowing them to virtually store and access data, applications, and technologies remotely, rather than on personal computers or local servers. Furthermore, it was shown that RAS is a set of attributes that must be considered when designing, manufacturing, purchasing, or using a computer product or component. It is important to ensure that cloud computing systems are built with reliability in mind and to consider the potential risks and benefits when deciding to use cloud services. It is necessary for companies to adopt cloud computing in order to remain competitive in the current global market. Moreover, the transition to the cloud will require careful planning and will not occur overnight. As businesses and enterprises consider the best approach for making your

solutions available through the cloud, it is also crucial to ensure that their products, services, and cloud infrastructure are able to scale, are reliable, and are accessible as needed.

References

1. G. Reese, Chapter 1, in: Cloud Application Architectures: Building Applications and Infrastructure in the Cloud, O'Reilly, Beijing, 2010.

2. SAAS vs. paas vs. iaas: Examples & how to tell them apart, BigCommerce. (n.d.). https://www.bigcommerce.com/articles/ecommerce/saas-vs-paas-vs-iaas/ (accessed January 2, 2023).

3. Types of cloud computing, Red Hat - We Make Open Source Technologies for the Enterprise. https://www.redhat.com/en/topics/cloud-computing/public-cloud-vs-private-cloud-and-hybrid-cl oud (accessed January 2, 2023).

4. Cloud computing 101: Scalability, reliability, and availability, Lucidchart. (2020). https://www.lucidchart.com/blog/reliability-availability-in-cloud-computing (accessed January 2, 2023).

5. J. McCrank, NASDAQ says software bug caused trading outage, Reuters. (2013). https://www.reuters.com/article/us-nasdaq-halt-glitch/nasdaq-says-software-bug-caused-trading- outage-idUSBRE97S11420130829 (accessed January 2, 2023).

6. Reliability terminology, Reliability Terminology | Reliability | Quality & Reliability | TI.com. (n.d.). https://www.ti.com/support-quality/reliability/reliability-terminology.html (accessed January 2, 2023).

7. M. Brown, The 4 types of cloud computing services, ExitCertified. https://www.exitcertified.com/blog/4-cloud-computing-services (accessed January 2, 2023).

Reliability, availability and serviceability, Reliability, Availability and Serviceability - The Linux Kernel Documentation. (n.d.). https://docs.kernel.org/admin-guide/ras.html (accessed January 2, 2023).