AN OVERVIEW OF ENVIRONMENTAL SCALABILITY AND SECURITY IN HYBRID CLOUD INFRASTRUCTURE DESIGNS

Research Article

i-PROCLAIM .my

Asia Pac. j. energy environ.

Sandesh Achar

Director of Cloud Engineering, Workday Inc., Pleasanton, California, **USA**

*Email for Correspondence: sandeshachar26@gmail.com

Manuscript Received: 05 June 2021 - Revised: 17 July 2021 - Accepted: 25 July 2021

Abstract

The practice of using IT resources via the internet with pay-as-you-go pricing is known as cloud computing. The cloud computing market has grown dramatically in the last few years. Most companies are transmitting from on-premises data centers to cloud computing. You can use the cloud as an alternative to purchasing and maintaining computer hardware and software. You can spend time, energy, or money doing everything yourself. This paper presents an overview of the architecture, scalability, economics, consistency-availability-partition, transactions, and security of cloud-based system designs. This study assesses the current cloud development ecosystem using various criteria.

Key words

Cloud-based System Design, Cloud Computing, Environmental Scalability, Architecture Design

INTRODUCTION

It is challenging to launch a business that focuses on the technical realm. Companies must purchase servers and other infrastructure through lengthy procurement procedures. Physical space is needed for the acquired systems, often a specific room with enough power and cooling support. Later, they require qualified employees to manage the plans after they have been configured and deployed. When demand increases or a company grows, it is challenging to scale this lengthy process. For example, businesses may purchase more computing resources than necessary, resulting in low utilization rates. By providing computing resources as scalable, on-demand services, cloud computing addresses these problems.

What is cloud computing?

Cloud computing is a practice of delivering IT resources via the internet with a pay-as-you-go pricing model. These IT resources include computing, storage, database, networking, and security tools and resources (Dillon et al., 2010). With cloud-based storage, files can be saved to a remote database instead of being kept in a hard drive or on-premises data center. For example, a computer or other electronic device may access data and the software needed to run it if it can access the internet. Many factors make cloud computing popular for individuals and companies, including high scalability, reduced cost, increased security, performance, and availability.

Companies prefer Cloud Computing due to three primary reasons:

- You can put work into managing or keeping it up without putting work into managing or keeping it up.
- Its size is practically endless, so you don't have to be concerned about it running out of room.
- Any device with an internet connection can access cloud-based applications and services.

TYPES OF CLOUD COMPUTING

In contrast to a microprocessor or a phone, cloud computing is not a single technology. The system is made up of three services: platform-as-a-service (PaaS), infrastructure-as-a-service (IaaS), and software-as-a-service (SaaS) (Achar, 2016).

- **Platform as a Service (PaaS):** Of the three cloud computing levels, PaaS is regarded as being the most sophisticated. While SaaS and PaaS are comparable in some ways, the main distinction is that SaaS is a platform for delivering software online. In contrast, PaaS is a platform for developing software provided online, for example, Heroku, Google App Engine, and Salesforce.
- **Infrastructure as a Service (IaaS):** IaaS refers to a technique for providing anything as part of an on-demand service over IP-based connectivity, including operating systems, servers, and storage. Clients can obtain software and servers through an on-demand, outsourced service rather than having to buy them outright. IaaS providers provide small, medium, big, extra-large, and memory- or compute-optimized instances for different workload requirements. For example, the IaaS cloud model is a remote data center for commercial users. Cloud computing platforms like Microsoft Azure, Amazon Web Service (AWS), and Google Cloud Platform (GCP) are some examples of IaaS.
- **Software as a Service (SaaS):** Software as a service (SaaS) entails granting licenses to users of the software. A pay-as-you-go or on-demand mechanism is generally used to provide appointments. For example, users can use a PC or mobile device with internet connectivity to access SaaS applications from anywhere, Microsoft Office 365.

CLOUD COMPUTING DEPLOYMENT MODELS

There are only a few cloud computing designs that work for everyone. What works for one business may not be appropriate for you and your needs as a business. However, the cloud's flexibility and adaptability enable companies to adjust to shifting markets or measurements quickly.

There are three different deployment models for cloud computing (Savu, 2011).

- **Public clouds:** Third-party cloud service providers run public clouds. They provide computing, security, storage, and network resources via the internet, allowing businesses to use on-demand resources following their needs and professional objectives.
- **Private clouds:** Private clouds, also called "on-premises," are designed and managed by a single company and privately housed in those companies' data centers. They give internal users access to a shared pool of resources while enhancing data control, security, and administration.
- **Hybrid clouds:** Companies can use public cloud services while maintaining the security and compliance features often found in private cloud architectures, thanks to hybrid clouds, which blend public and private cloud models. For example, companies can use the public cloud to accommodate workload surges or demand spikes while running sensitive applications on the private cloud. The objective of a hybrid cloud is to provide a unified, automated, scalable environment that makes the most of public cloud infrastructure while keeping sensitive data under your control.

Using cloud computing for file access is simply the tip of the iceberg. There are numerous other benefits. For example, users can access their files from any device and create backup cloud computing services, ensuring they will always have access. Large firms can save a ton of money using cloud computing. As a result, companies had to invest in costly information management infrastructure, technology purchases, construction, and maintenance before the cloud became a practical substitute (Achar, 2020b).

CLOUD COMPUTING ARCHITECTURE

The cloud environment has two sides. The user interface, or front end, is what the end user can see. The back-end infrastructure operates the cloud, which consists of applications, networks, and services. Middleware is used to connect the front end with the back end.

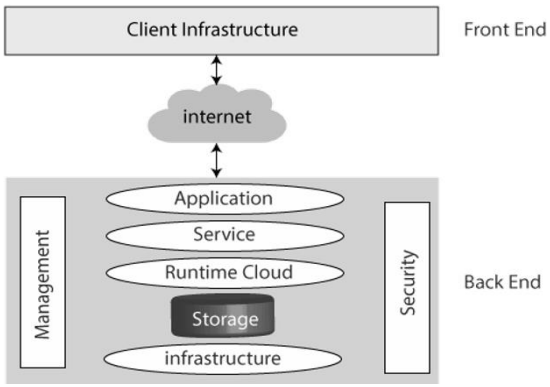

Figure 1: Cloud computing architecture

A cloud-based system is designed using a step-by-step process that begins with the collection of requirements and analysis, continues with the design of the architecture based on that analysis, evaluates improvements, delivers implementation, and culminates in the achievement of continuous operations.

- **Virtualization:** Virtualized servers, storage, and networks are the foundation of cloud computing.
- **Infrastructure:** All the servers, persistent storage, and networking equipment found in conventional data centers are included in cloud infrastructure.
- **Middleware:** These software elements allow networked computers, applications, and software to connect, much like in conventional data centers.
- **Management:** These technologies enable continuous monitoring of a cloud environment's efficiency and capacity. IT teams can monitor usage from a single console, roll out new apps, integrate data, and ensure disaster recovery.

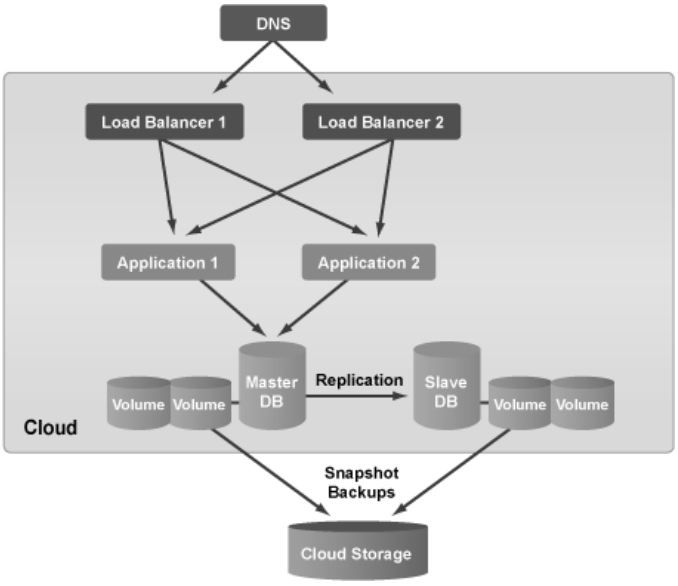

Figure 2: Redundant tier 3 cloud architecture

ENVIRONMENTAL SCALABILITY

Environmental scalability in cloud computing is a way of increasing or decreasing the use of cloud resources per environmental business's needs. Simply, it refers to a system's capacity to support heavier or lighter loads. For example, networking, storage, and computing resources may be scaled to meet fluctuating demand without interfering with industrial operations.

Types of scaling

Cloud scalability may be divided into two categories: horizontal and vertical. The decision between these two strategies should be based on the organization's and the product's present and future demands (Chieu et al., 2009).

1. **Vertical scaling (scale up and down):** Vertical scaling refers to increasing or decreasing the amount of power in a running instance. It focuses on enhancing memory, storage, or processor power to handle rising workloads. This strategy doesn't involve changing the code in any way. For example, the product's performance might be impacted, or there could be downtime. Vertical scaling enables more effective resource optimization, which can reduce cloud expenses.

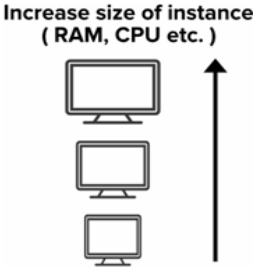

Figure 3: Vertical scaling in cloud computing

2. **Horizontal scaling (scale in and out):** It entails adding or removing servers to the system. By spreading the load across additional servers, we may improve the availability of an application. Horizontal scaling is simpler to complete without the downtime and is easier to maintain automatically. However, due to more instances, this approach can also guarantee improved performance during catastrophic technical breakdowns or natural disasters.

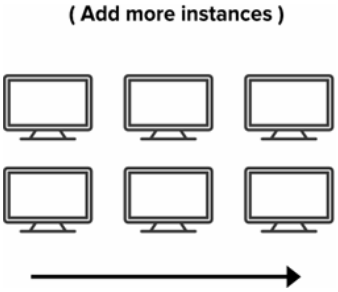

Figure 4: Horizontal scaling in cloud computing

These scaling options can be mixed if necessary and do not conflict with one another. Organizations can, for instance, grow up vertically until the server limit is met, then clone the server to add additional resources as needed. This version might be a smart choice for companies operating in more unstable situations since it enables you to maintain greater agility by scaling up and down as well as in and out.

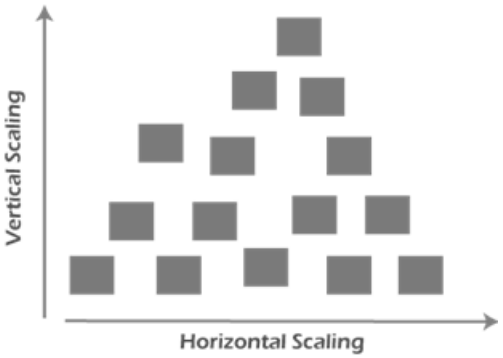

Figure 5: Diagonal scaling in cloud computing

ECONOMY

A continual evaluation of cloud expenses is a process called cloud cost optimization. Identifying and optimizing the weak points of architecture are prime goals of cloud cost optimization. About 35% of cloud spending is wasted, AWS acknowledged in 2017, costing them $6.4 billion in revenue. Therefore, it is essential to make cloud-based system designs economically optimized (Li et al., 2009). Controlling costs is essential and should be a continuous endeavor, whether you choose a fixed pricing strategy or a dynamic pricing approach.

- Identify and shut down unused resources: Searching for underutilized resources is the most straightforward approach to reducing cloud expenditures. It frequently happens that a developer would "spin up" a temporary server to carry out a task but neglect to shut it down after the work is done. Another frequent-use scenario is when the administrator needs to remember to detach the storage from the instances they terminate.
- As a result, costs for resources that a company had bought but is no longer utilizing will appear on its cloud invoices. Therefore, a method for cloud cost minimization should begin by locating and deleting any resources that are entirely unconnected and not in use.
- Use spot and reserved instances: Reserved instances enable organizations to save money on cloud resources by committing up front to using a specific capacity over a one- to three-year period. Comparing reserved instances to on-demand instances can result in savings of up to 80%, depending on other factors. This price option is available with Google Cloud Committed Use, Azure Reserved VM Instances, and Amazon EC2 Reserved Instances (RIs). Workloads with consistent, predictable usage are best suited for reserved instances. Spot Instances can reduce your cloud bill considerably. Spot Instances are up for auction, and if the price is reasonable, they can be bought for use immediately away. Spot instance purchasing possibilities, however, can vanish fast. Because of this, they work best in specific computing scenarios like batch jobs and jobs that can be swiftly canceled.

- Right-size capacity planning: While businesses can swiftly spin up and down cloud instances, they frequently pay for underutilized capacity. To handle unforeseen traffic surges and load swings, organizations must ensure adequate power, but only a little that they overspend on unneeded resources. The best part of capacity planning is selecting the appropriate cloud resources for your system. To fulfill each application's distinct processing, memory, storage, and performance requirements, compute instances are offered in various configurations. Given the variety of options, it's simple to oversize an example by giving it much more processing power, memory, and storage than the application requires. If those extra resources are employed, money is well-spent every month.
- Consider multi-cloud strategy vs. single cloud: Some businesses purposefully look for multi-cloud solutions to prevent vendor lock-in (Simarro et al., 2011). The availability and uptime can be increased with this technique, but these companies run the risk of losing any potential bulk cost reductions offered by a single cloud vendor. A corporation might be unable to achieve the $1 million tier with one vendor, for instance, if they spend $500,000 on AWS + $500,000 on Azure Platform. The benefits of that $1 million tier could include significant discounts on total cloud spending and special treatment from that specific vendor. However, a multi-cloud strategy may not be cost-effective due to the administrative burdens of moving between platforms, paying for network traffic across clouds, and requiring people to receive training on different clouds.

AVAILABILITY

The proportion of time that an infrastructure or system is available for its intended purpose while operating is typically referred to as availability. In cloud infrastructure, availability refers to the fraction of the service's paid duration during which the data center is reachable or provides the intended IT service.

The formula used to calculate availability is -

Percentage of availability = total elapsed time - the sum of downtime/ total elapsed time

The figures give a clear picture of the system availability, enabling businesses to know how much service uptime they can anticipate from IT service providers.

Availability %	Downtime per year	Downtime per month*	Downtime per week
90% ('one nine')	36.5 days	72 hours	16.8 hours
99% ('two nines')	3.65 days	7.20 hours	1.68 hours
99.5%	1.83 days	3.60 hours	50.4 minutes
99.9% ('three nines')	8.76 hours	43.8 minutes	10.1 minutes
99.95%	4.38 hours	21.56 minutes	5.04 minutes
99.99% ('four nines')	52.56 minutes	4.32 minutes	1.01 minutes
99.999% ('five nines')	5.26 minutes	25.9 seconds	6.05 seconds
99.9999% ('six nines')	31.5 seconds	2.59 seconds	0.605 seconds
99.99999% ('seven nines')	3.15 seconds	0.259 seconds	0.0605 seconds

Fig. 6 Availability and expected downtime comparison

The most crucial step in obtaining high availability is selecting cloud vendors that promise high availability. For example, cloud providers like AWS, Azure, and GCP keep their services at 99% availability or more elevated.

- Geographic redundancy: Geographic redundancy is a strong defense against service failure when facing catastrophic events like natural disasters. Geo-redundancy is accomplished by deploying numerous servers at geographically distant places, much like geo-replication. The goal is to select areas that are widely dispersed globally and not overly concentrated in one spot. To ensure that even if one of these remote locations fails, the others continue to function normally, you must execute independent application stacks across each.
- Load balancing: By using load balancing, you can improve the availability of your system. Instances are smoothly replaced in the event of a server failure, and traffic is immediately diverted to working servers. High availability and gradual scalability are made possible by load balancing.

- Failover solutions: Usually, a high-availability design consists of several loosely coupled servers with failover capabilities. When a primary system component goes down due to an unexpected failure or planned downtime, the functions of the secondary system immediately take over. This is known as failover. As a result, you may manage your failover solutions with DNS's assistance in a regulated setting.

PARTITION

A table, index, or index-organized table can be broken into smaller chunks using the partitioning technique, where each such database object's fragment is referred to as a partition. Data partitioning separates the data set and disperses it across various servers or shards. Each shard functions as a separate database, and all the fragments combined comprise a single database. The portioning aids scalability, manageability, performance, security, and high availability. Provider level, region level, zone level, or rack level are the possible partition levels in cloud-based systems.

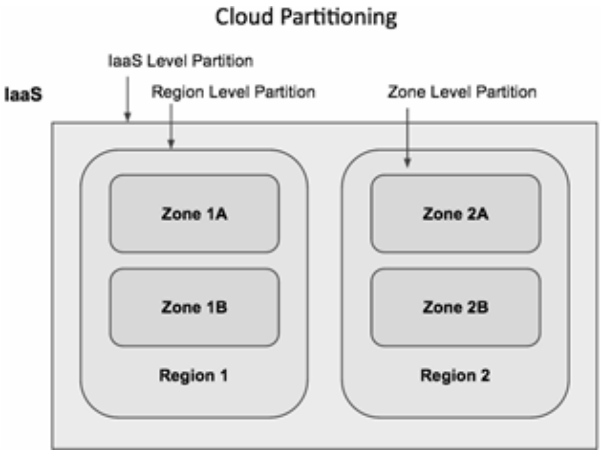

Figure 6: Cloud partitioning example

A section of an IaaS connected to a region or one network of an IaaS is known as a partition group, also known as a network partition. Each province will draw the boundaries of a distinct geographic area. Each area may have several remote, unconnected places known as zones. As a result, a network partition can contain several sections. For the goal of high availability, you can define numerous partition groups with various divisions if necessary.

TRANSACTION

A series of operations performed to complete a set of logical tasks is referred to as a transaction. For example, a transaction often indicates that the data in the database has changed. Therefore, one of the primary purposes of a DBMS is to safeguard user data from system failures. The main advantage of employing transactions is data integrity. To preserve consistency in data collection, the use of various databases necessitates storing data in either multiple tables or rows of a single table. Using transactions, it would be possible to guarantee that other connections to a similar database would either see all the updates or none.

SECURITY

The practice of safeguarding cloud computing infrastructures, applications, and data is referred to as cloud computing security (Achar, 2020c). Cloud security aims to protect cloud environments from dangers such as malware, hackers, DDOS attacks, illegal use, and access.

Challenges in the security of cloud-based systems

Due to the rise in cyber-attacks, saving your data from hackers and malicious attacks has become crucial. Here are some major challenges faced in the security of cloud-based systems (Shahzad, 2014).

- Multitenancy: Multiple client infrastructures are housed under one roof in public cloud settings; therefore, your hosted services could be penetrated by hostile attackers as collateral damage when they target other companies.
- Lack of visibility: It might be easy to lose track of who is viewing your data because many cloud services are accessible outside your company network and through third-party applications.
- Compliance: For businesses employing public or hybrid cloud installations, regulatory compliance management is sometimes a cause of complexity. The company is still ultimately responsible for data privacy and security and relying heavily on third-party solutions to manage this might result in expensive compliance problems.

- Misconfigurations: In 2019, 86% of data breaches involved misconfigured assets, making the accidental insider a significant problem for cloud computing settings. Misconfigurations might occur when the proper privacy settings are not created, or administrative passwords are left in place.

Cloud-based system security best practices

Cloud vendors host cloud computing resources on their servers (Achar, 2020a). They utilize cloud security techniques to keep customer data secret and securely stored because their business depends on consumer confidence. Cloud security, however, also partially depends on the client. Therefore, it is essential to comprehend both aspects for a successful cloud security solution.

The technologies, protocols, and best practices that safeguard cloud computing environments, cloud-based applications, and cloud-stored data collectively constitute cloud security. Understanding precisely what must be secured and the system components that must be handled is the first step in obtaining cloud resources.

- Multiple authentications: First, you must ensure that your cloud vendor offers effective authentication mechanisms, such as multi-factor authentication (MFA) and robust password management, to guarantee secure access to resources. Additionally, the Cloud vendor should allow single sign-on and MFA for internal and external users.
- Data encryption: Ensure that all your data in the cloud is encrypted in both transmit and rest states. A symmetric key is used to encrypt data at rest when it is written to storage. Using Transport Layer Security and a secure channel, data is encrypted while it travels through wired or wireless networks.
- Comply: Ensure that services offered by your cloud service provider are independently reviewed and verified to satisfy international requirements. A cloud vendor who strongly emphasizes ongoing compliance will safeguard your business from legal issues and ensure you're utilizing the most recent security procedures. Observe local or sectoral laws and regulations, including the GDPR, CCPA, FINRA, HIPAA, PCI, GxP, and FedRAMP.

Shift Security to left

- Shifting security left is driven through a few significant components in the Secure Software Development Lifecycle (SSDLC), like Threat Modelling and Accelerated Technical Security Review, New developer-focused Security tooling and automation, Stronger security patterns and frameworks.
- The Accelerated Technical security review involves principles like detecting security design flaws early and collaboration between Security and Development groups.

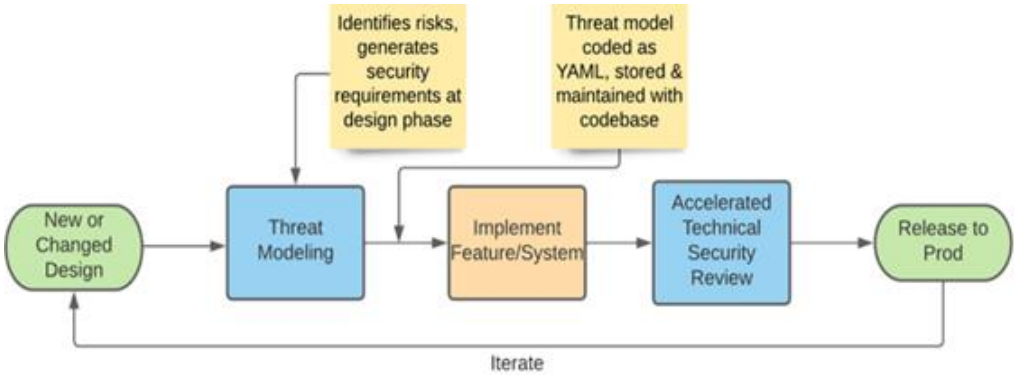

Figure 7: Shift Left Process Diagram

- Adopting a DevSecOps approach in SSDLC to promote continuous security baked in from the start rather than inspected at the end and describing security requirements driven by a practical understanding of risks to the Cloud-based system and the controls that counter those risks.
- Defense in depth by adopting an access control list, network firewall, allowing listing mechanism, and application-specific inspections into cloud-based architecture is a must.

CONCLUSION

Creating a cloud-ready application architecture necessitates paying attention to all aspects, from scalability and availability to security. Whether you work with public, private, or hybrid cloud does not matter. These concepts will help design a secure, highly available, and easy-to-scale cloud-based system.

REFERENCES

Achar, S. (2016). Software as a Service (SaaS) as Cloud Computing: Security and Risk vs. Technological Complexity. *Engineering International*, *4*(2), 79-88. https://doi.org/10.18034/ei.v4i2.633

Achar, S. (2020a). Cloud and HPC Headway for Next-Generation Management of Projects and Technologies. *Asian Business Review*, *10*(3), 187-192. https://doi.org/10.18034/abr.v10i3.637

Achar, S. (2020b). Influence of IoT Technology on Environmental Monitoring. *Asia Pacific Journal of Energy and Environment*, *7*(2), 87-92. https://doi.org/10.18034/apjee.v7i2.649

Achar, S. (2020c). Maximizing the Potential of Artificial Intelligence to Perform Evaluations in Ungauged Washbowls. *Engineering International*, *8*(2), 159-164. https://doi.org/10.18034/ei.v8i2.636

Chieu, T. C., Mohindra, A., Karve, A. A., and Segal, A. (2009). Dynamic Scaling of Web Applications in a Virtualized Cloud Computing Environment. 2009 IEEE International Conference on e-Business Engineering, 281-286, https://doi.org/10.1109/ICEBE.2009.45

Dillon, T., Wu, C., and Chang, E. (2010). *Cloud Computing: Issues and Challenges*. 2010 24th IEEE International Conference on Advanced Information Networking and Applications, 27-33, https://doi.org/10.1109/AINA.2010.187

Li, X., Li, Y., Liu, T., Qiu, J., and Wang, F. (2009). The Method and Tool of Cost Analysis for Cloud Computing. 2009 IEEE International Conference on Cloud Computing, 93-100, https://doi.org/10.1109/CLOUD.2009.84

Savu, L. (2011). *Cloud Computing: Deployment Models, Delivery Models, Risks and Research Challenges*. 2011 International Conference on Computer and Management (CAMAN). 1-4. https://doi.org/10.1109/CAMAN.2011.5778816

Shahzad, F. (2014). State-of-the-art Survey on Cloud Computing Security Challenges, Approaches, and Solutions. Procedia Computer Science, 37, 357-362. https://doi.org/10.1016/j.procs.2014.08.053

Simarro, J. L. L., Moreno-Vozmediano, R., Montero, R. S., and Llorente, I. M. (2011). Dynamic placement of virtual machines for cost optimization in multi-cloud environments. 2011 International Conference on High Performance Computing & Simulation, 1-7, https://doi.org/10.1109/HPCSim.2011.5999800

--0--