Available Online at http://www.journalajst.com

ASIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ISSN: 0976-3376

Asian Journal of Science and Technology
Vol. 16, Issue, 11, pp. 13981-13985, November, 2025

RESEARCH ARTICLE

REAL TIME DYNAMIC ALLOCATION OF VIRTUAL MACHINES IN CLOUD COMPUTING

Onyonkiton Theophile ABALLO, Aziz SAIBOU, Amadou T. SANDA MAHAMA, Arsène Narcisse DAGBA and Taohidi Alamou LAMIDI

LETIA/EPAC/UAC, University of Abomey-Calavi, Abomey-Calavi Republic of BENIN

ARTICLE INFO

Article History:
Received 19th August, 2025
Received in revised form
20th September, 2025
Accepted 15th October, 2025
Published online 30th November, 2025

Key words:
Real time Allocation, Virtual Machines, Data Center Resources, Cloud Computing.

***Corresponding author:**
Onyonkiton Theophile ABALLO

ABSTRACT

Information technology is increasingly evolving towards a model of standardized services provided in a way similar to traditional public utilities such as water, electricity and gas. Within this paradigm, a considerable number of users access services according to their needs, regardless of where they are hosted or how they are delivered. Consequently, effective resource management becomes crucial. Despite the number of virtual machine management algorithms, the proposal of new techniques is required in order to tackle issues related to the optimal use of resources. Our proposal is based on the dynamic allocation of virtual machines (VMs) to meet the needs of users. In this work, we propose an algorithm for creating, adjusting and removing VMs based on requests, while ensuring optimal use of data center resources. Our reel-time algorithm adjusts resources leading to cost reduction and therefore performance improvement.

Citation: ***Onyonkiton Theophile ABALLO, Aziz SAIBOU, Amadou T. SANDA MAHAMA, Arsène Narcisse DAGBA and Taohidi Alamou LAMIDI.*** **2025.** "Real Time Dynamic allocation of virtual machines in cloud computing.", *Asian Journal of Science and Technology*, 16, (11), xxx-xxx.

INTRODUCTION

Cloud computing has emerged as a paradigm that enables the reclamation of unused resources, the creation of shared resource pools, convenient access for users, and improved overall efficiency in resource utilization. In the current context where cloud services play a key role in the industry.The needs of service users vary in terms of computational demand, and many data is generated through online services. Cloud technology therefore appears as a solution to address the challenges posed by massive data generation. Nevertheless, resource management remains crucial.according to the definition of the National Institute of Standards and Technology (NIST), cloud computing is defined as on-demand network access, via telecommunications, to a shared pool of configurable computing resources, provided in a self-service manner. It is therefore a model for abstracting and virtualizing IT infrastructure (Casavant, 1994). The present work proposes a real-time dynamic allocation of virtual machines (VMs) to meets the needs of differents users. The main objective is to maximize efficiency by adjusting resources in real time, while reducing costs and optimizing performance. To structure this work, we review several concepts and approaches proposed in the scientific literature, highlight the contribution of the present study, and present the experiments and results.

STATE OF ART

The management of queues is a fundamental area of operational research and IT, aimed at modeling, analysing and optimizing the processing of query flows in environments characterized by random demand and limited resources.

Management of queues: The management of queues is a set of techniques and practices designed to organize, regulate and optimize customer waiting times or tasks in systems where resources are limited compared with demand. It applies in a variety of contexts, in business, in public services, health care, and in computer systems.

Queues

Queues can be organized in different ways

- Also known as First Come First Served (FCFS), the FIFO principle is to process applications in the exact order of arrival. It is often considered to be the fairest mode of service for all queue clients.

Also known as Last Come First Served (LCFS), the LIFO principle is to treat the latest arrival request as a priority. It is often compared to a warehouse where the items are stacked: it is easier to drop and remove those above the pile (https://dspace.mit.edu/ bitstream/handle/ 1721.1/92225/897471271-MIT.pdf?sequence=2&isAllowed=y).

Naor Model: The Naor model is a theoretical model used in computer science to optimize resource management, especially in systems where the demand for resources fluctuates. This model is based particularly on the balance between the cost of waiting for users and the cost of operating the service provider, in order to maximize overall efficiency.

Some assumptions of the Naor model are (Naor, 1969**):** single and constant queue: Naor's model generally considers a single queue of

clients waiting for access to a resource, with a variable but unified length;

- **Impatient customers:** In Naor's model, **customers may abandon the queue if waiting time is excessive, creating pressure to minimize delays.**
- Waiting costs for users: **Waiting costs for users:** The model includes a waiting cost borne by users, proportional to their waiting time before accessing resources;
- **Operating cost for service provider:** The service provider bears costs related to resource **provision** and management (such as infrastructure, energy, etc.). This cost increases with resource capacity, creating a trade-off between service level and operating cost;
- **Optimization of efficiency:** The model seeks to optimize the system by balancing user waiting costs with resource operation costs in order to minimize overall costs or maximize efficiency;
- **Arrival of clients according to a Poisson process:** The Naor model considers that customers arrive according to a Poisson process, i.e., arrivals are independent and follow a certain probability. This assumptions simplifies the calculation of wait times and the modelling of customer flow:
- **Exponential service time:** Finally, the model assumes that service time (the time a resource is occupied by a client) follows an exponential distribution. This means that the service is random, but with a known average, thus facilitating calculations.

Principle of selecting the overloaded server: The principle of selecting the overloaded server, also known as the super-node concept, proposed by researchers Lo and al. (2005) is based on a load balancing approach in a decentralized network. The idea is to select certain nodes **according to their capacity to handle** a large volume of data and simultaneous connections, and then use them to distribute the load among the different connected users or devices.

This approach enables

- **Identifying** nodes with **high capacities, called super-nodes,** in terms of processing power, memory and bandwidth.
- **Directing** traffic to these nodes to ease the load on other standard servers or nodes.
- **Ensuring** efficient network operation by allocating load, reducing latency and increasing reliability.

Cloud concepts

Definition: Cloud Computing is an approach that provides applications and computing resources as on-demand utilities over the Internet. It allows us to create, configure and customize online applications (https://www.tutorialspoint. com/cloud_computing/ cloud_computing_overview.htm). As defined by the National Institute of Standards and Technology (NIST), cloud computing on-demand network access, via telecommunications, to a shared pool of configurable computing resources, provided in a self-service manner. It is therefore a virtualization and abstraction of the Information Technologie infrastructure (https://fr.wikipedia.org/ wiki/ Cloud_computing). It refers the set of computing services (such as storage, processing, and data management) delivered remotely through the Internet. Instead of storing and managing data on local servers or personal computers, these services use remote servers located in specialized data centers.

Cloud computing deployment models: Cloud computing deployment models define how cloud services are accessible and used by organizations and individuals. They determine the type of cloud configuration best suited to specific requirements in terms of security, control, cost, and flexibility. Key cloud deployment models include (http://blog.3li.com/cloud-les-modeles-de-deploiement/):

- **Public Cloud:** Public Cloud: Offered by external providers such as Amazon Web Services (AWS), Microsoft Azure, or Google Cloud Platform, and made available to multiple customers over the Internet.
- **Private Cloud:** Infrastructure dedicated to a single organisation, offering enhanced security and control.
- **Hybrid cloud:** Combines public and private clouds, allowing flexibility by using public resources for variable workloads while retaining sensitive or critical data in the private cloud.

Cloud computing architecture

Cloud computing architecture consists of several layers and components that together provide IT services via the Internet. Its main components are as follows:

- Front-end (Customer Interface): Represents the visible part of the architecture through which users interact, typically via an application or a web browser.
- Backend (Cloud infrastructure): Includes servers, storage, databases, and services responsible for processing user requests and managing data storage.
- Networks: Ensure connectivity between customers and cloud resources, while guaranteeing secure and efficient data transmission
- Data storage: Provides storage for files, databases, and backups, with cloud systems designed to be both scalable and reliable.
- Management and security: Includes governance policies, monitoring tools, user management, and data security.

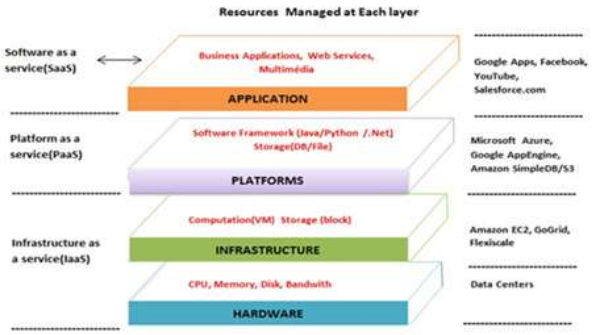

Figure 1. Computer Cloud Architecture Drawn from Zhang, Cheng and Boutaba (2010) (Zhang, 2010)

Virtualization in cloud computing: Virtualization is a key technology in cloud computing, which enables the creation of virtual versions of physical resources, such as servers, storage, or networks. It enables cloud providers to partition physical infrastructure into multiple independent virtual machines (VMs) that share the same underlying hardware (9).

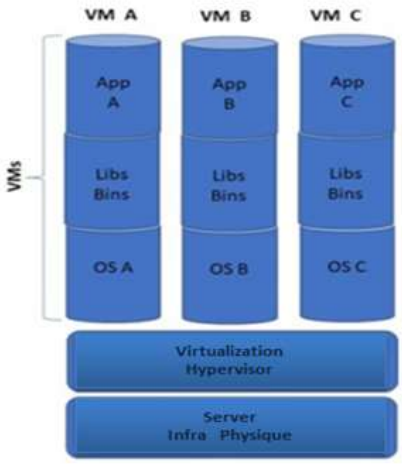

Fig. 1. Structure of a virtual machine

Hypervisor: Virtualization is based on the use of a hypervisor a software (or firmware) layer that creates and manages multiple virtual machines on a single physical host. There are two main categories:

- hyperisor type 1 (bar-metal): installed directly on the physical equipment. It offers high performance since no host operating system is required. Examples:Xen, VMware ESX and Proxmox.
- hypervisor type 2 (hosted): runs on top of a host operating system, convenient for testing or development but less efficient in production. Examples: VMware Player, VMware Workstation, Virtual Box (https://www.it-connect.fr/les-types-dhyperviseurs/).

Cloud computing concepts related to resource allocation

Invoicing resources for use

Pay-per-use invoicing (also known as utility or consumption-based billing) is a pricing model in which customers are charged according to their actual resource usage. This approach is increasingly adopted in information technology, telecommunications, and public services.

Scalability

Scalability refers to the ability of a system to manage an increase in workload or data volume by adding additional resources (Zhang, 2010). There are two main types of scalability:

- Vertical scalability: Adding resources (CPU, RAM, storage) to a single server. For example, increasing memory if an application requires more capacity.
- Horizontal scalability: Adding more servers or instances to the system. Although more complex to manage, it increases capacity without being limited by a single machine (https://theses.hal.science/tel-01136131/).

Elasticitis: Elasticity refers to the ability of a system to adapt automatically to changes in workload by increasing or reducing resources as needed. This optimizes costs by ensuring resources are allocated only when required. Key features include:

- Active provision: Resources are provisioned or released automatically according to peaks or drops in demand.
- Resource management: Elastic systems continuously monitor resource usage and adjust allocations in real time to maintain performance (https://theses.hal.science/tel-01136131/).

Distribution of costs

- The distribution of workload, or load balancing, is a technique used to efficiently spread tasks or requests across multiple servers, networks, or storage systems. Its main objectives are to optimize resource usage, improve response times, and prevent overload of a single component. We distinguish (https://aws.amazon.com/fr/compare/the-difference-between-type-1-and-type-2-hypervisors/) :
- hardware Load Balancing: Uses physical devices to distribute traffic.
- software Load Balancing: Uses software solutions to manage traffic across servers or applications.
- Some current distributions are:
- round Robin: Requests are distributed circularly between servers.
- least Connections: The request is sent to the server with the least active connections.
- IP Hashing: The distribution is based on the IP address of the request, ensuring that the same address is always directed to the same server.
- sticky Sessions : Sessions are kept on the same server to ensure continuity of data

Related work: Researchers have contributed to the improvement of resource allocation in the cloud. Zhen Xiao and al in have solved on the one hand the problem of servers that are turned on but little used in a cloud data center, which wastes energy and costs. And on the other hand, the massive consolidation of virtual machines on the same physical server, which could cause a decrease in performance. In fact, these authors introduced the concept of skewness to balance multi-resource use (CPU, memory, network) and developed a comprehensive automatic resource management system based on virtual machine migration. They also developed an efficient and simple prediction algorithm (FUSD), which reduces overload without too much migration (https://www.researchgate.net/ publication/ 260357894_Dynamic_Resource_Allocation_Using_Virtual_Machine s_for_Cloud_Computing_Environment). Etienne Michon focuses on dynamic resource allocation in the context of cloud computing, particularly on infrastructure as a service (IaaS). Its theme, which was supported in 2015, focused on optimizing the leasing of IT resources to maximize efficiency while at the same time reducing costs. It has developed a brokerage system capable of automating resource provisioning according to user-defined strategies and simulating execution to estimate costs and time requirements. This system is adaptable to various cloud providers and provisioning strategies, and has been tested on a large scale on several platforms (https://theses.hal.science/tel-01290235). Concerning Ashwini E in his paper (https://www.worldwidejournals.com/global-journal- for-research-analysis-GJRA/article/dynamic- allocation-of-virtual-machines -in-cloud-computing/NDgzNg==/?is=1) focuses on dynamic task allocation in a cloud environment. It offers a dispatch algorithm that optimizes the use of virtual machines according to their load, with concrete results on various tasks. The approach is pragmatic and focused on real-time performance. S.Jason in (https://www.researchgate.net/publication/371963527_Modeling_and _Simulation_of_Real-Time_Virtual_Machine_ Allocation_in_a_Cloud_Data_Center) addressed the issues of conventional planning methods often based on a single parameter (CPU) and neglecting memory and bandwidth, creating load imbalances (hotspots) and under-utilization of resources.As well as insufficient simulation tools adapted to the IaaS level to test real-time distribution and consolidation strategies. This author designed a light and extensible simulator (CloudSched) for the evaluation of IaaS allocation strategies and also introduced integrated multidimensional equilibrium measures (CPU, memory, bandwidth). Experimental results show that its CloudSched simualizer is effective in reducing imbalances, improving energy consumption, and increasing data centre efficiency. Chunhua Lin and al in (https://www.researchgate.net/publication/369616001_Dynamic_syst em_allocation_and_application_of_cloud_computing_virtual_resourc es_based_on_system_architecture) have also worked on dynamic allocation of virtual resources in cloud computing environments. The rationale for the topic is that cloud computing, as a method of system development, allows a large number of systems to be combined to provide services.

The problem is the need to manage a load balance between the different virtual resources to optimize their use and minimize waste. Scientific originality lies in the design of a system that integrates real-time monitoring of the cluster network topology for more efficient load balancing. Researchers P. R. Patil and al in (https://www.researchgate.net/publication/392843154_Analysis_of_V irtual_Machine_Allocation_Strategies_and_Development_of_a_Nove l_Policy) analyzed existing virtual machine allocation strategies. It is a static allocation that is simple to deploy but ineffective in the face of variations in load and dynamic allocation, a strategy that is more flexible but costly in latency and resources. These authors proposed a new smart allocation policy based on machine learning for workload prediction. They suggested a dynamic real-time allocation based on anticipated needs.J. Hu et al. (2010) in Scheduling Strategy on Virtual Machine Resource Load Balancing proposed a scheduling strategy on resource load balancing that considers old data collected versus current data. For these authors, it is a strategy that reduces migration by using a genetic algorithm. As a result, the problem of high migration cost and load imbalance is solved.

OUR PROPOSED ALGORITHM

The main objective is to maximize efficiency by adjusting resources in real time, reducing costs and optimizing migration and performance. Our solution allows you to create virtual machines on another server physical cloud environment in case of overuse rate processor (80%) of the physical machine performing ongoing tasks. On the other hand ,if the processor usage rate of a server is less than 20%, the active virtual machines are migrated to servers whose rate of use processor less than 80%. And finally it allows to remove immediately the iInactive virtual machines.Once the server is released,He is arrested or placed on standby.

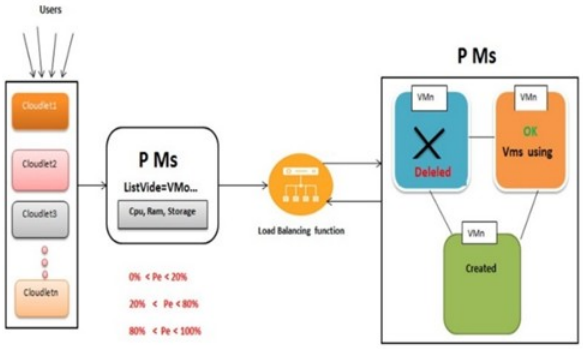

Fig. 3. Bloc diagram of our proposal

Modeling the problem

Basic Hypothesis

Creating VMs (overloading): When a physical server reaches a CPU usage rate greater than 80%, it is considered overloaded.

To avoid performance degradation, any new virtual machine is not created on this server, but allocated to another less loaded physical server. The host server is selected according to an allocation policy which may be:

The least loaded server in CPU, Or the one that minimizes the overall imbalance between resources (CPU, memory, bandwidth).

Under-use case (<20% CPU): When a PMk physical server has a CPU usage of less than 20%, it is considered in under-use status (coldspot). There are two situations:

Active virtual machines (active VMs). These VMs still host applications or processes running. They must be migrated to other less loaded physical servers in order to fully release PMk. If all other servers are already close to 80% (hotspot). We do not migrate immediately, because it would cause overload. Two possible solutions:

Turn on a new physical server (horizontal scalability) and transfer active VMs to it. Differentiate migration until an existing server returns below the threshold (e.g. after deleting inactive VMs or ending tasks). Once the migration is successful (and PMk is hosting more VM), the server can be: turned off to save energy, or placed on standby so that it can be restarted quickly. Inactive virtual machines (inactive VMs). A VM is said to be inactive when it does not perform payloads, i.e.:

CPU 0 % for a prolonged period (e.g. 5–10 min inactivity threshold). No incoming query or significant network activity. Memory only occupied by the basic system, without active processes.

Since these VMs do not bring any value to the system, they are deleted automatically.

Rating — sets & parameters

In this part, we defined the paramaters and the differents sets:

- P={1,……P} : all physical servers (PM)
- γ={1,……V} : all virtual machines (VMs) considered at the step of decision
- $C_i > 0$: CPU capacity of server i (standardised units)
- $d_v \geq 0$: VM v CPU request .
- TH_{hot} = 0.8: overload threshold (80%)
- TH_{cold}= 0.2 :under-use threshold (20%)
- $x_{v,i}^{old} \in \{0, 1\}$: placement known before decision
- $y_i^{old} \in \{0, 1\}$: working / stopping condition known before decision (parameter)
- $act_v \in \{0, 1\}$: observed VM activity indicator v (1 = active , 0 = inactive)

Variables of decision

- . $x_{v,i} \in \{0,1\}$:1 if VM v is placed after PM i decision
- $m_v \in \{0,1\}$: 1 if the VM v has been migrated (final placement - initial placement).
- $start_i \in \{0,1\}$: 1 if PM i is started now (it was turned off before).
- $stop_i \in \{0,1\}$: 1 if PM i is stopped now (it was on before).
- $u_i \in (0,1)$: relative CPU use of PM i after decision.
- $y_i \in \{0,1\}$: 1 if PM i is lit after decision

Real time Dynamic adjustment of VMs: The dynamic adjustment of VMs can be modelled as follows:

Placement unit for active VMs

$\forall v \in V : \sum_{i \in P} x_{v,i} = act_v$

No server placement off

$\forall v \in V , \forall i \in P: x_{v,i} \leq y_i$

CPU capacity and strict compliance with threshold 80% (operational rule)

$\forall i \in P : \sum_{v \in V} d_v x_{v,i} \leq TH_{hot} C_i = 0.8 C_i$

(For example, no PM exceeds 80% after reassignment.)

Definition of relative use

$\forall i \in P : u_i = \frac{1}{C_i} \sum_{v \in V} d_v x_{v,i}$

Detection of migration (change of placement)

$\forall v \in V: m_v \geq x_{v,i}^{old} - x_{v,i}$ et $m_v \geq x_{v,i} - x_{v,i}^{old} \quad \forall i.$

If the placement changes, m_v becomes 1.

Starting / stopping PM (links with previous states)

$\forall i \in P: start_i \geq y_i - y_i^{old} \quad , \quad stop_i \geq y_i^{old} - y_i$

Turn off a PM only if it is empty

$\forall i \in P: \sum_{v \in V} x_{v,i} \leq M\, y_i$

with M large (e.g. M=|V|). If the sum is 0, y_i may be 0

Forced migration outside Cold PMs (operational policy)

For each i as que $u_i^{old} < TH_{hot}$ (i.e. PM cold before decision), and for each VM v that was on i and is active ($x_{v,i}^{old} = 1$ et $act_v = 1$)

This forces the relocation of active VMs from Cold PMs.

$x_{v,i} = 1$

Objective function (simple, weighted version)

Minimize a compromise between number of lit PMs, migrations and start-ups:

min $\alpha \sum_{i \in P} y_i + \beta \sum_{v \in V} m_v + \gamma \sum_{i \in P} start_i$,

with α,β,γ>0 . Practical recommendations: α high (encourage consolidation), β medium (avoid unnecessary migration), γ low (start-up cost).

EXPERIMENTS AND RESULTS

Experimental environment

Materials

The algorithm was implemented on a PC machine with the following specifications:

- Processor: Intel (R) Core (TM) i5-2450M CPU @ 2.50 GHz
- RAM: 8 GB
- Operating system: Microsoft Windows 10 Professional 64 bits.

Software

Our test was implemented using the Java programming language, with the CloudSim framework, on the integrated development environment (IDE) Eclipse.

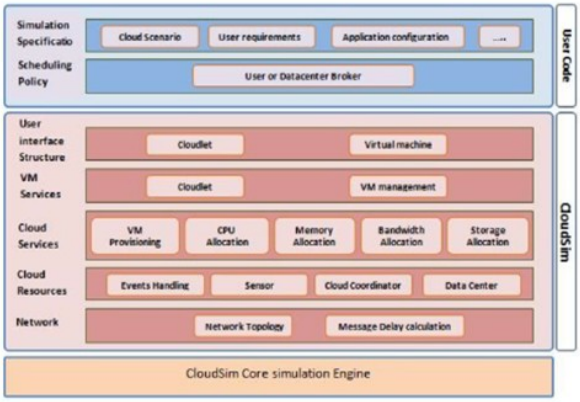

Fig. 4. CloudSim architecture

RESULTS

Evolution of cpu and virtual machines: The figure shows the evolution of the processor utilization rate (blue curve) and the number of active virtual machines (red curve) over time. When the CPU is less than 20%, inactive VMs are removed; Between 20% and 80%, the number of VMs remains stable; and beyond 80%, a new VM is created to absorb overload. This mechanism illustrates the system's ability to automatically adapt resources according to load, ensuring optimized use of virtual machines.

CONCLUSION

In a dynamic cloud environment, the efficient management of virtual machines (VMs) is essential to optimize resources and reduce costs. Allocation strategies dynamic and auto-scaling allow the system to adapt to load variations by automatically adjusting the number of VMs. Management of inactive VMs releases unused resources, maximizing the efficiency of the data centre. The adjustment of VMs to the needs of cloudlets ensures optimal allocation of resources. These techniques allow the system to maintain a balance between performance, availability and cost optimization, while adapting to unforeseen fluctuations.

REFERENCES

Casavant, T. J. Kuhl,A taxonomy of scheduling in general-purpose distributed computing systems, IEEE Transactions on Software Engineering, vol. 14, no. 2, pp. 141-153, 1994.

https://dspace.mit.edu/bitstream/handle/1721.1/92225/897471271-MIT.pdf?sequence=2&isAllowed=y.

P.Naor. «_eregulation of queue size by levying tolls ». In : Econometrica 37.2 (1969), p. 15-24.

Virginia Lo, Dayi Zhou, Yuhong Liu, Chris Gauthierdickey, and Jun Li. Scalable super node selection in peer-to-peer overlay networks. In In Proceedings of the 2nd International Workshop on Hot Topics in Peer-to-Peer Systems, La. IEEE, 2005. (Citen pages 55 et 56.)

https://www.tutorialspoint.com/cloud_computing/cloud_computing_overview.htm

https://fr.wikipedia.org/wiki/Cloud_computing

http://blog.3li.com/cloud-les-modeles-de-deploiement/

Zhang, Q., Cheng, L. et Boutaba, R. 2010. « Cloud computing: state-of-the-art and research challenges ». Journal of Internet Services and Applications, vol. 1, no 1, p. 7-18.

francois rivard,LAVOISIER.14,rue de provigny,94230,ISBN 987-7462-3815-2,ISSN 1635-7361, « cloud computing le systeme d'information sans limite ».

. https://www.it-connect.fr/les-types-dhyperviseurs/

https://theses.hal.science/tel-01136131/

https://aws.amazon.com/fr/compare/the-difference-between-type-1-and-type-2-hypervisors/

https://www.researchgate.net/publication/260357894_Dynamic_Resource_Allocation_Using_Virtual_Machines_for_Cloud_Computing_Environment

https://theses.hal.science/tel-01290235

https://www.worldwidejournals.com/global-journal-for-research-analysis-GJRA/article/dynamic-allocation-of-virtual-machines-in-cloud-computing/NDgzNg==/?is=1

https://www.researchgate.net/publication/371963527_Modeling_and_Simulation_of_Real-Time_Virtual_Machine_Allocation_in_a_Cloud_Data_Center

https://www.researchgate.net/publication/369616001_Dynamic_system_allocation_and_application_of_cloud_computing_virtual_resources_based_on_system_architecture

https://www.researchgate.net/publication/392843154_Analysis_of_Virtual_Machine_Allocation_Strategies_and_Development_of_a_Novel_Policy

Hu J, Gu J, Sun G, Zhao T. A scheduling strategy on load balancing of virtual machine resources in cloud computing environment. 2010 3rd International symposium on parallel architectures, algorithms and programming, IEEE, 2010; 89–96

****