UNIVERSITY OF CRAIOVA

FACULTY OF AUTOMATICS, COMPUTERS AND ELECTRONICS

PhD SCHOOL „CONSTANTIN BELEA”

PhD DOMAIN: COMPUTERS AND INFORMATION TECHNOLOGY

PhD THESIS

SOFTWARE PERFORMANCE OPTIMIZATION IN VIRTUAL COMPUTING

SUPERVISOR

Mihai MOCANU, PhD Prof. Eng.

PhD STUDENT

Cătălina MANCAȘ, Eng.

- 2018 –

CONTENTS

Introduction

Virtualization solutions used within Cloud computing have proven over time to address the problem of inefficient use of physical computing resources. When servers are virtualized, research shows a decrease of resource costs (hardware, air conditioning equipments, hosting space, etc.) as well as lower energy consumption. Virtualization can provide high-level availability to critical applications with, thus streamlining the operation of IT infrastructure and responding quickly to changes. Also, rapid backup in case of error is a crucial benefit. The major disadvantage, however, is the decrease in the performance of virtual machines when compared to physical machines. But how big is the cost/performance ratio? For many organizations, migration to a virtual environment is undoubtedly the first choice. However, an analysis of the virtualization performance is required.

This paper is a scientific approach to unifying methods, models and technologies in a homogeneous unitary methodology, capable of identifying optimized mapping solutions for complex applications on a virtualized, generic, highly parametrized hardware Cloud computing architecture (Cloud systems today are widespread high-performance heterogeneous computing resources), from the pre-design phase of the hardware-software package (co-design). The targeted optimizations aim to use the available resources as efficiently as possible, in order to achieve maximum performance, in restrictive energy consumed conditions, complexity etc., being treated as complex multi-criteria problems.

Cloud computing has been the subject of many research activities which, over the years, have been demonstrating results as: reduced hardware costs, reduced energy consumption, and efficient use of physical machines. Regarding the wider use of Virtualization-based Cloud Services, primarily driven by economic considerations, continuous virtualization research proves to be an useful activity in improving Cloud performance and reducing costs for Cloud service users The results show that many improvements can be made to virtualization-based computing systems and that Cloud computing is not only topical but also future-proof.

Thus, in this paper are studied various techniques and technologies of virtualization, using Cloud solutions and solutions for performanece optimization. The performance of various types of virtual machines has been evaluated by making recommendations based on the specifics of Cloud service users. A comparative analysis of the Private Cloud vs the Public Cloud was also carried out in order to recommend the appropriate solution according to the priority requirements of the users.

Starting from Jackson waiting queue theory, the paper proposes a mathematical model designed to meet the required quality of service (QoS). A policy of allocating virtual machines to physical machines was also proposed for the purpose of reducing electricity consumption and a policy of allocating workloads to available virtual machines to reduce the response time of the system was proposed too. The presented experiments are based on a rigorous analysis of the literature and aim to validate the proposed mathematical models and the proposed planning policies designed to reduce response time and cost by lowering electricity consumption.

Another aspect dealt with in this paper is the evaluation of the use of computing resources in a Cloud-based computing center and the scalability impact on the system's response time. Conclusions resulting from the experiments are in support of users in choosing solutions that meet a certain level of required QoS but with minimal cost.

Cloud performance has a significant impact on the performance of the future computing infrastructure. A thorough evaluation of the performance of Cloud services is crucial and beneficial to service providers as well as consumers, in such manner building an active research area. Some key technologies for Cloud computing, such as virtualization and Service Oriented Architecture (SOA), challenge performance appraisal. So far, the research community has invested substantial efforts to address these challenges, and the result has been remarkable. Among the Cloud performance analysis, assessment approaches developed through system modeling play an important role. However, recent papers concerned specific sections of the literature, leaving no general picture presenting the latest state of the art. The purpose of this paper is to contribute to the creation of the overall image of the current state of performance evaluation in virtualized systems through the results of performance evaluation experiments. This paper also analyzes the open issues and challenges of the assessment approaches made and identifies possible future research opportunities in this important field.

Virtualization has been revolutionized by the emergence of Cloud Computing, primarily through the provision of on-demand services. If a client needs computing power for a limited time, he can buy it for that period in the Cloud without having to acquire the physical resources needed to provide the required computing power. By purchasing Cloud services, the client has access to a wide range of shared resources without having knowledge of their physical location. The resources provided in the Cloud are scalable, immediately. Thus, the client can choose whether to increase or decrease the computing resources on the spot. This option introduces modularity property into the Cloud. In addition, Cloud resources are controllable, measurable, and optimized. Optimizing the use of physical machines is facilitated by virtualization, which has gained popularity with the emergence of the multi-core processor and virtualization support.

Cloud computing has been the subject of many research that has over the years been demonstrating results such as: reducing hardware costs, reducing energy consumption, and making efficient use of physical machines. Specialist literature attests that the adoption of Cloud computing solutions occurs primarily for economic reasons (Wang & Varela, 2011) (Indrani, Sudhakar, & Lizy, 2012), (Kundu, Rangaswami, Dutta, & Zhao, 2010), (Zhang, și alții, 2013). Cloud computing allows customers to reduce their costs of computing and power consumption because customers only pay for the resources they use. Another conclusion that comes out is that many physical machines are used inefficiently because they are insufficiently loaded. Using Cloud computing in conjunction with virtualization is a solution to server downtime, and today's data centers are fast moving towards using server virtualization as the preferred way to share hardware resources between a single server between multiple virtual instances which host different applications (Abirami & Shalini, 2012), (Chang, Tsai, & Chen, 2013). However, virtualization performance in Cloud computing can not equal native performance.

There are several types of Clouds that can be used to meet different needs. Of these, the public Cloud - managed by a service provider accessible to customers via the Internet, and the private Cloud - inside a company's network, are the most popular and serve as the medium for the experiments in this paper.

Cloud computing is a recent technology, but more and more popular since more and more companies are starting to use it. For this reason, an intensive research activity is invested in this direction. Also, the relationship between Cloud computing and virtualization is a current topic of research aimed at improving virtualization performance in Cloud computing (Chang, Tsai, & Chen, 2013), (Diogo & Ferraz, 2012), (Indrani, Sudhakar, & Lizy, 2012). The objective of the field studies, as shown by current research, is to achieve the virtualized performances as close as possible to native performance.

Server virtualization deals with masking the physical resources of the server, including the number and the identity of individual physical servers, processors, and operating systems. The server administrator uses a specific software to divide a physical server into multiple isolated virtual environments. These virtual environments are also known as guests, instances, containers, or emulations. Virtualization is a method by which multiple operating systems can be run on a single physical machine. There are three popular server virtualization approaches: full virtualization, operating system based virtualization, and hardware based virtualization.

Due to the popularity of Cloud computing, many server virtualization solutions have emerged on the market, including VMware, KVM, Microsoft, Xen, or Oracle. These solutions show different performance outcomes, so that a current topic of interest is to identify the recommended hypervisor according to the specifics of the user's applications (Graniszewski & Arciszewski, 2016). This topic served as starting point for some performance analyzes in this paper. Thus, the paper presents a comparative study of the performance of VMware vSphere, Microsoft Hyper-V and KVM, in terms of CPU, memory and I/O usage.

Using virtualization introduces a native performance degradation, affecting the degree of CPU, network, memory, and storage space or I/O devices (Johnsen, Schlatte R., & Tapia, 2012). The purpose of this paper is to evaluate the performance of server virtualization in Cloud computing, and one of the goals of the paper is to recommend the type of virtual machine that affects the least the native performance, depending on the specificity of the user's application. For this, several performance measurement tools have been used with various metrics. Among the tools chosen to measure performance are: UNIXBench (Yamato, 2017), IOzone, (Tarasov, Bhanage, Zadok, & Seltzer, 2011) or RAMSpeed[1].

Any Cloud computing system is built up of two components called Front-End and Back-End. The Front-End is the component accessible to the Cloud service client. This includes the application used to access the Cloud through an interface such as the web browser and the client-accessible network. The Back-End is the back component hidden from the client and includes various servers, storage devices, network equipment, etc. In theory, a Cloud incorporates all types of dedicated servers to host virtually any type of application. A central server manages the system, monitors the system, traffic, and client requests to ensure any required functionality. Additionally, there are sets of rules and middleware that help networking computing resources that require communication between them. The paper proposes a mathematical model for the Back-End of a Cloud system, implemented and tested by discrete event-based simulation. The existence of virtual machines is reflected in Cloud performance, because running virtual machines increases the number of context changes, consuming considerably the power of the processor to translate addresses. The CPU load monitoring aims at improving performance and is a useful action in performance investigations because the CPU load rating has critical implications for other system performance metrics, such as power consumption.

In terms of Cloud Virtualization Management, virtualized servers' hypervisors use various planning policies to manage the efficient use of virtual instances. The performances of the alliance running in a virtual environment are directly influenced by the parameters of the planning policy used. Thus, varying the parameters of the planning policy can affect the performance of the application. For example, the authors of (Chang, Tsai, & Chen, 2013) have demonstrated that I/O performance can be improved by virtual machine task planning. Thus, one of the contributions of this paper is the proposal of a two-tier structured planning policy which incorporates both the virtual machine allocation and physical allocation of virtual machines to workloads. The proposed policy was evaluated by discrete, event-based simulation, and the results were compared to those of a popular planning policy - based on the Round-robin algorithm.

In a real-life scenario, Cloud computing resources are shared among customers who pay for services based on resource usage. Generally, through virtualization, the details of the computing resources become transparent to customers. Observed from the user's perspective, the services are identical in terms of functionality and interface. But the allocation of resources to perform the tasks required by various system users, as well as Cloud scalability, are crucial factors for optimizing system response time (Vinothina, Sridaran, & Ganapathi, 2012), (Falatah & Batarfi, 2014), (Goswami & Saha, 2013). If resource allocation is inefficiently managed, the degree of system service is affected, implicitly by QoS, an important factor in

[1] RAMSpeed, a cache and memory benchmarking tool (http://alasir.com/software/ramspeed/)

evaluating the performance of virtualized systems. The paper studies how response time is influenced by resource allocation by scaling the system.

To evaluate the performance of different virtualization methods from the design phase, there is need for a hardware-software model that allows performance to be estimated, without using the application code and the concrete choices of hardware architecture choice, considered only as a concept (generic, parameterized Cloud architecture). The model must represent the assembly in a uniform manner so that it can be described by associating the competing components of the application with the elementary or aggregate execution nodes at the Cloud components level, and at the same time can highlight the extent to which mapping exploits certain features such as existing hardware parallelism, or restrictions such as jurisdictional.

Synthetic presentation of the chapters of the doctoral thesis

The doctoral thesis is structured in seven chapters. The introductory chapter presents the motivation behind the paper and the state of the art of the domain, the objectives and the structure and organization of the thesis. Starting from the current state of research and the needs of the Cloud computing market, there have been identified the research directions useful in raising the level of knowledge in the field. The proposed objectives are meant to support Cloud services users by facilitating their choice of the service that best suits the specifics of their applications, and on the other hand, proposing solutions designed to contribute to the efficiency of resource utilization and lower costs. The emergence of various Cloud service types on the market: public Cloud, private Cloud, or hybrid Cloud, represents the motivation to support the increasing number of users in choosing the type of service that best suits their needs.

Virtualization has been a subject of constant interest in the research work developed over the last few years (Mancaș C. , 2015), (Mancaș C. , 2014), (Mancaș, Mocanu, & Mancaș, 2013). Research results have shown that the performance of the virtualized environment and the Cloud system depend significantly on the planning of the tasks and of the workload of the system. This argument, along with the desire to find solutions for reducing system response time and electricity consumption, was one of the reasons for pursuing inquiries in this direction. Also, the need for Cloud services to meet a certain level of QoS has represented the motivation behind the research idea of modeling Virtualization-based Cloud Systems in order to deliver the required level of Cloud service quality, at a minimum cost.

This paper aims at evaluating the performance of virtual computing, especially the performance of a Cloud environment, with the aim of proposing new solutions to optimize virtual computing in order to increase performance and reduce costs. The paper evaluates the performance of existing solutions on the market and makes recommendations to Cloud service users to choose the optimal solutions from the point of view of their application specificity, but also proposes new solutions in order to reduce the response time of the system or to reduce electricity consumption. The experiments included in the performance evaluations are performed in both real-world testing and simulation environments.

The main objective of the paper: *Performance optimization in virtual computing* is decomposed into sub-objectives.

1. Suggest recommendations for the use of KVM, VMware and Microsoft Hyper-V hypervisors, depending on the requirements and specifics of the user's applications. For this purpose, the performance of the virtual machines KVM, VMware and Hyper-V was evaluated, using various test scenarios.

2. Formulate recommendations on the type of Cloud service that is recommended to be adopted according to the specific requirements of the users. In order to achieve this goal, a comparative analysis was performed between a public Cloud and a private one regarding the performance of the computing resources: processor, memory and I/O system.

3. Reducing costs by reducing electricity consumption. In this respect, a planning policy for the mapping process of virtual machines to physical machines has been developed and implemented.

4. Reduce the total runtime of Cloud infrastructure applications.

5. Reduce the execution time of parallel workloads to ensure the required QoS level. Thus, a planning policy has been proposed for mapping saricins in virtual machines, in order to reduce the execution time of parallel tasks and increase the overall performance of the Cloud computing system.

6. Improve the performance of a Cloud system by analyzing the impact of computing resources utilization and Cloud infrastructure scalability on the total response time of the system.

The second chapter of the thesis Current Virtualization Technologies and Solutions presents general concepts of virtualization, starting with the first forms of virtualization and providing a retrospective view of the virtualization concept. Chapter 2 subchapters include theoretical notions about: CPU virtualization, notably the virtualization of the x86 processor and its limitations, memory virtualization and I/O virtualization, virtualization techniques, implications of the different virtualization types on performance, virtualization technologies among which of interest is server virtualization and current server virtualization solutions.

The chapter presents the terminology of virtualization, explaining concepts such as physical machine, virtual machine, virtual machine monitor (VMM), server virtualization, or computing resources. Lowering the level of detail, Chapter 2 also discusses the processor virtualization and details how the processor virtualization is performed, with the features and limitations of VMware's x86 processor virtualization. VMware marks the history of virtualization with the launch of the first CPU virtualization solution, but the limitations of the x86 family have spawned several virtualization techniques: full virtulization, paravirtualization, or OS-assisted virtualization, and hardware-assisted virtualization. The processor virtualization introduces a key abstraction level in the virtualization domain, which is called the hypervisor. To run multiple virtual machines on a single physical system, processor virtualization is not enough. Additionally, other levels of virtualization need to be introduced: memory virtualization and virtualization of I/O devices. The key component of memory virtualization is the Memory Management Unit, responsible for mapping physical physical memory to physical memory, while virtualization of I/O devices is made possible through the hypervisor.

The performance of CPU virtualization, memory virtualization, and I/O virtualization is the subject of this study. In the performance evaluation experiments, virtualization of computing resources was performed using virtualization solutions: VMware vSphere, Microsoft Hyper-V and KVM. These solutions were chosen conducting a literature study and a market study that led to the motivation to try new performance analysis experiments. The results of the study and the current situation of the three virtualization solutions are also described in Chapter 2, and the results of this study represent a necessary starting point in the performance assessments presented in Chapter 6 of the thesis.

In Chapter 3 *Cloud Computing: reference models and architectures*, the focus moves from general to specific, from virtualization concept to its component: Cloud computing. In this case too, it is desirable to formulate definitions and review a terminology. The most popular definition of Cloud computing is the standardization of the ISO / IEC 17788: 2014 (ISO-International Standard for Organizations, 2014). Early efforts in standardizing Cloud calculations were, however, carried out by NIST (The National Institute of Standards and Technologies of the US Government, Mell & Grance, 2011). This model described by NIST includes five essential features, three service models, and four models of development. The five features are described in the subchapter Cloud computing Standardization. The service models formulated by NIST are described in the subchapter *Service Models, Functionality Types and Service Categories in Cloud*, which are incorporated in ISO/IEC Standard 17788 : 2014. However, the standard chooses to fit the three models proposed by NIST into another classification called *Cloud Services.* A category may include functionalities of all three types mentioned above, so the standard clearly defines Cloud service categories, functional types Infrastructure as a Service (IaaS), Network as a Service (NaaS), Platform as a Service (PaaS), Software as a Service (SaaS), etc. In addition, new categories of services are expected to be standardized in 2019, currently in the informal category. These informal categories may serve as future research directions.

The subchapter *Cloud Implementation Models* deals with Cloud standardized models, such as: public Cloud, private Cloud, hybrid Cloud, and community Cloud. Of these, performance analysis evaluates the first two models: the public Cloud and the private Cloud. Last but not least, Chapter 3 lists the cross-cutting aspects of standardized Cloud computing, of which performance is the main aspect of this paper. In order to perform a rigorous performance analysis, it is necessary to define the parameters to be measured and to be evaluated.

A complete definition of Cloud computing is reflected by the Cloud Computing Reference Architecture (CCRA) reference architecture described in the *Cloud Computing Reference Architecture* subchapter. This architectural model meets the description of roles and sub-roles, activities, cross-functional aspects and functional components of Cloud computation, as well as the relationships between them. CCRA is a reference tool in the Cloud computing International community, focusing on what Cloud services provide. From an architectural point of view, Cloud systems can be viewed and described from several angles. In this work, the Cloud reference architecture is described from the user's point of view and from the functional point of view.

CCRA has led to the development of architectural models of the two Cloud infrastructures: public and private, used in performance assesment experiments. The two architectural models are presented in the subchapter *Architectural Models of the Infrastructure in the Performance Evaluation Experiments.*

Starting from the architectural model of the Private Cloud, the mathematical model of this infrastructure was designed to perform more experiments. Mathematical modeling is based on Queue Theory Theory and Jackson Open Network Theory. The mathematical foundation required to generate a mathematical model and the proposed models can be consulted in Chapter 4. Mathematical modeling presents a basic version, but also a virtualized extension, both validated by event-based simulation.

Chapter 5 *Preliminaries and Testbed Environments of the Performance Evaluation Experiments* describes the environment in which performance evaluation experiments were conducted with the aim of identifying real scenarios where a Cloud system exhibits high performance results. The most relevant performance indicator is the response time of the system. In order to reduce the total response time, but also the execution time of the parallel tasks, several experiments were carried out. Another goal is to reduce energy consumption. To achieve these objectives, two planning policies have been proposed. The first proposed policy concerns the planning of the virtual machine mapping process in physical machines and aims at reducing the system's energy consumption. This policy is described in the The *High-Level Planning Policy – Virtual Scheduler* subchapter and is evaluated in the *The Virtual Scheduler Performance Evaluation* section. The second policy aims at reducing the total response time of the system, but also the execution time of parallel tasks, and proposes an algorithm of assigning tasks to virtual machines. This algorithm is described in the *The Low Level Scheduling Policy – Job Scheduler* section and the policy is evaluated in the *The Job Scheduler Performance Evaluation* section. The results and conclusions that come out of the evaluation of the two proposed policies can be reviewed in the *Conclusions* section.

Datorită expansiunii adoptării de soluții de vitualizare s-a dovedit necesară o analiză a soluțiilor de virtualizare din categoria liderilor de pe piață, în vederea formulării de recomandări cu privire la performanțele soluțiilor în funcție de specificul aplicației utilizatorului. Astfel, au fost evaluate performanțele hipervizorilor KVM, VMware și Hyper-V, din punct de vedere al comportamentului procesorului, al operațiilor asupra memoriei și al operațiilor asupra sistemului de fișiere. Experimentul a fost realizat într-un Cloud privat, utilizând diferite aplicații open-source pentru măsurarea unei game variate de metrici cum ar fi: rata de Input/Output (I/O), viteză la scriere/citire, latența, creare/ștergere fișiere, viteză de scriere/citire memorie, viteză CPU (Unitatea Aritmetico-Logică și Unitatea în virgulă mobilă) etc. Indicatorii de performanță, uneltele utilizate, precum și rezultatele experimentului sunt descrise în secțiunea ***Error! Reference source not found.***. Rezultatele obținute sunt descrise în secțiunea *REF _Ref523605343 \h * MERGEFORMAT* ***Error! Reference source not found.***, ***Error! Reference source not found.*** și ***Error! Reference source not found.***. Concluziile și recomandările formulate în urma rezultatelor obținute, precum și direcțiile de cercetare ce se desprind în urma studiului de performanță se regăsesc în capitolul **Error! Reference source not found.**, și urmăresc îndeplinirea obiectivelor formulate în capitolul 1.

Due to the expansion of the adoption of virtualization solutions, an analysis of the market leaders virtualization solutions it is compulsory in order to formulate recommendations on the performance of the solutions according to the specifics of the user's application. Thus, the performances of KVM, VMware and Hyper-V hypervisors was evaluated from the point of view of the processor's behavior, memory operations and operations on the file system. The experiment was conducted in a private Cloud using various open-source applications to measure a variety of metrics such as Input/Output (I/O), files Read/Write speed, latency, creation/deletion of files, memory Read/Write speed, CPU speed (Arithmetic Logic Unit and Floating Point Unit), etc. The performance indicators, the tools used and the results of the experiment are described in the subchapter *Hypervisors Performance Evaluation Preliminaries*. The results obtained are described in *Hypervisors Performance Evaluation* subchapter. The conclusions and recommendations of the results obtained, as well as the research directions that come out the performance study, are found in Chapter 7 and aim at achieving the objectives set out in Chapter 1.

Another topic of interest in the paper was to identify the Cloud implementation model (see *Cloud Implementation Models* chapter) recommended to be adopted according to user requirements. In this respect, the experiment has compared the performance obtained in a private Cloud with those obtained in a public Cloud (*Public Cloud vs. Private Cloud Performance Evaluation*). Performance has been evaluated for processor, memory, and I/O behavior, and the public Cloud Provider has been Rackspace. The two infrastructures are described in the introductory section of chapter 5 *Testbed Environments: work infrastructures description.*

The response time of the system is also analyzed from the point of view of resource utilization and of the scalability of data centers. The performance analysis, as well as its results, are described in the section *Resource Allocation and Scalability Performance Evaluation*.

The final chapter, *Conclusions and future research* incorporates the results and the conclusions of all the evaluation experiments described in Chapter 6. These experiments address the objectives identified in the introductory chapter and also serve as starting point for future research directions described in the final chapter.

Thesis contribution and results

A first contribution regards providing recommendations for adopting a virtualization solution. Three of the most popular virtualization solutions have been subject to several performance evaluation experiments, with the purpose of satisfying user's requirements. The emergence of various Cloud models on the market represented the motivation for supporting the growing number of users in choosing the type of service that best suits their needs. Thus, the experiments conducted in both public and private Cloud systems result in recommendations that support users when deciding to turn to Cloud services. Starting from the architectural model of the Private Cloud, the mathematical model of this infrastructure was designed with the purpose of performing more evaluation experiments. Mathematical modeling is based on the Queuing Theory and on Jackson's Open Networks Theory.

The need to extend the functionality of the proposed mathematical models led to the proposal of an event-based planning policy, structured on two levels: the level of physical machines and the level of virtual machines. Tasks are assigned to virtual machines, which in turn are mapped to computing resources (processors, memory, secondary memory, etc.). The purpose of the proposed policy is to map the tasks to processing nodes, according to the nodes' processing power and availability.

In virtualization, performance depends significantly on how the tasks are handled and on the workload of the system. This argument, along with the will to find solutions for reducing the system response time and the electricity consumption, are the research reasons.

For the described performance tests, various test environments based on both hardware and simulation were used. The simulation environments used are SAGEMath, CloudSim and Eucalyptus. From the hardware point of view, the testbed environment is made up of two Cloud infrastructures: a private Cloud infrastructure and a public Rackspace infrastructure.

Evaluation of KVM, VMware and Hyper-V hypervisors in private Cloud infrastructure

The performance of virtualized systems is described by the behavior of virtualized computing resources. In order to study this behavior, it is necessary to measure certain parameters, metrics or indicators. Evaluations were performed on different types of virtual machines: KVM, VMware and Hyper-V in the following configuration: 2 core (4 threads), 8 GB RAM, 1 TB storage capacity. Virtual machines were created on a physical machine in the private Cloud infrastructure, with the following configuration: CPU: 2x Intel XEON e5 2690v2 at 3 GHz, 20 core (40 threads), 256 GB RAM, 18 TB storage capacity. Performance evaluation was performed on a Linux operating system using the UnixBench tool. UnixBench can be used to evaluate performance both when running a single task or multiple tasks, and for parallel processing.

The UnixBench metrics evaluated in the current experiment are:

1. *Dhrystone:* is a metric used to measure and compare performance in non-floating-point applications.
2. *Whetstone:* is a synthetic metric that measures performance in terms of speed and efficiency for floating point operations.
3. *Execl Throughput:* measures the number of "execl" calls per second. The test results lead to the conclusion that although the performance of the three types of virtual machines are close, VMware performs better than Hyper-V and KVM.

UnixBench tests designed to evaluate process manipulation were:

1. *Pipe Throughput:* measures the number of instances per second in which a process can write and read 512 bytes.
2. *Process creation:* measures how many times in a time unit a process can create and terminate sub-processes.
3. *Shell Scripts:* measures the number of instances per minute, in which a process can manipulate a series of eight competing copies of a Shell script that applies a series of transformations over a data file.
4. *System call overtime:* estimates the cost of inputs and outputs to and from the operating system kernel, the cost of making a system call.

The test results reflect that the Hyper-V virtual machine performs slightly weaker than the VMware and KVM competitors. This could be explained by running Hyper-V on a Linux operating system (with drivers).

Following Pipe Troughput, Shell Creation and Shell Scripts test, VMware stands out significantly from its counter-candidates. VMware is followed by KVM, then by Hyper-V. The latter two show similar results in the Create Process. In the behavioral test scenario, VMware is slightly outperforming KVM, but Hyper-V remains the last.

Memory evaluation tests have been designed to measure the speed of reading and writing operations performed on memory. Of interest are the read/write operations of integer data blocks, as well as read/write operations of floating-point data blocks. The results show that in the case of reading from RAM of integer data blocks, the performance of all three virtual machines decreases with the increase in the data block size. Results show similar performance of the three types of virtual machines. However, in the case of small-sizes blocks (16, 256, 1024 bytes), the same classification pattern is observed: VMware is demonstrating the lowest speed, followed by KVM, then by Hyper-V. In case of floating-point data block readings from RAM, Hyper-V again behaves best for small data blocks. When the data block size is more than 65536 bytes, the highest speed is shown by KVM. The test results show that Hyper-V is more powerful than the other two when writing integer data blocks. In this situation, VMware visibly performs the weakest. In case of writing floating point blocks, the evaluation demonstrates similar performance for the three types of virtual machines with a small gain for VMware.

In testing the performance of the file system, the IOzone tool was used. Iozone generates and measures a wide range of file operations, out of which we test: writing, overwriting, reading, random read, random write, reverse reading (reads the file from the beginning to the beginning). In case of file writing, KVM behaves generally better than the VMware and Hyper-V, but Hyper-V shows best results for larger size files. Overall, the trend is that the performance of all virtual machines drops with the increase in file size. Compared with the write operation results, in the case of file rewriting, the performance of all virtual machines increases significantly. The KVM virtual machine is the most powerful, being exceeded by Hyper-V only for large-sized files. In case of reading, KVM and Hyper-V offer comparable performance while the VMware virtual machine exhibits, this time, weaker performance. In the case of random readings, all three types of virtual machines analyzed have similar performance, with a small gain for VMware and Hyper-V, whereas in case of random writing, Hyper-V and VMware behave significantly better than KVM. The results of the performance evaluation experiments reflect the fact that there is no universal solution that proves maximum performance in all three points of interest. Depending on the specifics of the users' applications, it is recommended to use a certain solution.

Regarding the CPU performance, the recommended solution to get the best performance results is VMware. KVM and Hyper-V show similar results in memory performance, but in case of file system operations, the recommendation is Microsoft Hyper-V. VMware demonstrated the best results in terms of both, CPU performance and in process handling. So, in case of CPU power requirements, VMware is the optimal solution, while KVM ranks second. On the other hand, Hyper-V stands out significantly from the two in both test scenarios. This may be caused to some extent by the FreeBSD Integration Services drivers installed to make Hyper-V working. VMware manifests the worst behavior in terms of the speed of read and write operations on memory. When writing floating-point blocks, VMware performance is good, but in the case of integer blocks, VMware drops significantly. KVM and Hyper-V are in permanent competition for the first place in memory operations. Thus, when reading from memory, KVM is recommended when operating with large blocks, and Hyper-V is preferred for small-block read and write operations of integer blocks. In case of floating-point blocks, KVM and Hyper-V show similar results. From the point of view of file operations, Hyper-V works better than KVM and VMware. KVM is the leader only for overwriting operations, and VMware for random read operations. For writing operation, KVM and Hyper-V show similar results.

Performance evaluation comparison: public Cloud vs. Private Cloud

In terms of CPU performance, in the private Cloud all tested virtual machines show similar results, while case of public Cloud differences are observed. This may be caused by variable machine loads in the public Cloud, while in the private Cloud the physical machine was used exclusively for performance evaluation, being loaded much under its capacity. This is also the explanation for the fact that, overall, the score obtained on virtual machines in the public Cloud is lower to that obtained in the private Cloud.

The memory read and write operations in floating point were configured to read or write 64Kb floating-point data blocks. In case of the public Cloud, there are small differences in performance between the virtual machines and the overall performance of virtual machines running in the public Cloud is slightly lower than the performance of the virtual machines running in the private Cloud.

To evaluate the performance of the file system, tests have read and written 1 MB files. Both, in reading and writing files, the private Cloud exhibits better performance than the public Cloud. Although the results of the performed test, in absolute value, can sometimes be considered not to be fully relevant because physical resources in the public Cloud can not be controlled, the comparative analysis between the two Cloud types is fully relevant, being very useful for users when deciding on the type of the Cloud service recommended for the specifics of their applications. The lower performance of virtual machines in the public Cloud is explained by sharing physical resources between multiple clients, resulting in CPU load fluctuations, memory access time, and I/O traffic.

In all experiments, the private Cloud proved to be superior to the public Cloud. At the same time, in case of the public Cloud, the experiment revealed an unreasonable variation in performance between virtual machines with identical configurations. The reason for this performance instability could be caused by the sharing of public Cloud resources between different users, resulting in uneven machine loads. The conclusion of these performance ratings is that when performance and security are the most important requirements, it is recommended to use a private Cloud. If priority issues are cost and scalability, it is recommended to use a public Cloud.

Evaluation of the proposed mathematical models

Mathematical modeling is certainly a useful way to improve the performance of virtualized systems, as it is possible to draw interesting conclusions about the virtualized behavior of many practical scenarios by following the implementation of various mathematical models.

The mathematical models proposed in Chapter 4 aimed at determining how the variance of the arrival rate λ affects the response time of the system. An accelerated arrival rate requires a larger number of processing nodes to optimize the response time. However, the simulation performed demonstrates that when the arrival rate increases, the use of the system decreases asymptotically with the increase in the number of processing nodes and there is a limit after which the increase in the number of processing nodes will no longer decrease the response time. This result is contrary to the expectation that by increasing the number of processing nodes, the response time will decrease exponentially. The response time stabilizes with the servers utilization. When using virtualization, the response time of the system is slightly higher than that of the base model. This behavior is explained by the processing the queues in two steps, initially at the physical machines level, then at the virtual machines level. In a real scenario, additional consumption of virtualization resources will also have to be considered.

Evaluation of the proposed scheduling policies

From the need to extend the functionality of the proposed mathematical models, a two-level event-based scheduling policy has been proposed. Its purpose is to map the tasks to the processing nodes, according to the nodes processing power and availability. The high-level policy assigns virtual machines to physical machines, and the low-level policy assigns tasks to virtual machines. These two policies were evaluated by simulation using the CloudSim simulator.

The proposal of a planning policy was motivated by the minimization of energy consumption. Ideally, the total response time for tasks assigned to virtual machines should not be affected by the virtual system scheduler. This assertion is based on the argument that each virtual machine presents its own computational requirements, and no matter which physical machine is chosen to host the virtual machine, the virtual machine will present these minimum requirements. This theory is aimed at building the algorithm on which planning policy is based.

The way virtual machines are allocated to physical hosts determines the energy consumption of the entire system. Energy consumption is not linear and its modeling would lead to a complex analytical problem. The power and performance characteristics of server computing devices have been measured using a framework designed to compare power and performance between different servers and serve as a toolkit to improve server efficiency.

The most common way of high-level planning is a circular one based on the Round-robin algorithm. Because this algorithm does not consider minimizing energy consumption, a customized planning policy has been proposed, which aims to allocate virtual machines according to the number of available cores. The principle behind this policy is the following: for each virtual machine to be allocated, the physical machine with the least number of available cores is selected. In this way, hosts with a lower capacity are used. The decrease in energy consumption is achieved by leaving the physical machines stronger inactive for as long as possible. The way virtual machines are allocated to host physical machines determines the energy consumption of each host and, ultimately, the energy efficiency of the entire data center.

The proposed policy was evaluated by simulation and compared to Round-robin policy. The results demonstrate a significant improvement in the datacenter power consumption when using the proposed planning policy. Reducing consumption, compared to the Round-robin implicit policy, is 7.9%. This shows how efficient virtual machine allocation can lead to a more energy-efficient system. Assigning virtual machines to hosts does not affect the response times of tasks, which will then be allocated to each virtual machine.

The virtual machine-task mapping policy has a direct impact on the response time of the system and determines the completion time of each task. A balanced planning policy has been proposed in this direction. The results of the simulation evaluation of the proposed policy are compared to those of the Round-robin scheduling simulation and demonstrate that system performance can increase by up to 25% if factors such as availability and capacity of the virtual machine are considered. This improvement is due not only to the virtual machine's computing capability but also to its current state. It is recommended to load the stronger virtual machines avoiding overloading and leaving the least powerful, overloaded.

Evaluation of resource utilization and of the scalability in Cloud

The interest in evaluating the use of computing resources (processor, RAM, and bandwidth) led to an experiment in which the degree of use of the computing resources was tested. Initially, the data center behavior experienced a peak in CPU usage, which then stabilized over time. This is explained by the fact that, in the first phase, at the arrival of the tasks in the system, planning and assignment of tasks to virtual machines must be performed. Then, the amount of resources needed to maintain the execution of the tasks decreased until it stabilized. Memory and processor usage are closely related, but the processor has been used on a larger scale. As far as network traffic capacity is concerned, it can be appreciated that there was no communication between workloads. This is the distinctive feature of the parallel applications that have been used throughout the experiment. To test the scalability feature, two data centers with features similar to the first data center were added and a significant decrease in total execution time was recorded. In this case, CloudSim used a Round-robin circular completion policy and thus runtime tends to stabilize.

Scalability was further tested by initially using three data centers, then increasing their number to eight data centers with similar features. Execution time has been found to have dropped significantly when adding the first data centers after which, when new data centers were added, decreases in execution time lowered and tended to stabilize at a certain amount. The presence of multiple data centers ensures a high level of robustness in the sense that, if one of the centers becomes unavailable, the execution of incoming tasks is ensured, and they are redirected to another center by applying a heuristic policy. This behavior is reflected in an increased tolerance of system error. However, the response time is significantly reduced to a certain limit, after which it tends to stabilize, and the addition of the number of data centers no longer leads to a decrease.

Future research directions

Experimentele de evaluare a performanței demonstrează rezultate precum: reducerea costurilor hardware, reducerea consumului de energie și utilizarea eficientă a mașinilor fizice ș.a. Având în vedere utilizarea pe scară tot mai largă a serviciilor de tip Cloud bazate pe virtualizare, determinată în primul rând de considerente economice, continuarea cercetării în domeniul virtualizării în vederea îmbunătățirii performanțelor sistemelor Cloud și a reducerii costurilor pentru utilizatorii de servicii Cloud este deosebit de utilă. Rezultatele obținute, prezentate în această lucrare, dovedesc că se pot aduce multe îmbunătățiri sistemelor de calcul bazate pe virtualizare și că domeniul calculului de tip Cloud nu este doar de mare actualitate, dar și de viitor.

Performance evaluation experiments demonstrate results such as: reducing hardware costs, reducing energy consumption and efficient use of physical machines, and more. Given the increasingly widespread use of virtualization-based Cloud services, primarily driven by economic considerations, continuing the virtualization research in order to improve Cloud performance and to reduce costs for Cloud service users is particularly useful. The results, presented in this paper, prove that many improvements can be provided to virtualization-based computing systems and that the field of Cloud computing is a future subject of interest.

Research issues of interest in optimizing the performance of Cloud computing are many: the availability of services, energy, supply and management, virtual machine management, load balancing, security, etc. These issues open up several research directions, which also introduce many challenges. Studies on virtualization and its use in Cloud systems presented in the paper can be expanded in various directions.

Considering the results of experiments performance analysis presented in this paper the future development of new planning policies taking into account the new criteria and scenarios involving complex task synchronization and communication between tasks components, would be of great interest. It would also be useful to develop new models that take into account data replication and the use of multiple storage servers.

Moreover, in the near future, it is intended to design and to implement the proposals presented in this paper in real Cloud systems through OpenStack platform. OpenStack platform allows the control of a diverse range of computing, storage and communication resources, being the ideal platform for heterogeneous infrastructures.

Another future direction of research is the scalability of Cloud services and the limitations imposed by the available hardware resources. In such case, a Cloud service provider might delegate some of the workloads to another Cloud service provider in a transparent way for users. Such a scenario requires SLA (Service Level Agreement) research and investigation of various SLA issues such as QoS monitoring, SLA violation, financial implications, etc.

References

Abirami, S., & Shalini, R. (2012). Linear Scheduling Strategy for Resource Allocation in Cloud Environment. *International Journal on Cloud Computing: Services and Architecture, Vol. 2, No. 1*.

Chang, B., Tsai, H.-F., & Chen, C.-M. (2013). Evaluation of Virtual Machine Performance and Virtualized Consolidation Ratio in Cloud Computing System. *Journal of Information Hiding and Multimedia Signal Processing*.

Diogo, M. M., & Ferraz, L. G. (2012). Virtual Network Performance Evaluation for Future Internet Architectures. *Journal of Emerging Technologies in Web Intelligence, Vol. 4, No. 4*.

Falatah, M., & Batarfi, O. (2014). Cloud Scalability Considerations. *International Journal of Computer Science & Engineering Survey (IJCSES), Vol.5, No.4*.

Goswami, B., & Saha, S. (2013). Resource Allocation Modeling in Abstraction using Predator-Prey dynamics: A qualitative analysis. *International Journal of Computer Applications, Vol. 61, Nr. 6*.

Graniszewski, W., & Arciszewski, A. (2016). Performance analysis of selected hypervisors (Virtual Machine Monitors - VMMs). *International Journal of Electronics and Telecommunications*, 231-236.

Indrani, P., Sudhakar, Y., & Lizy, J. K. (2012). Performance Impact of Virtual Machine Placement in a Datacenter. *IEEE 31st International Performance Computing and Communications Conference*, (pp. 424-431). Austin, TX.

ISO-International Standard for Organizations. (2014, 10 15). Information technology - Cloud computing - Overview and vocabulary, ISO/IEC17788.

Johnsen, E., Schlatte R., R., & Tapia, S. T. (2012). Modeling Resource-Aware Virtualized Applications for the Cloud in Real-Time ABS. *ICFEM 2012. Lecture Notes in Computer Science, vol 7635.* Ber: Springer, Berlin, Heidelberg.

Kundu, S., Rangaswami, R., Dutta, K., & Zhao, M. (2010). *Application Performance Modeling in a Virtualized Environment.* Florida, USA: School of Computing & Information Sciences, College of Business Administration Florida International University.

Mancaș, C. (2015). Performance Improvement through Virtualization. *RoEduNet 13th International Conference: Networking in Education and Research.* Craiova.

Mancaș, C. (2014). Best Practices in Distributed Learning Environments. *The International Scientific Conference eLearning and Software for Education, Vol. 3* (pp. 272-279). București: "Carol I" National Defence University.

Mancaș, C., Mocanu, M., & Mancaș, D. (2013). Congestion Avoidance in Multimedia Networks. *RoEduNet 11th International Conference: Networking in Education and Research.* Sinaia, România.

Mell, P., & Grance, T. (2011). *The NIST Definition of Cloud Computing.* Gaithersburg: Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology, United States Department of Commerce.

Tarasov, V., Bhanage, S., Zadok, E., & Seltzer, M. (2011). Benchmarking File System Benchmarking: It *IS* Rocket Science. HotOS.

Vinothina, V., Sridaran, R., & Ganapathi, P. (2012). A Survey on Resource Allocation Strategies in Cloud Computing. *International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 3, No. 6*.

Wang, Q., & Varela, C. A. (2011). Impact of Cloud Computing Strategies on Workloads Performance. *IEEE Computer Society*, 130-137.

Yamato, Y. (2017). Performance-aware server architecture recommendation and automatic performance verification technology on IaaS Cloud. *Service Oriented Computing and Applications, Vol. 11, No. 2*, 121-135.

Zhang, X., Tune, E., Hagmann, R., Jnagal , R., Gokhale, V., & Wilkes, J. (2013). *CPI: CPU performance isolation for shared compute clusters.* ACM, Google Inc.