Volume 4, Issue 4, April 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Review of Cloud Systems and Resource Allocation Strategies

Nisha Sharma
M.Tech Scholar,CSE Dept, JMIT Radaur
KURUKSHETRA UNIVERSITY, India

Mamta Dhanda
Asst. Professor,CSE Dept, JMIT Radaur
KURUKSHETRA UNIVERSITY, India

Abstract— Cloud computing is on demand service as it offers dynamic, flexible and efficient resource allocation for reliable and guaranteed services in pay-as-you-use manner to the customers. Cloud computing offers multiple cloud users requesting number of cloud services simultaneously, so there must be a provision that all resources are made available to requesting user in efficient manner to satisfy their need without compromising on the performance of the resources. One of the major challenges in cloud computing is related to optimizing the resources being allocated over various virtual machines. This Paper presents design implementations, and evaluates resource management for cloud computing services.

Keywords— Cloud Computing, Cloud Services,Virtual machines, Virtuallization, Resource Allocation.

I. INTRODUCTION

Pardeep Kumar, AmandeepVerma, Independent Task Scheduling in Cloud Cloud Computing, the long-held dream of computing as a utility, has the potential to transform a large part of the IT industry, making software even more attractive as a service and shaping the way IT hardware is designed and purchased. Developers with innovative ideas for new Internet services no longer require the large capital out lays in hardware to deploy their service or the human expense to operate it. They need not be concerned about over provisioning for a service whose popularity does not meet their predictions, thus wasting costly resources, or under provisioning for one that becomes wildly popular, thus missing potential customers and revenue. Moreover, companies with large batch-oriented tasks can get results as quickly as their programs can scale, since using 1000 servers for one hour costs no more than using one server for 1000 hours. This elasticity of resources, without paying a premium for large scale, is unprecedented in the history of IT. Cloud Computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services. The services themselves have long been referred to as Software as a Service (SaaS). The datacenter hardware and software is what we will call a Cloud. When a Cloud is made available in a pay-as-you-go manner to the general public, we call it a Public Cloud; the service being sold is Utility Computing. We use the term Private Cloud to refer to internal datacenters of a business or other organization, not made available to the general public. Thus, Cloud Computing is the sum of SaaS and Utility Computing, but does not include Private Clouds. People can be users or providers of SaaS, or users or providers of Utility Computing. We focus on SaaS Providers (Cloud Users) and Cloud Providers, which have received less attention than SaaS Users. Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics**,** three service models, and four deployment models.

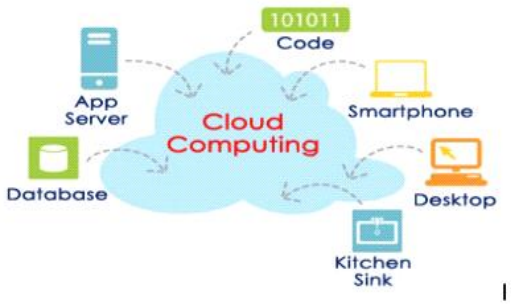

Figure 1: Cloud Platform on the web

Essential Characteristics:

On demand self-service:-A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.

Broad network access: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g. mobile phones, tablets, laptops, and workstations).

Resource pooling: The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand.

There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth.

Rapid elasticity*:* Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

Measured service: Cloud systems automatically control and optimize resource use by leveraging metering capability1 at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

Service Models:

Software as a Service (SaaS). The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user specific application configuration settings.

Platform as a Service (PaaS). The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming.

Typically this is done on a pay-per-use or charge-per-use basis.

A cloud infrastructure is the collection of hardware and software that enables the five essential characteristics of cloud computing. The cloud infrastructure can be viewed as containing both a physical layer and an abstraction layer. The physical layer consists of the hardware resources that are necessary to support the cloud services being provided, and typically includes server, storage and network components. The abstraction layer consists of the software deployed across the physical layer, which manifests the essential cloud characteristics. Conceptually the abstraction layer sits above the physical layer. Languages, libraries, services, and tools supported by the provider.

The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.

Infrastructure as a Service (IaaS).The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls).

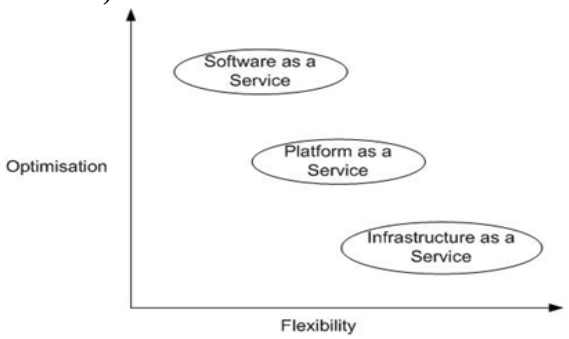

Fig 2 – Software, Platform and Infrastructure Services

II. Virtualization

In a virtualized cloud computing environment, customers may never know exactly where their data is stored. In fact, data may be stored across multiple data centers in an effort to improve reliability, increase performance, and provide redundancies. This geographic dispersion may make it more difficult to ascertain legal jurisdiction if disputes arise

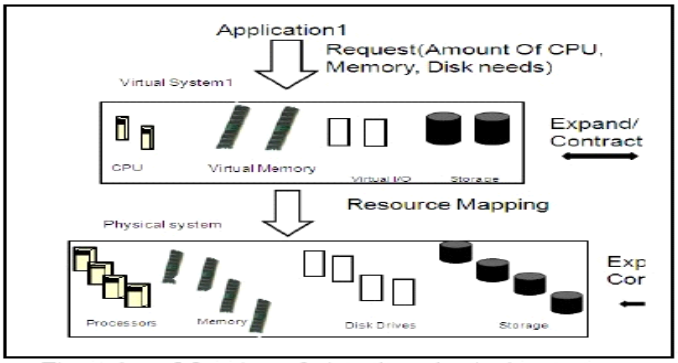

Fig 2: Mapping of physical to Virtual Resources

Virtual Machine.

As discussed earlier, a host can simultaneously instantiate multiple VMs and allocate cores based on predefined processor sharing policies (space-shared, time-shared). Every VM component has access to a component that stores the characteristics related to a VM, such as memory, processor, storage, and the VM's internal scheduling policy, which is extended from the abstract component called VM Scheduling.

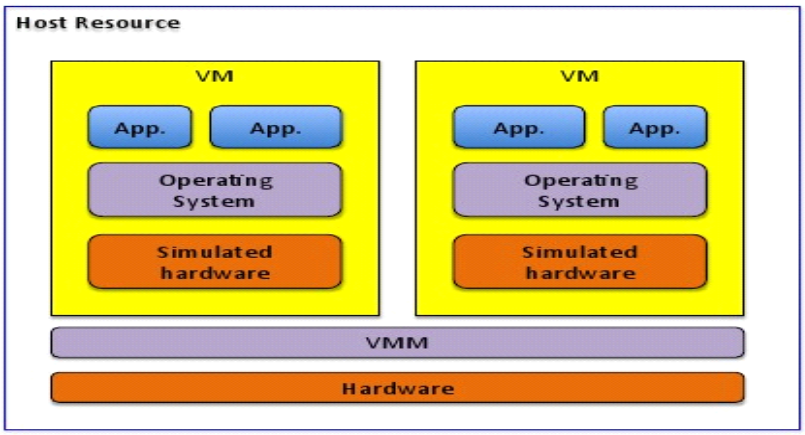

Figure3: Virtual Machine Architecture

A system which can automatically scale its infrastructure resources is designed in. The system composed of a virtual network of virtual machines capable of live migration across multi- domain physical infrastructure. By using dynamic availability of infrastructure resources and dynamic application demand, a virtual computation environment is able to automatically relocate itself across the infrastructure and scale its resources. But the above work considers only the non-preempt able scheduling policy. Several researchers have developed efficient resource allocations for real time tasks on multiprocessor system. But the studies, scheduled tasks on fixed number of processors.

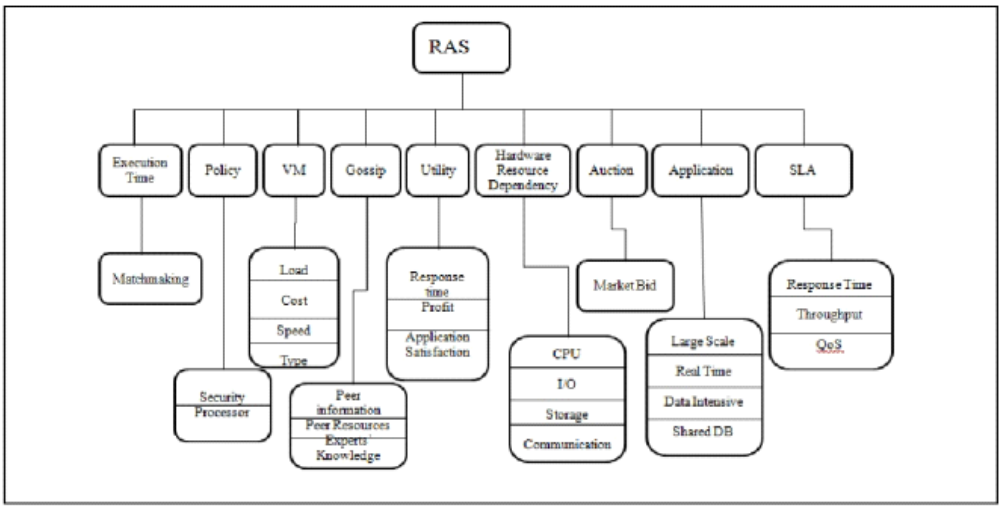

Fig4: Resource allocation Strategies

Hence it is lacks in scalability feature of cloud computing. Recent studies on allocating cloud VMs for real time tasks focus on different aspects like infrastructures to enable real-time tasks on VMs and selection of VMs for power management in the data center. But the allocation of resources based on the speed and cost of different VMs in IaaS. It differs from other related works, by allowing the user to select VMs and reduces cost for the user.

III. Applications

Virtual infrastructure allocation strategies are designed for workflow based applications where resources are allocated based on the workflow representation of the application. For work flow based applications, the application logic can be interpreted and exploited to produce an execution schedule estimate. This helps the user to estimate the exact amount of resources that will be consumed for each run of the application.

Four strategies such as Naive, FIFO, Optimized and services group optimization are designed to allocate resources and schedule computing tasks.

Real time application which collects and analyzes real time data from external service or applications has a deadline for completing the task. This kind of application has a light weight web interface and resource intensive back end. To enable dynamic allocation of cloud resources for back-end, a prototype system is implemented and evaluated for both static and adaptive allocation with a test bed cloud to allocate resources to the application.

The system also accommodates new requests despite a-priori undefined resource utilization requirements. This prototype monitoring the CPU usage of each virtual machine and adaptively invoking additional virtual machines as required by the system. Have suggested the integration of high bandwidth radar sensor networks with computational and storage resources in the cloud to design end-to-end data intensive cloud systems.

IV. RELATED WORK

N.Krishnaveni and G.Sivakumar ,et al in **"** Survey on Dynamic Resource Allocation Strategy in Cloud Computing Environment"[1] presents that Cloud computing becomes quite popular among cloud users by offering a variety of resources. This is an on demand service because it offers dynamic flexible resource allocation and guaranteed services in pay as-you-use manner to public. The author presents the several dynamic resource allocation techniques and its performance. The author in this paper provides detailed description of the dynamic resource allocation technique in cloud for cloud users and comparative study provides the clear detail about the different techniques. The author in this paper addresses the theoretic study of various dynamic resource allocation techniques in cloud environment. The author provides the detail description of the techniques is summarized and also summarizes the advantages with parameters of the various techniques in cloud computing environment.

Shabnam Khan, et al in **"**A survey on scheduling based resource allocation in Cloud Computing"[2] presents that Cloud computing is the next generation of technology which unifies everything into one. It is an on demand service because it offers dynamic flexible resource allocation for reliable and guaranteed services in pay as-you-use manner to public. In Cloud computing multiple cloud users can request number of cloud services simultaneously. So there must be a provision that all resources are made available to requesting user in efficient manner to satisfy their need. The author in this paper a review of various policies for dynamic resource allocation in cloud computing is shown based on Topology Aware Resource Allocation (TARA), Linear Scheduling Strategy for Resource Allocation and Dynamic Resource Allocation for Parallel Data Processing. Moreover, significance, advantages and limitations of using Resource Allocation in Cloud computing systems is also discussed. The other challenges of resource allocation are meeting customer demands and application requirements. The author in this paper discusses various resource allocation strategies and their challenges. It is believed that this paper would benefit both cloud users and researchers in overcoming the challenges faced. Scheduling is one of the most important tasks in cloud computing environment. The author have analyzed various scheduling algorithm and tabulated various parameter. The author has noticed that disk space management is critical issue in virtual environment. Existing scheduling algorithm gives high throughput and cost effective but they do not consider reliability and availability. So the author need algorithm that improves availability and reliability in cloud computing environment. In future enhancement will propose a new algorithm for resource scheduling and comparative with existing algorithms.

Vaghela Ankita , et al in **"** A survey on various resource allocation policies in cloud computing environment"[3] presents Cloud computing is bringing a revolution in computing environment replacing traditional software installations, licensing issues into complete on-demand services through internet. In Cloud computing multiple cloud users can request number of cloud services simultaneously. So there must be a provision that all resources are made available to requesting user in efficient manner to satisfy their need. Resource allocation is based on quality of service and service level agreement. In cloud computing environment, to allocate resources to the user there are several methods but provider should consider the efficient way to guarantee that the applications' requirements are attended to correctly and satisfy the user's need the author survey different resource allocation policies used in cloud computing environment. Cloud computing is the delivery of computing as a service rather than a product, whereby shared resources, software and information are provided to users over the network. It is an on demand service because it offers dynamic flexible resource allocation for reliable and guaranteed services in pay as-you-use manner to public. The author discusses various resource allocation policies based on service level agreement and quality of service. All above methods used different techniques for allocating resource to the user. Time driven gives better response time and increase resource utilization. Multidimensional SLA based algorithm save resources and increase resource utilization. Multidimensional algorithm increase resource utilization and reduce cost of data center. SLA-based policy minimizes the saas provider's cost and the number of SLA violations. Adaptive resource allocation policy increase resource utilization. Policy based resource allocation maximize resource utilization.

Anshul Rai, Ranjita Bhagwan, Saikat Guha,et al in "Generalized Resource Allocation for the Cloud"[4] presents that Resource allocation is an integral, evolving part of many data center management problems such as virtual machine placement in data canters, network virtualization, and multi-path

Network routing. Since the problems are inherently NP-Hard, most existing systems use custom-designed heuristics to find a suitable solution. However, such heuristics are often rigid, making it difficult to extend them as requirements change. The authors present a novel approach to resource allocation that permits the problem specification to evolve with ease. The authors have built Wrasse, a generic and extensible tool that cloud environments can use to solve their specific allocation problem. Wrasse provides a simple yet expressive specification language that captures a wide range of resource allocation
Problems. At the back-end, it leverages the power of GPUs to provide solutions to the allocation problems in a fast and timely manner. We show the extensibility of Wrasse by expressing several allocation problems in its specification

language. Our experiments show that Wrasse's solution quality is as good as with heuristics, and sometimes even better, while maintaining good performance. In one case, Wrasse packed 71% more instances than a custom heuristic.

V. CONCLUSION

We have discussed the in this paper one of the major challenges in cloud computing is related to optimizing the resources being allocated over various virtual machines. This Paper presents design implementations, and evaluates resource management for cloud computing services. We have reviewed Cloud Computing, the long-held dream of computing as a utility, has the potential to transform a large part of the IT industry, making software even more attractive as a service and shaping the way IT hardware is designed and purchased. Developers with innovative ideas for new Internet services no longer require the large capital out lays in hardware to deploy their service or the human expense to operate it.

FUTURE SCOPE

Due to its major advantages, service-oriented architecture has been adopted in various distributed systems, such as web services, grid computing systems, utility computing systems and cloud computing systems. In order to properly use these systems, especially cloud systems in various applications, one major challenge which must be addressed is to manage the resource allocation and related policies to satisfy users' requirements. In our Future work we will implement resource allocation strategies under cloud simulation framework.

REFERENCES

[1] N.Krishnaveni, G.Sivakumar "Survey on Dynamic Resource Allocation Strategy in Cloud Computing Environment" Dept. of CSE Erode Sengunthar Engineering College Thudupathi, India, International Journal of Computer Applications Technology and Research Volume 2– Issue 6, 731 - 737, 2013.

[2] Shabnam Khan "A survey on scheduling based resource allocation in cloud computing" Computer Science and Engineering Department, Sobhasaria Engineering College, Sikar, Rajasthan, India, International Journal For Technological Research In Engineering Vol. 1, Issue. 1, Sep – 2013.

[3] Vaghela Ankita **"** A survey on various resource allocation policies in cloud computing environment" Department of Computer Engineering, Alpha College of Engineering and Technology, Gujarat, India, Volume. 2, Issue. 5, ISSN: 2319 – 1163, 760 – 763.

[4] Anshul Rai, Ranjita Bhagwan, Saikat Guha, "Generalized Resource Allocation for the Cloud" Microsoft Research India.

[5] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I.Stoica and M. Zaharia. ?Above the Clouds: A Berkeley View of Cloud Computing?. UC Berkeley Reliable Adaptive Distributed Systems Laboratory, 2009.

[6] S.M. Hashemi, A.Kh. Bardsiri, ?Cloud Computing Vs. Grid Computing?. ARPN Journal of Systems and Software, Vol. 2, No 5, pp. 188-194, 2012.

[7] J. Geelan, ?Twenty one experts define cloud computing. Virtualization?. Electronic Magazine, article availableathttp://virtualization.syscon.com/node/612375, 2008.

[8] Y. Zhang, A. Mandal, C. Koelbel1 and K. Cooper, ?Combined Fault Tolerance and Scheduling Techniques for Workflow Applications on Computational Grids?. in 9th IEEE/ACM international symposium on clustering and grid, pp. 244-251 , 2009.

[9] K. Liu, J. Chen, Y. Yang, H. Jin, ?A throughput maximization strategy for scheduling transaction intensiveworkflows on SwinDeW -G?. Concurrent Computing, Vol. 20, No. 15, pp. 1807–1820, 2008.

[10] M. Wang, R. Kotagiri, J. Chen ?Trust-based robust scheduling and runtime adaptation of scientific workflow?.Concurrent Computing, Vol. 21, No. 16, pp. 1982–1998, 2009.

[11] E. Deelman, D. Gannon, M. Shields, I. Taylor, ?Workflows and e-science: an overview of workflow system features and capabilities?. Future GenerComputSyst, Vol. 25, No. 6, pp. 528–540, 2008.

