Exemplifying Practical issues of Resource Management in Cloud Computing

Akash Rana*, Dr. Gundeep Tanwar**
* Department of Computer Applications, E-max Group of Institutions Ambala, Haryana
** Department of Computer Science & Engg., BRCM College of Engineering & Technology, Bhiwani, Haryana

***Abstract*— Cloud computing is a budding business infrastructure archetype that guarantees to eradicate the necessitates for maintaining exclusive computing hardware. Even so, the prospective of using Cloud computing infrastructure to sustain computational and data-intensive scientific applications has not yet been adequately addressed. Resource management for an essentially complex system such as cloud computing requires dissimilar ways of measuring and allocating resources. Managing resources at huge scale while providing performance isolation and efficient use of underlying hardware is a key confront for any cloud management software. The majority implicit machine (VM) resource management systems like VMware DRS clusters, Microsoft PRO and Eucalyptus, do not presently scale to the quantity of hosts and VMs needed by cloud offerings to sustain the suppleness required handling peak demand. In addition to scale, other problems a cloud-level resource management layer needs to resolve include heterogeneity of systems, compatibility constraints between virtual machines and underlying hardware, islands of resources created due to storage and network connectivity and imperfect scale of storage resources.**
In this paper, we get rid of several foundation challenges in building a cloud-scale resource management system based on past research and shipping cluster resource management products. Additionally, we converse various techniques to grip these challenges, along with the pros and cons of each technique. We expect to stimulate future research in this area to extend practical solutions to these issues.

Keywords*— *Clouding Computing, Resource Management, Practical issue

I. INTRODUCTION

Resource management is a foundation job required of any man-made system. It affects the three fundamental criteria for system evaluation as given below:

1. Performance
2. Functionality
3. Cost

Inefficient resource management has an unswerving unconstructive consequence on performance and cost. It can also indirectly influence system functionality. Some functions the system provides might become too costly or unproductive due to pitiable performance. A cloud computing infrastructure is a multifaceted system with a huge number of collective resources. These are subject to unpredictable requests and can be affected by exterior events beyond user control.

Figure 1

The cloud resource management requires multifarious policies and decisions for multi-objective optimization. It is enormously challenging because of the complication of the system, which makes it impracticable to have precise comprehensive state information.
It is also subject to unremitting and impulsive interactions with the environment. The strategies for cloud resource management connected with the three cloud delivery models as given below:

- Infrastructure as a Service (IaaS)
- Platform as a Service (PaaS)
- Software as a Service (SaaS)

They differ from one another. In all cases, the cloud services providers are faced with massive, unpredictable loads that tackle to preserve of cloud suppleness. In some cases, when they can estimate a spike can be predicted, they can prerequisite assets in advance. For example, seasonal Web services may be subject to spikes. For an unplanned impale, the situation is slightly more complicated. The user can employ Auto Scaling for unintentional impale loads, provided there's a pool of resources the users can release or allocate on demand and a monitoring system that lets user decide in real time to budge resources. Auto Scaling is supported by PaaS services such as Google App Engine. Auto Scaling for IaaS is complicated due to the lack of standards.
In the cloud, where changes are frequent and unpredictable, centralized control is unlikely to provide continuous service

and performance guarantees. Indeed, centralized control can't provide adequate solutions to the host of cloud management policies user have to enforce. Autonomic policies are of great interest due to the scale of the system, the large number of service requests, the large user population and the unpredictability of the load. The ratio of the mean to the peak resource needs can be large.

II. POLICIES AND MECHANISMS

A policy typically refers to the principal guiding decisions, whereas mechanisms represent the means to implement policies. Separating policies from mechanisms is a guiding principle in computer science. Butler Lampson and Per Brinch Hansen offer solid arguments for this separation in the context of OS design. User can loosely group cloud resource management policies into five classes:

The explicit goal of an admission control policy is to prevent the system from accepting workloads in violation of high-level system policies. For example, a system may not accept an additional workload that would prevent it from completing work already in progress or contracted. Limiting the workload requires some knowledge of the global system state. In a dynamic system, this information is often obsolete at best. The capacity allocation means allocating resources for individual instances. An instance is service activation. Locating resources that are subject to multiple global optimization constraints requires user to a search a large space when the state of individual systems is changing so rapidly. User can perform load balancing and energy optimization locally, but global load-balancing and energy-optimization policies encounter the same difficulties as the ones already discussed. Load balancing and energy optimization are correlated and affect the cost of providing the services.

The common meaning of the term load balancing is that of evenly distributing the load to a set of servers. In cloud computing, a critical goal is minimizing the cost of providing the service. In particular, this also means minimizing energy consumption. This leads to a different meaning of the term load balancing. Instead of having the load evenly distributed among all servers, we want to concentrate it and use the smallest number of servers while switching the others to standby mode, a state in which a server uses less energy. In our example, the load from D will migrate to A and the load from C will migrate to B. Thus, A and B will be loaded at full capacity, whereas C and D will be switched to standby mode. Quality of service is that aspect of resource management that's probably the most difficult to address and, at the same time, possibly the most critical to the future of cloud computing. Resource management strategies often jointly target performance and power consumption.

Dynamic voltage and frequency scaling (DVFS) techniques such as Intel SpeedStep and AMD PowerNow lower the voltage and the frequency to decrease power consumption. Motivated initially by the need to save power for mobile devices, these techniques have migrated to virtually all processors, including those used in high-performance servers. As a result of lower voltages and frequencies, the processor performance decreases. However, it does so at a substantially slower rate than the energy consumption. Virtually all optimal or near-optimal mechanisms to address the five policy classes don't scale up. They typically target a single aspect of resource management, such as admission control, but ignore energy conservation. Many require complex computations that can't be done effectively in the time available to respond. Performance models are complex, analytical solutions are intractable, and the monitoring systems used to gather state information for these models can be too intrusive and unable to provide accurate data.

Therefore, many techniques are concentrated on system performance in terms of throughput and time in system. They rarely include energy tradeoffs or QoS guarantees. Some techniques are based on unrealistic assumptions. For example, capacity allocation is viewed as an optimization problem, but under the assumption that servers are protected from overload.

III. RELATED WORK

Kandalintsev et al. (2012) stated that software methods did not have control over low-level hardware circuit modules. Built-in hardware methods had very fine-grained control, but their impact was limited to a specific microchip unit. In this study they seemed to address this problem by developing algorithms that improve interoperability and combine the benefits of both software and hardware approaches delivering significant energy savings.

Rathore et al. (2011) stated that In case of the High Performance Computing (HPC), providing adequate resources for user applications was crucial. For instance, a computing center that a user has access to cannot handle the user applications with short deadlines due to limited computing infrastructure in the center. Therefore, to get the application completed by the deadline, users usually tried to get access to several computing centers (resources). However, managing several resources, potentially with different architectures, was difficult for users. Another difficulty was optimally scheduling applications in such environment. In this paper we were giving the strategy how the resource managed in cloud environment.

Irwin et al. (2010) argued that the cloud paradigm was also well suited for handling data-intensive applications, characterized by the processing and storage of data produced by high-bandwidth sensors or streaming applications. The data rates and the processing demands varied over time for many such applications, making the on-demand cloud paradigm a good match for their needs. However, today's cloud platforms needed to evolve to meet the storage, communication, and processing demands of data-intensive applications. We presented an ongoing GENI project to connect high-bandwidth radar sensor networks with computational and storage resources in the cloud and used this example to highlight the opportunities and challenges in designing end-to-end data-intensive cloud systems.

Gulati et al. (2011) shed light on some of the key issues in building cloud-scale resource management systems, based on five years of research and shipping cluster resource management products. Furthermore, they discussed various techniques to provide large scale resource management, along with the pros and cons of each technique. they hoped to motivate future research in this area to develop practical solutions to these issues.

Hu et al. (2010) argued that the resource provisioning for cloud computing, an important issue was how resources may

be allocated to an application mix such that the service level agreements (SLAs) of all applications are met. A performance model with two interactive job classes was used to determine the smallest number of servers required to meet the SLAs of both classes. For each class, the SLA is specified by the relationship: Prob [response time $\leq$ x] $\geq$ y. Two server allocation strategies are considered: shared allocation (SA) and dedicated allocation (DA). For the case of FCFS scheduling, analytic results for response time distribution were used to develop a heuristic algorithm that determined an allocation strategy (SA or DA) that required the smallest number of servers. The effectiveness of this algorithm was evaluated over a range of operating conditions. The performance of SA with non-FCFS scheduling was also investigated. Among the scheduling disciplines considered, a new discipline called probability dependent priority was found to have the best performance in terms of requiring the smallest number of servers.

Sasitharagai et al. (2013) stated that the problem of dynamic resource management for a large-scale cloud environment was mitigated with optimized high throughput performance. The resource management framework consisted of, Gossip protocol that ensured fair resource allocation among sites by calculating Memory Load Factor and CPU Load Factor and routing table for dynamically managing the tasks. A request partitioning approach based on gossip protocol was proposed that facilitates the cost-efficient and online splitting of user requests among eligible Cloud Service Providers within a networked cloud environment. Following the outcome of the request partitioning phase, the embedding phase - where the actual mapping of requested virtual to physical resources was performed that allows for efficient and balanced allocation of cloud resources. Finally, a thorough evaluation of the overall framework on a simulated cloud environment was made, which offers reliable and dynamic resource management.

Buyya et al. (2010) promised to offer subscription-oriented, enterprise-quality computing services to users worldwide. With the increased demand for delivering services to a large number of users, they needed to offer differentiated services to users and meet their quality expectations. Existing resource management systems in data centers are yet to support Service Level Agreement (SLA)-oriented resource allocation, and thus needed to be enhanced to realize cloud computing and utility computing.

IV. SYSTEM MODEL

The system model for provisioning cloud resources is shown in figure 2.

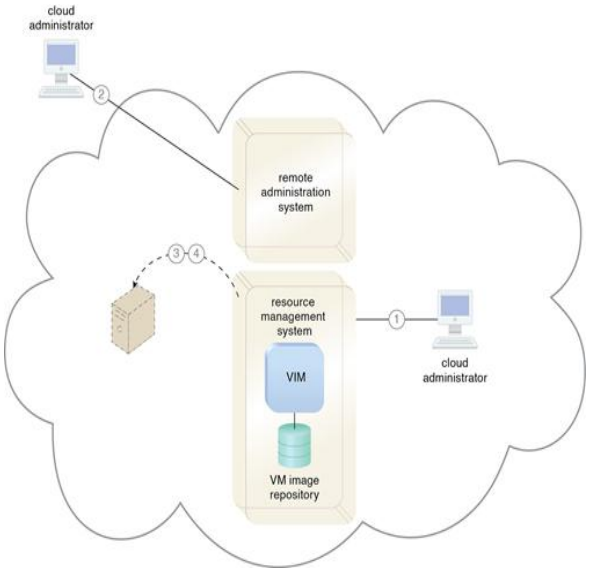

Figure 2

It consists of computing resources and Virtual Machines. The RCRP algorithm is used to provision resources. This algorithm is implemented in the cloud agent. The cloud agent will collect resources from the repository of each cloud consumer. The VMs which are stored in the VM repository should be allocated to appropriate provider. This is done by the cloud agent. The RCRP uses optimization technique to find the appropriate cloud provider. Optimization is done by calculating four uncertainty parameters viz., wait-time, idle-time, cost and profit. wait-time- The time at which the users have to wait for before getting the requested resources allocated. idle-time- The time at which the cloud consumer has to wait after allocating the requested VM to a particular loud provider.

V. RISKS IN CLOUD COMPUTING

The security risks in cloud computing must be identified by the company in order to get a clear picture about the proper internal controls and related responses that a company should take to ensure the continued smooth operation of the company without fear of data disruption or compromise. Cloud computing is now an accepted part of the array of technology available to accountants. Cloud computing can offer efficiency and cost cutting benefits. Before using cloud technology, however companies should understand the risks and security issues inherent in this new technology. By taking a systematic approach to risk assessment, including creating effective policies for cloud usage and a risk response plan, companies can take advantage of this new technology to increase operational efficiency. All organizations should have policies to establish controls to prevent and detect the unauthorized procurement and use of cloud services, regardless of management's position on venturing into cloud computing. Due to the low cost of initiating cloud services relative to traditional technology purchases, current controls such as expenditure limits may not trigger appropriate attention from management.

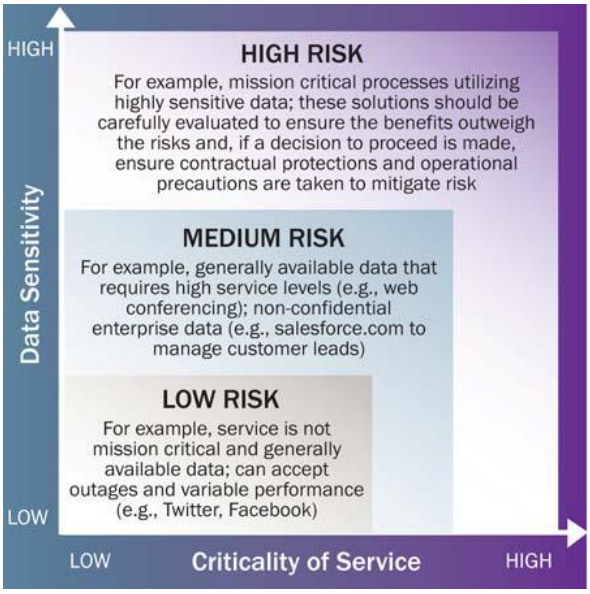

Figure 3

Deciding whether to adopt cloud computing requires management to evaluate the internal environment – including the state of business operations, process standardization, IT costs, and the backlog of IT projects – along with the external environment – which includes laws and regulations and the competition's adoption of cloud computing. As management contemplates its cloud computing position and strategies, it should address some key questions, including:

- What is management's stance on outsourcing functions?
- Does the organization anticipate rapid growth that might require using cloud solutions?
- Is the organization in a mature market that might require using cloud computing to save costs to remain competitive?
- Are the organization's operational functions and processes mature and formalized enough to allow for a change in the underlying technology platform?
- What is the capability and maturity of the organization's current IT function?
- How should the organization prepare for cloud computing?
- Should cloud computing be embraced, to capitalize on its benefits, or rejected, to avoid risks such as data breaches or noncompliance with complex e-discovery requirements?
- Who should be involved in the evaluation process, and who makes the decisions?
- How can the organization manage its risks adequately while operating in a business environment with cloud computing?

The variables to be considered when making decisions about cloud computing solutions include business processes to be supported, specific deployment models, specific service delivery models, and the specific vendors that could become service providers.

VI. CONCLUSION

Efficient management of resources at cloud scale while providing proper performance isolation, higher consolidation and elastic use of underlying hardware resources is key to a successful cloud deployment. Existing approaches either provide poor management controls, or low consolidation ratios, or do not scale well. Based on years of experience shipping the VMware DRS resource management solution and prototypes to increase its scale, we have presented some use cases for powerful controls, key challenges in providing those controls at large scale, and an initial taxonomy of techniques available to do so.

REFERENCES

[1]. B. Alexander, "Web 2.0: A New Wave of Innovation for Teaching and Learning?" Learning, vol. 41, no. 2, pp. 32–44, 2006.
[2]. R. Buyya, C. Yeo, and S. Venugopal, "Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities," in Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications (HPCC-08, IEEE CS Press, Los Alamitos, CA, USA), 2008.
[3]. I. Foster, Y. Zhao, "Cloud Computing and Grid Computing 360-Degree Compared," in Grid Computing Environments Workshop, 2008. GCE'08, 2008, pp. 1–10.
[4]. W. Forrest, "How to cut data centre carbon emissions?" Website, December2008. [Online]. Zvailable:http://www.computerweekly.com/Articles/008/12/05/233748/how-tocut-data-centre-carbon-emissions.htm
[5]. J. Koomey, "Estimating total power consumption by servers in the US and the world," Final report. February, vol. 15, 2007.
[6]. L. Wang, G. von Laszewski, "Cloud computing: a perspective study," New Generation Computing, vol. WangLYAH, to appear in 2010.
[7]. G. von Laszewski, A. Younge, X., "Experiment and Workflow Management Using Cyberaide Shell," in 4th International Workshop on Workflow Systems in e-Science (WSES 09) in conjunction with 9th IEEE International Symposium on Cluster Computing and the Grid. IEEE, 2009. [Online].
[8]. G. von Laszewski, F. Wang, "Cyberaide JavaScript: A JavaScript Commodity Grid Kit," in GCE08 at SC'08. Austin, TX: IEEE, Nov. 16 2008
[9]. P. Barham, B. Dragovic, "Xen and the art of virtualization,"in Proceedings of the 19th ACM Symposium on OperatingSystems Principles, New York, U. S. A., Oct. 2003, pp. 164–177.
[10]. VMware, "Understanding Full Virtualization, Paravirtualization, andHardware Assis," VMware, Tech. Rep., 2007. [Online]. Available: http://www.vmware.com/files/pdf/paravirtualization.pdf
[11]. Amazon, "Elastic Compute Cloud." [Online]. Available: http://aws.amazon.com/ec2/
[12]. D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L.Youseff, and D. Zagorodnov, "The Eucalyptus Open-source Cloudcomputing System," Proceedings of Cloud Computing and Its Applications, 2008.