A Survey on Resource Allocation and Monitoring in Cloud Computing

Mohd Hairy Mohamaddiah, Azizol Abdullah, Shamala Subramaniam, and Masnida Hussin

Abstract—**The cloud provider plays a major role especially providing resources such as computing power for the cloud subscriber to deploy their applications on multiple platforms anywhere; anytime. Hence the cloud users still having problem for resource management in receiving the guaranteed computing resources on time. This will impact the service time and the service level agreements for various users in multiple applications. Therefore there is a need for a new resolution to resolve this problem. This survey paper conducts a study in resource allocation and monitoring in the cloud computing environment. We describe cloud computing and its properties, research issues in resource management mainly in resource allocation and monitoring and finally solutions approach for resource allocation and monitoring. It is believed that this paper would benefit both cloud users and researchers for further knowledge on resource management in cloud computing.**

Index Terms—**Resource allocation, resource monitoring, cloud computing, resource management.**

I. Introduction

Cloud Computing, a service model was introduced to provide computing resources such as computing power, storage and bandwidth to deliver the IT services to the organization. This service model is rapidly being adopted by many organizations because it offers lots of business opportunity especially in terms of financial investment and human capital. With cloud services, the organization may opt out to set up a data center for its IT infrastructure or to procure hardware and software for its business applications as they can lease the resources from the cloud service provider. Initial providers in cloud computing such as Amazon EC2 [1], Google Apps Engine [2] , Microsoft Azure [3], SalesForce.com [4] offers great business value to the interested organizations who want to subscribe the services with a concept pay-per-use, on-demand and a defined Service level agreements (SLAs). For example, Google Apps Engine offers monthly uptime percentage: 99.00% – < 99.95% for client for its covered services. For the downtime period, they offer a period of five consecutive minutes of downtime. This value is an attractive package for the client, because they manage to control the usage of resources and demands more, faster and reliable infrastructure.

Manuscript received October 11, 2013; revised December 7, 2013.

The authors are with the Department of Communication Technology and Network, Faculty of Computer Science and Information Technology,University Putra Malaysia, 43400 UPM Serdang Selangor Malaysia (Corresponding author: Azizol Abdullah; e-mail: kelatedotcom@gmail.com, {azizol, shamala, masnida}@fsktm.upm.edu.my).

On the other hand, the resources on the cloud are pooled in order to serve multiple subscribers. The provider use multi-tenancy model where the resources (physical and virtual) are reassigned dynamically based on the tenant requirement [5]. The assigning of the resources will be based on the lease and SLA agreement, whereby different clients will need more or less amount of virtual resources. Subsequently, the growth of demands for cloud services is bringing more challenge for the provider to provide the resources to the client subscriber. Therefore, in this paper we provide a review on cloud computing which focus on resource management: allocation and monitoring. Our methodologies for this review are as follows:

- We provide a cloud computing taxonomy covers the cloud definitions, characteristics and deployment models.
- We then analyze the literatures and discuss about resource management, the process and the elements.
- We then concentrate literatures on resource allocation and monitoring. We derived the problems, challenge and the approach solution for resource allocation and monitoring in the cloud.

This paper organizes as follows: Section II introduces an overview of Cloud Computing, Section III discuss about resource management and its processes, Section IV discuss about related work with resource management in the cloud, section, Section V describes about approach solution to resource allocation and monitoring and, finally Section VI concludes the paper.

II. Overview of Cloud Computing

Cloud computing is defined as a model 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 [6] . In paper [7] refers cloud computing as:

- The applications delivered as services over the Internet
- The hardware and systems software in the datacenters that provide services that is software as a service.

Its characteristics comprise of a broad network access which has the ability to access the network via heterogeneous platforms, on demand self-service, when it provision the computing power automatically. It also has features of the service oriented which is called measures service, with its metering capability, resources can be monitored, controlled and reported and this is transparent to the provider and the customer. Furthermore, it is also equipped with an elastic feature where the scale-out or scale-in provisioning of resources can be implemented rapidly in

DOI: 10.7763/IJMLC.2014.V4.382

order to make the resources available, unlimited and can be purchased at any time by the cloud subscriber.

The main deployment models of the Cloud are the private cloud, public cloud, hybrid cloud and community cloud. The private cloud is where the physical hardware and software are set up at the client's premise whereas the public cloud is where all the resources are set up at the cloud provider. The cloud service model has three types of service model; Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). For SaaS service model, it is related to software or applications that the customer can subscribe and access from the cloud. For the developers, PaaS is the service which is suitable for them as it provide development tools than can be pledged in order to finish their chores. IaaS is the bottom layer of cloud reference model. This service model provides the resources (with computing power: CPU, memory, storage and network) and abstract into numbers of virtual machines for the cloud subscribers.

The Cloud Provider or we refer as the service provider (SP) delivers the physical computing provided by the infrastructure provider (IP). The IP offers computing, storage, and network resources required for hosting services. Their goal is to capitalize on the provider from tenants by making efficient use of their infrastructures. Possibly, by outsourcing partial workloads to partnering providers[8]. Then the SP will run the cloud software in order to provision request resources via multiple services to the customers. The Service Provider offers economically efficient services using hardware resources provisioned by infrastructure providers and it can be directly accessed by end-users or orchestrated by other service providers [8]. This provisioning process is abstracted into virtual machines by using Virtualization technology. The provisioned resources will be used by the client/subscriber to deploy its required applications in the provision platform. They have full access to the provisioned resources while the provider control on the physical hardware and services layer. This is to enable them to monitor the performance of the resources [9]. Fig. 1 below illustrates a visualization of matters of a cloud reference architecture developed by the National Institute of Standard Technology (NIST) which depicts the services flow of the cloud service model. The reference architecture[6] consists of 5 main section consists of:

Cloud provider: Person, organization or entity responsible for making a service available to Cloud Consumers

Cloud broker: An entity that manages the use, performance and delivery of cloud services, and negotiates relationships between Cloud Providers and Cloud Consumers.

Cloud consumer: Person or organization that maintains a business relationship with, and uses a service from Cloud Providers.

Cloud carrier: The intermediary that provides connectivity and transport of cloud services from Cloud Providers to Cloud Consumers.

Cloud auditor: A party that can conduct an independent assessment of cloud services, information system operations, performance and security of the cloud implementation.

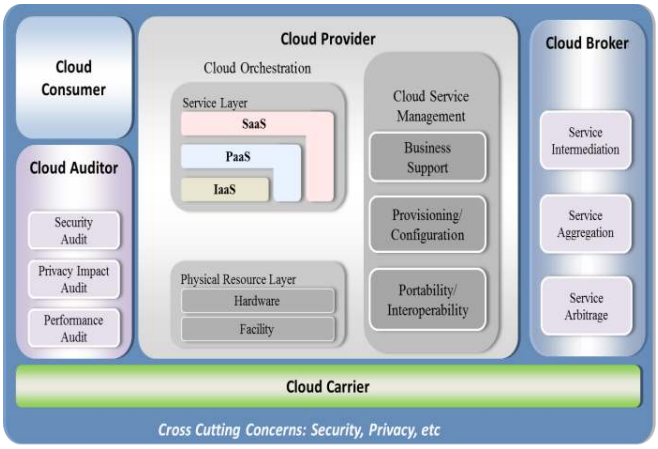

Fig. 1. Cloud reference architecture by NIST.

NIST developed the requirements of standards from data portability, service interoperability, security and integration within the cloud providers. For example, cloud consumer is reflecting to the users who subscribe the cloud services, but have a relationship with the provider. Therefore, consumers who use Platform as a Service (PaaS), need to test and deploy and manage their applications in the cloud environment. The role of provider will provide development or administration tools and middleware to the consumer for them to develop and deploy their business applications. The reference architecture is a guide for the cloud community in order to give or to obtain the best services from the cloud. Furthermore, each section also provides research challenges for the researchers to acquire knowledge advancement in the cloud computing area.

III. Resource Management at a Glance

A. Resource Management

From our perspective, research management includes resource discovery, allocation and monitoring process as illustrated in Fig. 2 below. These processes manage physical resources such as CPU cores, disk space, and network bandwidth. This resources must be sliced and shared between virtual machines running potentially heterogeneous workloads. We outline the taxonomy for resource management elements in Fig. 2 below:

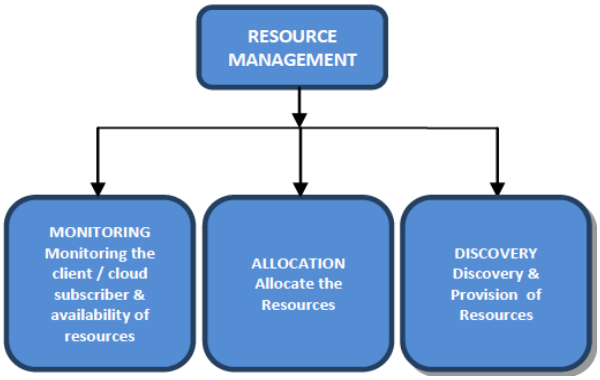

Fig. 2. Elements of resource management.

The fundamental element of resource management is the discovery process. It involves searching for the appropriate resource types available that match the application

requirements[10]. The process is managed by the cloud service provider. This process is being taken by the resource broker or user broker to discover available resources. Discovery consists of detailed description of resources available. According to [11], resource discovery provide a way for a resource management system (RMS) to determine the state of the resources that are managed by it and other RMSs that interoperate with it. The resource discovery works with dissemination of resources to provide information about the state of resources to the information server.

The Allocation process is the process of assigning an available source needed cloud applications over the internet [12]. These resources are allocated based on user request and pay-per-use method. In this process, scheduling and dispatching method is being applied to allocate the resources. The scheduler will schedule assigned resources to the client. Then, the dispatcher will allocate the assigned resources to the client.

Resource monitoring as defined in paper [13] is a key tool for controlling and managing hardware and software infrastructures. It also provides information and Key Performance Indicators (KPIs) for both platforms and applications in cloud to be used for data collection to assist in decision method of allocating the resources. It's also one key component to monitor the state of the resources in the event of failure whether at the physical layer or at the services layer.

From our point of view, we would like to group allocation and discovery as the provisioning process, while monitoring is one single process. However these 3 processes are interrelated with each other in order to provision resources to the users. Next sub-section will elaborate about the resource provisioning and the resource monitoring process flow in details.

B. Resource Provisioning and Monitoring Process

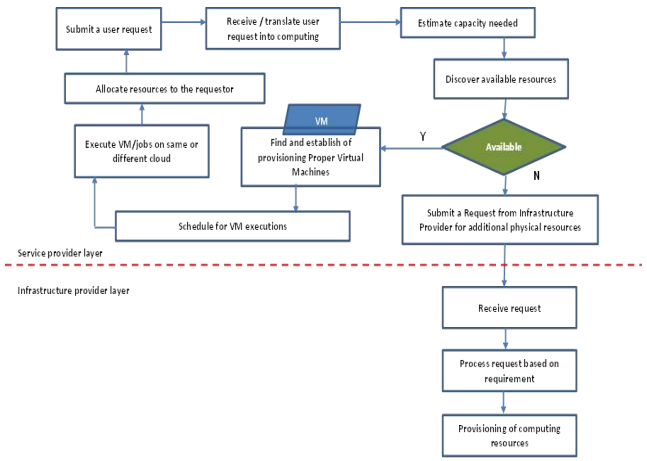

Fig. 3. Resource discovery and allocation process.

In Fig. 3, we outline the resource provisioning processes which consist of discovery, allocation and monitoring. The discovery and allocation process is done at the service provider layer while at the infrastructure provider is only processing the physical resource request by the service provider. There is no direct interaction between the consumer and the infrastructure provider. The service provider will provision Virtual machines from the existing pool of physical resources. If in the event of inadequate of resources, it will request from the infrastructure provider.

In Fig. 4 below depicts the monitoring process undertaken by the service and infrastructure provider. We conclude that both parties play a major role for monitoring available resources that had been established to make sure that the service conveyed at the desired service level agreements. In the event of a crash or over-provision, the providers have a right to hold the scheduler process and migrate or expand the virtual resources to other available physical hosts.

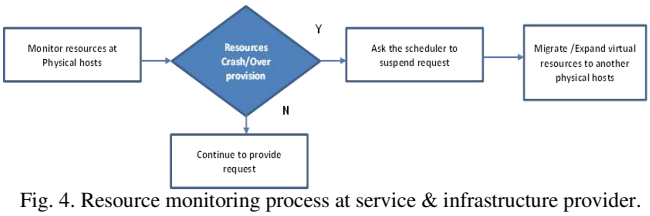

Fig. 4. Resource monitoring process at service & infrastructure provider.

In Fig. 5 is the ongoing monitoring process done by the service provider. This process is being done to optimize the resources available. If there is underutilized computing power, the service provider will perform a load balancing exercise that will result a better utilization among the available resources.

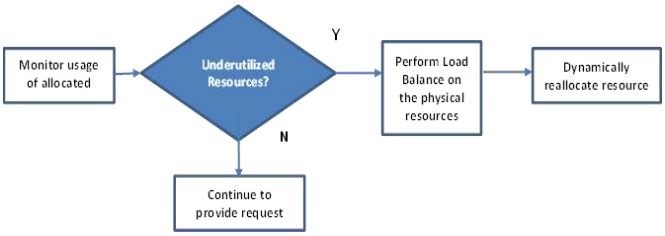

Fig. 5. Resource monitoring with load balancing approach.

The processes elaborate in figures above bring the challenges to the cloud provider to provide the best services. In addition, it had produced numerous researches in order to solve resource allocation and monitoring problems. Moreover, it developed new knowledge to the cloud computing field. The cloud subscriber as the client wants a service on-demand, dynamic, able to utilize change in a real-time environment and accommodate the user's need. For example, event of resources become over utilized or vice versa cannot be tolerated. The infrastructure and service provider have to meet the Service Level Agreement (SLA), where they have an ability to allocate resources for mix environment or application to the fulfill mix of SLA of applications or users [8].

In the next Section, we will elaborate about related work in resource management which includes the issues of allocation and monitoring process in cloud computing.

IV. Related Work in Resource Management

The services from the cloud are used to run a workload from such a complex system or a high transaction business application. From the literature outlines numerous problems in resource management in cloud computing not limited to the computing power availability but also on the other elements such as cost and energy consumption. These problems happen whether at the abstraction, service or

physical layer. At the physical resource layer, the Infrastructure Provider (IP) is responsible for hardware and software resource configuration, provisioning/allocation, allocation and monitoring for the user. This is the main functional area of the physical resource layer at the Infrastructure as a Service (IaaS) service layer. Nevertheless, the cloud users are still having problem to gain the resource in a guaranteed computing resources on time, especially for modern e-Science and high Technology applications which require a high performance infrastructure [14] to support the complex application requirements. This resource contention also happens in high computational hybrid scientific applications [15] at the cloud, in the event of the resources are becoming exhausted and in limited usage to the users in the cloud. With the requirement of the high workloads to be processed, the users require a service on demand (real-time) and ability to utilize change. Therefore the need to put the current infrastructure readily to be agile [16] is a must for the service and infrastructure provider.

The source of the computing power for the infrastructure provider comes from consolidated data centers. If the scenario of contention for the server consolidation in a consolidate data center on shared platform resources [17] happens, this will result the physical resources becoming drained, and this will trouble the whole organization of infrastructure provider for the shortage of capitals. Therefore, the Infrastructure Provider needs a more agile in infrastructure for resource pooling and virtualization technology to simplify the service provisioning for on-demand resources and to improve their service offering and business model to Virtual Infrastructure provisioning [14]. The resources that both (SPs & IPs) have to provide to the user should then be optimized to cope with current demand from the users especially for real-time or high transaction applications.

J. Espadas *et al.* in paper [18] discuss about over and under provisioning of cloud resources, although the peak loads can be successfully predicted, thus without having an effective elasticity model, costly resources are wasted during nonpeak times (under-utilization) or revenues from potential customers are lost after experiencing a poor service. The control of the leased resources (virtual machines, storage) falls under the client responsibilities. The provider only monitors the usage of the whole resources (CPU, memory, bandwidth etc.), while the user or the subscriber monitors the business transactions, therefore there should be an advisory system which simultaneously linked between the applications from the user site and the physical monitoring resources. If there is no reliable monitoring system, the over provisioning of computing powers will be a waste or a risk for the consumers to deploy and maintain their business applications in cloud environments.

V. Vinothina *et al.* [12] surveyed that resource contention, scarcity of resources, resource fragmentation and over provisioning should be avoided, so that allocation and usage of the resources will be optimal. There is a need of a new strategy to find the cause the computing resources become exhausted and out of capacity to serve the cloud subscriber. A new solution must provide a new method that will trigger and enhance the computing power to provide the required resources on time to users without starving its current capacity. The proposed strategy may be suitable at the infrastructure provider site since the main resources for the cloud is controlled by the service provider.

Resources or computing power is an abstracting process from physical layer where it is residing in Infrastructure Provider's data center. Issues of maintaining a data center are quite an expensive with a high energy costs and huge carbon footprints. The Issues of energy usage in resource management for the cloud operations [19] is still wandering along even though the performance of the computing increased by reducing time for processing [20] but the power efficiency becoming lower. R. Basmadjian *et al*. proposed power consumption models based on prediction in order to save the energy cost in cloud environment [21] in the private cloud computing data center. The prediction is based on power calculation module which depicts the power consumption for every component in a data center such as servers (and its component), power module, storage and others in a private cloud. Thus, the prediction model can be improved and enhanced in public cloud which encompasses more heterogeneous components and will benefit the inter-cloud service provider and the consumer as well.

In paper [22] states that several scheduling algorithms with different allocation strategy may produce a saving on the energy consumption in the operational kind but with condition whereby on-demand requests do not face the unpleasant blocking probability. As a result study for power consumption will contribute to energy efficiency and cost savings. From the financial view, the provider will get a cost reduction on the expenses for the utility while the user will benefit the fee reduction towards the usage of resources. On the other hand this will also give more support to the green initiative which is being initiated by many countries. With this new solution it should also make the computing power to be scalable and rapidly elastic without increasing the power usage in the current cloud infrastructure and improve the power efficiency.

Based on the stated problems, we believe that the problem occurs at the allocation of the resources, which this will impact the services to the client. Some of the allocation scheme is not efficient enough to cater the user's need. Without a good monitoring system, especially at the resource provider, there is a possibility of a failure in delivering the service, even though the cloud has its high availability features. Our next section will elaborate more about existing solution or mechanism been done for resource allocation and monitoring in delivering the services for cloud users.

V. Approach in Resource Allocation and Monitoring

A. Resource Allocation

In resource allocation, scheduling algorithm is one of the widely use approach in order to optimize the usage of resources in cloud such as maximizing CPU and memory utilizations. As we know, in a uniprocessor only one process is running. Then this process migrates between various scheduling queues. The process of selecting processes to be

run from among these queues is carried out by a scheduler [23]. The aim of processor scheduling is to assign processes to be executed by the processor. This scenario affect the performance of the system, especially when there is a delay during a situation of determining which process to be run by the processor and which will wait. One of the earliest researches used scheduling algorithm, where the first come first serve [8] method is applied, by determining the smallest number of servers required to meet the SLAs for shared allocation and dedicated allocation resources. Consequently, many of the researches embark on scheduling algorithm such as in [24] proposed an energy-conscious scheduling algorithm, while in paper [20], [25], [26] use gang scheduling resource allocation, In paper [25] implemented multi-tiered resource scheduling allocation scheme using virtual clusters to develop a more efficient algorithm. In addition, it is also to determine the preferred allocation strategy of resources. Based on the research works, in paper [24] it was proven that its proposed algorithm contribute to reduce the energy consumption in cloud resources with the adoption of Dynamic Voltage Scaling (DVS) technique.

In [26] the algorithm improves the waiting time especially when the workload had increased. However, the works only run in one scenario and not for different scenarios with variety of workload especially for High Performance Application (HPC). M. Stillwell *et al.* [25] conducted a study on resource allocation using virtual clusters which run multiple competing tasks. They have established optimal algorithms which compute allocations of task for different number of nodes. Users can choose which algorithm suits the allocation to gain the best resource available. For parallel tasks (a single workload is divided into multiple task), they consider a homogenous multiple jobs. These jobs have already had its identical resources. The work imposes an equal allocation for every task (although some might require lower allocation) by scheduling the tasks using sequential job algorithms. Thus, the result will be better by applying load balancing approach since it is tested in parallel applications which run on a virtual cluster environment. For dynamic workload, they work on pragmatic formulation for the adaptation of allocation of the resources. However, this solution is based on the current allocation whereby the features of dynamic workload may be different and changes rapidly. Yet, the proposed method proved to be very fast in order to adapt the change, and rapid allocation of the needed resources.

Do *et al.* works in paper [22] proved that several scheduling algorithm had assisted various resource allocation scheme to be implemented for the infrastructure provider. The impact of scheduling implemented in [8] creates a scenario when a very small process will have to wait for its turn to utilize the CPU. A short process located behind a long process will result in lowering the CPU utilization. Moreover, it will surely be a waste and the cost also will be higher to the client. From the point of resource allocation, however, by implementing prioritization of any workload, any short process should be grouped in one workload. Therefore it should be run together so that the workload will make use all of the available resources at the same time. This will result a better CPU utilization. On the contrary, longer process which requires a bigger source should be dispatched to another scheduler. In addition, the scheduler must able to run automatically to allocate available resources to run the workload faster.

Paper [27] focus on aggressive and selective backfilling algorithm combined with brokering strategies to resolve the problem of resource failures due to the increasing functionality and complexity of hybrid Cloud computing. They aim to improve the Quality of Service (QoS) for the users' requests by utilizing the public Cloud resources. However, based on the work, the proposed algorithm is only measured in independent task and was not tested in a mixed task to prove the scalability of the algorithm. A work in [28] also imposed scheduling algorithm using min-min and list scheduling algorithm in pre-emptible task. The scheduler's works are based on the updated information about actual task executions such as feedback information. The dynamic procedure with updated information does not impact the application execution time significantly. But then, their proposed algorithm proved to have a shorter average execution time for allocating the resources. On the contrary, there is a risk whether the feedback information is reliable for the scheduler to count on. There should be another reliability checking mechanism which should be faster in terms of resource allocation. Therefore, in the event of false information, the scheduler will eliminate the faulty to allocate the resources required by the cloud subscriber especially in a heterogeneous workload. However, from their work, they proved that their proposed algorithms can reduce energy consumption compared to the First Come First Serve algorithm.

Other computer science domain is also being implemented for resource allocation mechanism. For example, knowledge management (KM) domain is being applied to study the adaptive actions by the virtual machine [29]. They developed KM agnostic- simulation engine to simulate executed actions and evaluate quality response to the workload. The works then construct a decision system and structured all reallocation process using knowledge management techniques. The proposed work had impacted the event for the provisioning of the virtual machine. The result of the work had improved the SLA for resource allocation seamlessly.

Basically there are several theories being adopted to formulate a new framework or model for many resource allocation researches in the cloud. In power consumption field, most of the previous researches incorporate resource management with the adoption of power consumption theory dynamic voltage and frequency scaling (DVFS). DVFS is a commonly-used power-management technique where the clock frequency of a processor is decreased to allow a corresponding reduction in the supply voltage [30]. This theory scheme can be used to reduce the power consumption which applied to a physical server [24] but it depends on the load and a situation which there is no virtual machine allocated to the physical server. The power scheme is also being adopted to develop an adaptive model-free approach in multi-tier cloud to resolve the energy issues in cloud [20]. However, works in paper [30] revealed that, the latest generations of processors in an idle state, has clearly more impact on energy savings in servers, and dynamic voltage

and frequency scaling has only a minimal impact on the reduction of energy consumption. In addition, operators provisioning IaaS have a limited scope to change the CPU frequency of processors since the computing capabilities explicitly expressed in equivalent CPU frequency. Moreover it is normally offered to customers in practice. They also stated that the most viable method to reduce the energy consumption is to increase of the utilization of server farms whereby switching off unused servers [29].

Several studies are also adopting another theory, one is the optimization theory. In paper [20], they proposed an online admission control, routing and resource allocation for a virtualized data center on time varying workload and heterogeneous applications. The optimization theory is adopted to unravel the problem of maximizing a joint utility of the long term throughput of the hosted applications and the average of the power expenditure in a virtualize data center. However, it will better if the solution uses estimation predictive theories and statistic of resources coming, which this will analyze the resources and allocation method should be faster.

The development of a framework and model is also one of the approaches in resource allocation. We conclude that some study will develop an algorithm for its framework or model; some will develop and use service oriented for the verification method in their methodology. Paper [14] embark on service oriented architecture and develop an architectural framework to combine on demand network and grid/cloud resources for service provisioning. The proposed Infrastructure Services Modeling Framework (ISMF) is part of the total solution of the on-demand architectural framework for on demand service provisioning to support e-science application and a high performance computing resources. The ISMF is formulating from a resources modeling where the main function of the model is able to describe physical resources in detail and abstracting the virtual resource layer. The description format which is based on semantic technology, will allow the extension and dynamic change of the resources in the framework.

Paper [15] formulates a model by combining and integrates the cloud and grid resources. Their solution is by outsourcing workload to the cloud when the resources on the grid become exhausted. Meta scheduling analysis is used to analyze the current state of the grid and schedule the integration with the cloud. As a result, the proposed method had utilized the resources in the hybrid infrastructure (cloud and grid).The proposed model in paper [31] is being developed using time varying workload and DVFS theory in a multi-tier cloud platform which resulted in saving the energy and cost. In paper [18], the researcher proposed a model using multi-tenancy characteristic to solve the over and under provisioning of the cloud resources problem. They used a tenant-based isolation approach, which encapsulates the execution of each tenant. A tenant-based load balancing, this distributes requests according to the tenant information also being implemented. To determine the number of VM instances needed, they used a tenant-based VM instance allocation. The method is based on VM capacity and tenant context weight. From these approaches, there is an improvement in the server based hours for virtual machine provisioning, which reduce the response time and cost. Their work proved to be success in a multi-tenancy environment and creating a cost-effective tenancy.

We outline a mind mapping diagram to summarize the proposed mechanism with its benefits concerning about resource allocation in Fig. 6 below.

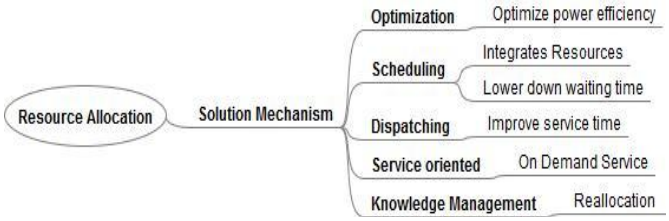

Fig. 6. Resource allocation and its mechanism.

B. Resource Monitoring

In this sub-section, we review and analyze various monitoring mechanism that is being implemented. There are several studies being conducted to address the heterogeneous and monitoring resources which denote an intuitive representation of a running cloud by focusing on common monitoring concerns.

Raw monitoring of data is gathered by multiple monitoring techniques to create a more intuitive cloud profile. Any changes in the real running system will be reflected in the model immediately [32]. Therefore human roles can inspect runtime status of the cloud via the model, instead of investigating disordered runtime data. Monitoring also reflect the service level management in cloud services. An architectural model proposed in [9] uses service monitoring and process of IT service management to realize management of Cloud services resources. The paper provides a foundation for technology to ensure the quality of service of Cloud services. One study conducted for resource monitoring apply a method of event trigger [33] to monitor the current state of the resources and integrate with fixed polling of resources.

There is not much significant study about the failure of resources itself. The current software technology in the market such as VMWare [34] has the features of load balancing or fault tolerance, but this applies only to the physical hosts not to the virtual servers. Furthermore, if a resource failure such as virtual machine is not identified and addressed in time, it could cause permanent damage to the resources, bring down critical IT services, causing massive data loss, and propel maintenance costs. This will herald the SLAs with the client.

High availability (HA) feature is also one of the features currently available in virtualization software. This feature is used by the Service Provider for resource failure or disaster recovery for bigger event failure. Although the HA features do make a positive outcome, especially when the situation of resource contention arise but, the SLA with the client is cramping the provider. A fault tolerance mechanism should be incorporated with the monitoring suite for the resources.

In paper [35] used a Discrete Time Markov Chain as a fault detection mechanism in their proposed solution. They also incorporate a long-term condition monitoring of equipment in the complex systems. Their works intend to describe states and apply directly the state transition in the failure existence. They had proved to make transition of the

application of the unsuspected potential failures, which might be difficult to detect through a non-automated means. Application environment which uses cloud facilities are dynamic and have various constraint. Therefore, this solution proved that detecting and fixing any problem on time can increase service uptime and enhance client satisfaction.

In paper [36] proposed a decomposition model that applies scheduling on traces for server consolidations on shared resources platform. It introduces the Virtual Platform Architecture (VPA) to estimate the resources, and enable transparent of shared resources management. The model also incorporates a heuristic approach in order to get the sampling data for analysis. Monitoring scheme of cache space and memory bandwidth is one of the major measurement parts of the model. The monitoring scheme in the VPA is associated and enforced with the client. They had shown that monitoring is needed for shared resource management and performance isolation in a consolidation scenario.

Performance metrics also play a major role in monitoring techniques for resource management. One of the most common metrics used is time. Time is used to show the improvement of the SLA based on scheduling algorithm, expanding utilization of resources, increasing of the quality of service, scalable and elastic the infrastructures, improve the waiting time, and improving the processing time. In paper [37] incorporate time via a run-time monitor for mapping rules based on the services being provisioned. They introduce a framework to manage the mappings of the Low-level resource Metrics to High-level SLAs called LoM2HiS framework. The framework assists in autonomic SLA management and enforcement. The framework is capable of detecting future SLA violation threats based on predefined threat thresholds.

Fig. 7 below depicts the summarize mechanism for resource monitoring and its solution based on our analysis. Cloud monitoring is very important for cloud provider because it supports the process of provisioning of resources.

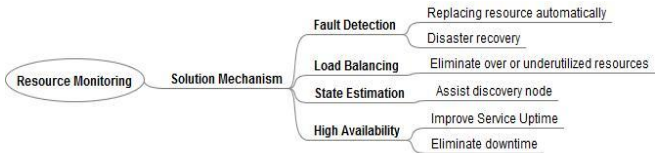

Fig. 7. Resource monitoring and its mechanism.

VI. Conclusion

In this paper, we have discussed about resource management in general, the existing resource allocation and monitoring strategies from the current research works. This paper has summarized different method (algorithms technique) and theory which being used to formulate framework and model, derived to provide a better resource allocation and monitoring process in terms of a better performance, competitive and efficiency to meet the required SLA, improved the resource performance and lowered the power consumption. We hope this paper will motivate researchers to explore and formulate a new mechanism to solve issues in allocating and monitoring resources in cloud computing.

Acknowledgment

The authors wish to thank Faculty of Computer Science and Information Technology for its continuous support in finishing this paper.

References

[1] Amazon Elastic Compute Cloud (EC2). [Online]. Available: http://www.amazon.com/ec2/
[2] Google App Engine. [Online]. Available: http://www.appengine.google.com
[3] Windows Azure. [Online]. Available: http://www.microsoft.com/azure
[4] Salesforce CRM. [Online]. Available: http://www.salesforce.com/platform
[5] T. Dillon, C. Wu, and E. Chang, "Cloud computing: issues and challenges," in *Proc. 2010 24th IEEE International Conference on Advanced Information Networking and Applications*, IEEE, 2010, pp. 27-33.
[6] R. B. Bohn, J. Messina, F. Liu, J. Tong, and J. Mao, "NIST cloud computing reference architecture," in *Proc. 2011 IEEE World Congress on Services*, 2011, pp. 594-596.
[7] M. Armbrust, A. D. Joseph, R. H. Katz, and D. A. Patterson. (February 2009). Above the clouds : A berkeley view of cloud computing. *Electrical Engineering and Computer Sciences*. 53(UCB/EECS-2009-28). [Online]. pp. 7-13. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html
[8] Y. Hu, J. Wong, G. Iszlai, and M. Litoiu, "Resource provisioning for cloud computing," in *Proc. the 2009 Conference of the Center for Advanced Studies on Collaborative Research*, 2009, pp. 101–111.
[9] Y. Sun, Z. Xiao, and D. Bao, "An architecture model of management and monitoring on cloud services resources," in *Proc. 2010 3rd International Conference on Advanced Computer Theory and Engineering*, 2010, pp. 207-211.
[10] R. Buyya and R. Ranjan, "Special section: Federated resource management in grid and cloud computing systems," *Future Generation Computer Systems*, vol. 26, no. 8, pp. 1189–1191, June, 2010.
[11] K. Krauter, R. Buyya, and M. Maheswaran, "A taxonomy and survey of grid resource management systems for distributed computing," *Software: Practice and Experience*, vol. 32, no. 2, pp. 135–164. 2002.
[12] V. Vinothina, R. Sridaran, and P. Ganapathi, "A survey on resource allocation strategies in cloud computing," *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 6, pp. 97–104, 2012.
[13] G. Aceto, A. Botta, W. De. Donato, and A. Pescapè, "Cloud monitoring: A survey," *Computer Networks*, vol. 57, no. 9, pp. 2093–2115, 2013.
[14] Y. Demchenko, J. V. der Ham, V. Yakovenko, C. D. Laat, M. Ghijsen, and M. Cristea, "On-demand provisioning of cloud and grid based infrastructure services for collaborative projects and groups," in *Proc. 2011 International Conference on Collaboration Technologies and Systems*, 23-27 May, 2011, pp. 134-142.
[15] A. Calatrava, G. Molto, and V. Hernandez, "Combining grid and cloud resources for hybrid scientific computing executions," in *Proc. 2011 IEEE Third International Conference on Cloud Computing Technology and Science*, 2011, pp. 494-501.
[16] V. Sarathy, P. Narayan, and R. Mikkilineni, "Next generation cloud computing architecture: enabling real-time dynamism for shared distributed physical infrastructure," in *Proc. 2010 19th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises*, 2010, pp. 48–53.
[17] R. Iyer, R. Illikkal, O. Tickoo, L. Zhao, P. Apparao, and D. Newell, "VM3: Measuring, modeling and managing VM shared resources," *Computer Networks*, vol. 53, no. 17, pp. 2873-2887, 2009.
[18] J. Espadas, A. Molina, G. Jiménez, M. Molina, R. Ramírez, and D. Concha, "A tenant-based resource allocation model for scaling software-as-a-service applications over cloud computing infrastructures," *Future Generation Computer Systems*, vol. 29, no. 1, pp. 273-286, 2013.
[19] X. Wang, Z. Du, and Y. Chen, "An adaptive model-free resource and power management approach for multi-tier cloud environments," *Journal of Systems and Software*, vol. 85, no. 5, pp. 1135-1146, 2012.
[20] R. Urgaonkar, U. C. Kozat, K. Igarashi, and M. J. Neely, *Dynamic Resource Allocation and Power Management in Virtualized Data Centers*, 2010, pp. 479–486.
[21] R. Basmadjian, H. De Meer, R. Lent, and G. Giuliani, "Cloud computing and its interest in saving energy: the use case of a private

cloud," *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 1, no. 1, pp. 5, 2012.
[22] T. V. Do and C. Rotter, "Comparison of scheduling schemes for on-demand IaaS requests," *Journal of Systems and Software*, vol. 85, no. 6, pp. 1400–1408, 2012.
[23] Types of Scheduling-Go4Expert. [Online]. Available: http://www.go4expert.com/articles/types-of-scheduling-t22307/
[24] Y. C. Lee and A. Y. Zomaya, "Energy conscious scheduling for distributed computing systems under different operating conditions," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 8, pp. 1374-1381, 2011.
[25] M. Stillwell, D. Schanzenbach, F. Vivien, and H. Casanova, "Resource allocation using virtual clusters," in *Proc. 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, (Section III), 2009, pp. 260–267.
[26] I. A. Moschakis, and H. D. Karatza, "Evaluation of gang scheduling performance and cost in a cloud computing system," *The Journal of Supercomputing*, vol. 59, no. 2, pp. 975–992, 2010.
[27] B. Javadi, J. Abawajy, and R. Buyya, "Failure-aware resource provisioning for hybrid Cloud infrastructure," *Journal of Parallel and Distributed Computing*, vol. 72, no. 10, pp. 1318–1331, 2012.
[28] J. Li, M. Qiu, Z. Ming, G. Quan, X. Qin, and Z. Gu, "Online optimization for scheduling preemptable tasks on IaaS cloud systems," *Journal of Parallel and Distributed Computing*, vol. 72, no. 5, pp. 666–677, 2012.
[29] M. Maurer, I. Brandic, and R. Sakellariou, "Adaptive resource configuration for Cloud infrastructure management," *Future Generation Computer Systems*, vol. 29, no. 2, pp. 472–487, 2013.
[30] E. Sueur, G. Heiser, "Dynamic voltage and frequency scaling: the laws of diminishing returns" in *Proc. the 2010 Workshop on Power Aware Computing and Systems (HotPower'10)*, Vancouver, Canada, October 2010, pp. 1–5.
[31] S. Srikantaiah, A. Kansal, and F. Zhao, in "Energy aware consolidation for cloud computing," *Power Aware Computing*, 2008.
[32] J. Shao, H. Wei, Q. X. Wang, and H. Mei, "A runtime model based monitoring approach for cloud," in *Proc. 2010 IEEE 3rd International Conference on Cloud Computing*, 2010, pp. 313-320.
[33] Y. Zhu, and W. Xu, "Research of grid resource monitoring based on event-trigger and fixed polling," in *Proc. 2010 International Conference on Financial Theory and Engineering*, 2010, pp. 108–111.
[34] VMWare. [Online]. Available: http://www.vmware.com/ap/cloud-computing/overview.html
[35] C. Dabrowski and F. Hunt, "Identifying failure scenarios in complex system by perturbing markov chain analysis models," in *Proc. the 2011 Pressure Vessels & Piping Division Conference*, 2011, pp. 1–24.
[36] R. Iyer, R. Illikkal, O. Tickoo, L. Zhao, P. Apparao, and D. Newell, "VM3: Measuring, modeling and managing VM shared resources," *Computer Networks*, vol. 53, no. 17, pp. 2873-2887, 2009.
[37] V. C. Emeakaroha, I. Brandic, M. Maurer, and S. Dustdar, "Low level Metrics to High level SLAs - LoM2HiS framework: Bridging the gap between monitored metrics and SLA parameters in cloud environments," in *Proc. 2010 International Conference on High Performance Computing and Simulation*, 2010, pp. 48-54.

Mohd Hairy Mohamaddiah is currently a PhD candidate in Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, (UPM). He has received his master degree in Computer Science from National University of Malaysia (UKM) in 2010. His research interests include cloud computing and resource management.

Azizol Abdullah is a senior lecturer and head of Department at Department Communication Technology and Network, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. He had received his PhD degree from the Universiti Putra Malaysia(UPM) in 2010. He obtained his master of science in engineering (telematics) from the University of Sheffield, UK in 1996. His main research areas include grid and cloud computing, peer-to-peer computing, wireless and mobile computing, computer networks, Computer Support Collaborative Workgroup (CSCW) and Computer Support Collaborative Learning (CSCL).

Shamala Subramaniam is an associate professor at the Department Communication Technology and Network, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. She received the B.S. degree in computer science from University Putra Malaysia (UPM), in 1996, M.S. (UPM), in 1999, PhD (UPM) in 2002. Her main research interests are computer networks, simulation and modeling, scheduling and real time system.

Masnida Hussin is a senior lecturer at the Department Communication Technology and Network, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM). She had received her PhD degree from the Universiti Of Sydney Australia in 2012. She obtained her master of science from the University Putra Malaysia (UPM), in 2006. Her main research interests are computer networks, energy efficient, grid computing and cloud computing..