Artificial Intelligence (AI) Backed Cloud Resource Management Approach for Infrastructure as a Service (IAAS)

Article by Md. Shahidul Hasan[1], Balamurugan E[2], Mohammad Shawkat Akbar Almamun[3], Sangeetha K[4]
[1]*Research Scholar, Texila American University*
[2,4]*University of Africa, Nigeria*
[3]*Research Scholar, Texila American University*
E-mail: hasan_1027@yahoo.ca[1], rethinbs@gmail.com[2], liks18@gmail.com[3]

Abstract

Cloud Computing (CC) and Artificial Intelligence (AI) marks the dawn of a new era of transformation, where customers can avail resources from service providers, that offer users or machines pay per use computers as virtual machines, raw (block) storage, firewalls, load balancers, and network devices enabling smart solutions; promptly, efficiently and economically. The resulting application from the twin-technologies of cloud computing and artificial intelligence could be combined to significantly enhance resource management services such as allocation, provisioning, requirement mapping, adaptation, discovery, estimation, and modeling. The conclusion that follows is scalability, quality of service, optimal utility, reduced overheads, improved throughput, reduced latency, specialized environment, cost effectiveness and simplified interface. This study aims to improve the performance of AI in cloud resource management for best optimization. The rest of the paper is organized as follows as Introduction to Artificial intelligence, cloud computing, Review of Literature, Resource Management in OpenStack, Issues and challenges and conclusion.

Keywords: *Artificial Intelligence, Cloud Computing, OpenStack, Resource Management*

Introduction

In today's business environment, Cloud Computing is viewed as a comparatively young discipline which is driven by practical advances, tendencies and applications to transform the traditional network communication architectures into a new singular model, aiming to improve agility and reduce operation costs. Although we find a great many research studies explore, solution and service oriented fundamental problems in cloud computing. Most of the cloud computing research struggles to adopt the principles of quantitative and/or qualitative paradigms in ubiquitous access to shared pools of configurable system resources and higher-level services that can be rapidly provisioned with minimal management effort, often over the Internet. However, due to the huge amount of data from storage, transactions and connecting devices that become new sources for competitive advantage to support and sustain in highly competitive environment, the cloud computing services often fail to provide full control on security, storage type and size. Some of the biggest challenge for a cloud provider are the issues of equipment failure, availability outages, loss of internet connection as well as provider outage. There is certainly a need for more research on cloud, considering the preparations of research patterns as well as the conjectural and trans-disciplinary foundations of cloud computing as a practical solution. Cloud Computing is a modern computing paradigm that is providing IT infrastructure and a very essential requirement for the IT companies. Cloud Computing providing essential service i.e.

1. Infrastructure as a Service (IaaS),
2. Network as a Service (NaaS),
3. Platform as a Service (PaaS),
4. Software as a Service (SaaS).

IaaS refers to a combination of hosting, hardware provisioning and basic services needed to run a cloud. PaaS refers to the provision of a computing platform and the provision and deployment of the associated set of software applications (called a solution stack) to an enterprise by a cloud provider.

Software as a Service (SaaS) is a software distribution model in which applications are hosted by a vendor or service provider and made available to customers over a network. Cloud computing is a model for enabling ubiquitous, on-demand network access, to a shared pool of configurable computing resources such as network, servers, storage, applications, and services, that can be rapidly provisioned and released with minimal management effort. Cloud clients can access and use the services of cloud applications using browsers, mobile devices, while all the data as well as software is stored on servers at a remote location, which are also used to perform all the heavy duty processing.

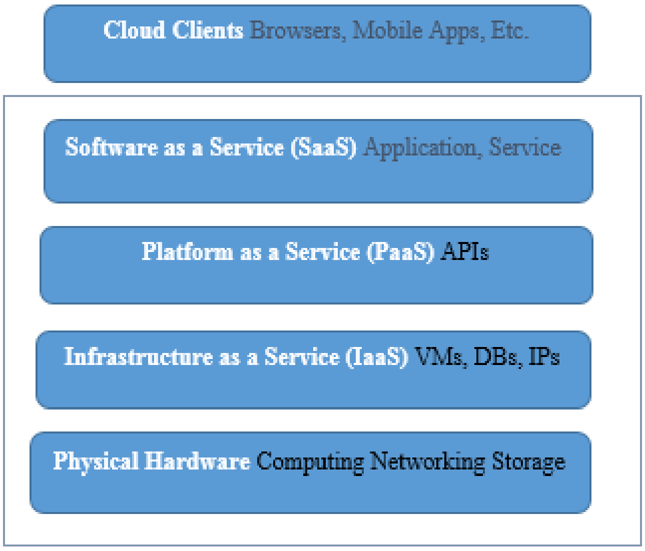

Cloud Clients Browsers, Mobile Apps, Etc.
Software as a Service (SaaS) Application, Service
Platform as a Service (PaaS) APIs
Infrastructure as a Service (IaaS) VMs, DBs, IPs
Physical Hardware Computing Networking Storage

Figure 1. Cloud computing layered model

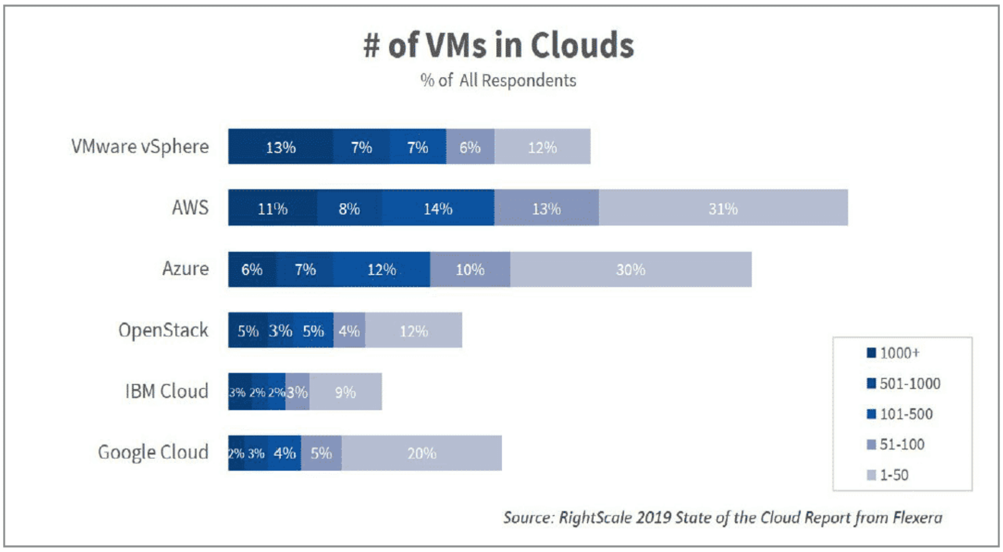
Literature

A significant number of papers have appeared in the literature examining IaaS for cloud computing, AI approaches on global applications and Cloud Resource allocation Issues and challenges for OpenStack. In recent years, extensive efforts have been put into the research of performance modeling, for virtualized systems presented a new resource management scheme, that integrates the Kalman filter into feedback controllers for dynamically allocating the CPU resources of multi-tier virtualized servers.

Sunil kumar S et al (2014) proposed a survey on resource management issues related with IaaS with cloud computing. In the recent times, AI has gained the critical-mass necessary for it to become the center of attention for technology investment and we see that bigger players are heavily investing in various AI technologies. Cloud computing and artificial intelligence technologies are introduced to various cloud platforms to improve performance, productivity and to increase satisfaction. Adopting these CC and AI technologies in telecom sectors poses a huge challenge in collecting of techniques inspired by natural intelligence. Nonetheless, using CC and AI will improve several areas performance and productivity, reduce costs, improve sustainability, and increase users' satisfaction, retention and loyalty in the long run. However, new threats may have to be taken care of, for example, security, privacy, legal conflicts and risky reputation, if it is used inappropriately.

Firstly, AI has languished itself as an "intangible brain in a jar", separated from the real world. Today we have reached the point where digital channels are becoming the standard; providing the brain with all the senses and members it needs, where customers and suppliers can communicate with each other electronically, substituting the human interface in the loop to make decisions. AI supports IT team, by detecting and predicting system failure, and thereby provides instant corrective action. Furthermore, AI can tackle customer service, with its highly complex computing skills, it can efficiently support instant online trouble shooting from a virtually remote station. Secondly, as data is made available and accessible when and where it is required, the new ("digital fuel") data pools are popping up everywhere and Application Programming interfaces (API)'s are crawling the internet available to all who seek. But such access to streams of data, is generating a huge yet perceptive "digital exhaust".

In a research, it was found that 90% of companies are already on the cloud. This stat comes to show the cloud is already mainstream in 2019. Furthermore, experts say 60% of workloads will be running on a hosted cloud service in 2019. For reference, the cloud hosted 45% of workloads in 2018. (Source: Flexera). Forbes suggest that 83% of enterprise workloads will be in the cloud by 2020. The prediction is that 41% of enterprise workload will be run on public cloud platforms by 2020. Another 20% will be private-cloud-based, while 22% will rely on hybrid cloud adoption. Further data cited by Forbes, suggests only 27% of workloads will be on premise by 2020. This would spell a 10% drop in absolute terms in just one year – as the same number for 2019 is at 37%. Cloud adoption trends suggest there's an advantage to using both public and private cloud solutions as this gives more flexibility and variety of options. Just 22% use the public cloud exclusively, and only 3% use a private one exclusively. According to Right Scale's annual State of the Cloud Report for 2019, 91% of businesses use public cloud and 72% use a private one. Most enterprises actually utilize both options – with 69% of them opting for a hybrid cloud solution



Source: RightScale 2019 State of the Cloud Report from Flexera

Figure 2. Cloud computing trends 2019

The service industry is currently struggling with various operational issues such as designing, maintenance and management, whereby with AI, industry stands to several benefits. Cloud computing service industry need intelligent decisions to manage their complex and dynamic operations. Through deep learning, machines can use existing training data in industry to analyze large amounts of data retrieved through data mining. AI gradually masters this data, as traffic increases, the accuracy of decision made also enhances. Using AI, system security has also increased whereby attacks can be detected automatically through machine learning. This has resulted in less attacks adding value to both the users and firms

Resource management in open stack

OpenStack is the fastest growing free open source software, announced in July 2010. OpenStack is a collection of open source software project that cloud computing technologist can use to setup and run their cloud computer and storage infrastructure. OpenStack mainly consist of three core software project which are OpenStack Compute Infrastructure (Nova), OpenStack Object Storage Infrastructure (Swift) and OpenStack Image Service Infrastructure (Glance). OpenStack open source software supports for creating private and public clouds, built and disseminated by a large and democratic community of developers, in collaboration with users. OpenStack is mostly deployed as Infrastructure-as-a-Service (IaaS), whereby virtual servers and other resources are made available to customers. The software platform consists of interrelated components that control diverse, multi-vendor hardware pools of processing, storage and networking resources throughout a data center.

DOI: 10.21522/TIJAR.2014.SE.19.02.Art010
ISSN: 2520-3088

The cloud is mainly providing computing features for end users in a remote environment, where the actual software runs as a service on reliable, scalable servers, rather than on each end users computer. OpenStack give facilities for deploying virtual machines (VMs) and other instances which handle different tasks for managing a cloud environment. It provides horizontal scaling with ease, which means that tasks which benefit from running concurrently can easily serve more as well as less users on the fly, by just spinning up more instances.

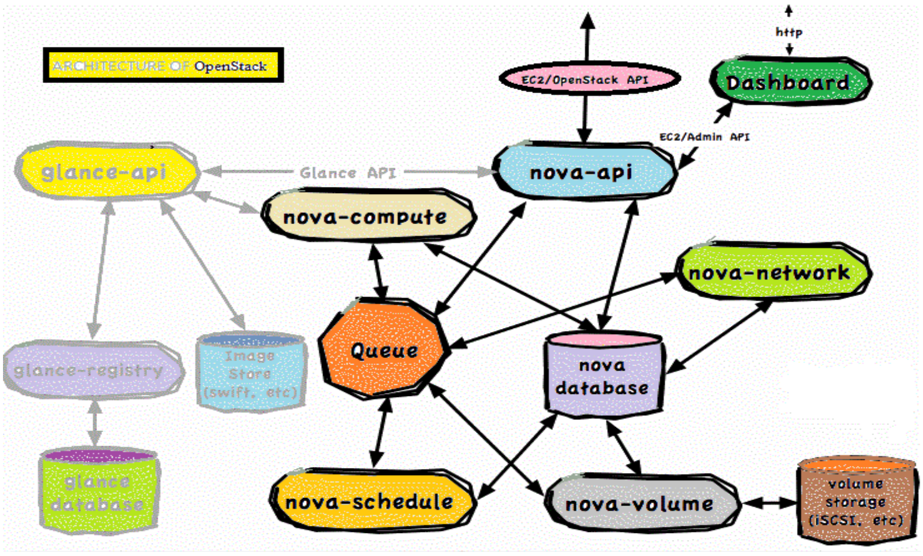


Figure 3. Open stack architecture

Components of open stack

1. Compute (Nova): OpenStack Compute (Nova) is a cloud computing fabric controller, which is used for deploying and managing large numbers of Virtual Machines and other instances to handle computing tasks.
2. Object Storage (Swift): OpenStack Object Storage (Swift) is a scalable redundant storage system for objects and files. Objects as well as files are written to multiple disk drives spread throughout servers in the data center, OpenStack software is only responsible for ensuring data replication and integrity across the cluster.
3. Block Storage (Cinder): OpenStack Block Storage (Cinder) is a block storage component, which is more analogous to the traditional notion of a computer being able to access specific locations on a disk drive as well as, it provides persistent block-level storage devices for use with OpenStack compute instances. In OpenStack, the block storage system manages the creation, attaching, detaching of the block devices to servers.
4. Networking (Neutron): OpenStack Networking (Neutron) provides the networking capability for OpenStack and it is a system for managing networks and IP addresses easily, quickly and efficiently.
5. Dashboard (Horizon): OpenStack Dashboard (Horizon) is the dashboard behind OpenStack which provides administrators and users a graphical interface to access, provision and automate cloud-based resources.
6. Identity Service (Keystone): OpenStack Identity (Keystone) provides identity services for OpenStack, for it is a central directory of users mapped to the OpenStack services they can access. It provides multiple means of access, and acts as a common authentication system across the cloud operating system and can integrate with existing backend directory services like LDAP.
7. Image Service (Glance): OpenStack Image Service (Glance) provides image services to OpenStack, discovery, registration and delivery services for disk and server images, it also allows these images to be used as templates when deploying new virtual machine instances.
8. Telemetry (Ceilometer): OpenStack Telemetry Service (Ceilometer) provides telemetry services, which allow the cloud to provide billing services to individual users of the cloud, it keeps a

verifiable count of each user’s system usage of each of the various components of an OpenStack cloud.

9. Orchestration (Heat): OpenStack Orchestration (Heat) is a service which allows developers to store the requirements of a cloud application in a file that defines what resources are necessary for that application.
10. Database (Trove): OpenStack (Trove) is a database as a service which provides relational and non-relational database engines.

Issues and challenges

Resource scheduling is a hotspot area of research in cloud due to large execution time and resource cost. Different resource scheduling criteria and parameters are directed to different categories of Resource Scheduling Algorithms (RSAs). There are several problems to be considered while managing resources, such as, type of resource required physical as well as logical allocation, brokering, provisioning, mapping, adaptation. The major issues which is commonly associate with IaaS in cloud systems are virtualization and multitenancy, resource management, infrastructure management, data management, APIs, interoperability etc.

Virtualization and multitenancy

Virtualization is an essential technological characteristic of clouds, which hides the technological complexity from the user and enables enhanced flexibility (through aggregation, routing and translation). In a multitenancy environment, multiple customers share the same application, running on the same operating system, on the same hardware, with the same data –storage mechanism. The distinction between the customers is achieved during application design, thus customers do not share or see each other's data. In case of virtualization, components are abstracted enabling each customer application to appear to run on a separate physical machine

Related to the problem of resource allocation in VMs.

It provides a concise overview of various existing techniques for resource allocation in VMs. These can be helpful for developing detailed designs, specifications and performance evaluation techniques for VMs. However, this paper does not attempt to provide an exhaustive survey on the resource allocation/management techniques in VMs.

Multi-tenancy

It is a highly essential issue in cloud systems, where the location of code and/or data is principally unknown and the same resource may be assigned to multiple users. This affects infrastructure resources as well as data/applications/services that are hosted on shared resources but need to be made available in multiple isolated instances. Multi-tenancy implies a lot of potential issues, ranging from data protection to legislator issues. While hardware-based virtualization has many benefits, it lacks from a high level of scalability required to offer cost effective cloud computing for masses. Multi-tenant virtualization remedies this bottleneck by focusing on software-based virtualization.

Resource management

At any instance, resources are to be allocated to effectively handle workload fluctuations, while providing QoS guarantees, to the end users. The computing and network resources are limited and have to be efficiently shared among the users in virtual manner. In order to perform effective resource management, we need to consider the issues such as resource mapping, resource provisioning, resource allocation and resource adaptation. The lack of mature virtualization tools and powerful processors, have prevented growth of cloud computing. The main challenge is to determine the resource demand, of each application at its current request load level, and to allocate resources in most efficient way. Metering of any kind of resource and service consumption is essential in order to offer elastic pricing, charging and billing. It is therefore a precondition for the elasticity of clouds. The issue here is to see that the users are charged only for the services that they use for the specific period of time. Cloud computing alone will not help an organization to determine who will pay for what resource, but it can help provide a platform for an infrastructure ensign that establishes a charge-back model for metering and billing.

Network infrastructure management

Managing millions of network components (hubs, bridges, switches etc.) leads to unsustainable administrator costs, requiring automated methods for typical system management tasks. These automated methods need to deal with increased monitoring data size of several orders of magnitudes higher than current systems. Chiaraviglio and Matta (2010) have proposed cooperation between ISPs (Internet Service Providers) and content providers that allow the achievement of an efficient simultaneous allocation of compute resources and network paths that minimize energy consumption.

Security, privacy and compliance

They are obviously essential in all systems dealing with potentially sensitive data and code. To ensure adequate security in cloud computing, various security issues, such as authentication, data confidentiality, integrity, and non-repudiation need to be considered.

Data management

It is an essential aspect in particular, for storage clouds, where data is flexibly distributed across multiple resources. Implicitly, data consistency needs to be maintained over a wide distribution of replicated data sources. At the same time, the system always needs to be aware of the data location (when replicating across data centers) taking latencies and work load into consideration.

APIs and/or programming enhancements are essential to exploit the cloud features. Common programming models require that the developer takes care of the scalability and autonomic capabilities, whilst a cloud environment, provides the features in a fashion that allows the user, to leave such management to the system.

The important issues identified in resource management are resource provisioning, resource allocation, resource adaptation, resource mapping, resource modelling and selection, resource brokering, and resource scheduling. Their definitions are highlighted in Table 1. AI approach can play a vital role to resolve this issue.

Table 1

S. No.	Issue	Definition
1	Resource provisioning	It is the allocation of a service provider's resources to a customer
2	Resource allocation	It is the distribution of resources economically among competing groups of people or programs
3	Resource adaptation	Resource adaptation It is the ability or capacity of that system to adjust the resources dynamically to fulfill the requirements of the user
4	Resource mapping	It is a correspondence between resources required by the users and resources available with the provider
5	Resource modeling	Resource modeling is based on detailed information of transmission network elements, resources and entities participating in the network. It is a framework that illustrates the most important attributes of resource management: states, transitions, inputs and outputs within a given environment. Resource modeling helps to predict the resource requirements in subsequent time intervals
6	Resource scheduling	A resource schedule is a timetable of events and resources. Shared resources are available at certain times and events are planned during these times. In other words, It is determining when an activity should start or end, depending on its duration, predecessor activities, predecessor relationships, and resources allocated

Conclusion

The movement toward cloud computing has made a massive progress with most service provider companies aiming to transform legacy network to modernized network dependent on network function virtualization with software-defined networking to compete and sustain in competitive pressure of a fast-changing environment. Tech Giants are heavily investing in various AI technologies. These investments are quite strategic and went beyond simple R&D extensions of existing products. However, just moving to cloud is not enough, as intelligent decisions to manage complex and dynamic operations become necessary and impossible for humans. The monitoring tools currently available are inadequate to allocate more resources to provide administrative service without interruption or without shutting down. With AI, importance comes to add value to cloud, leading to better traffic classification, more accurate network fault predictions, time optimization and heightened customer services. The findings are based heavily on literature reading, analyzing results, relative comparisons of the results in different research papers, that conclude that the AI and cloud computing model fits well to analyze consumption pattern of resources by analyzing large amount of data to classify traffic, have more accurate predictions, and detect anomalies, leading to better decision making on resource allocation or upgradation operations at any given time.

References

[1]. Sunil Kumar S, Manvi A, Gopal Krishna Shyamb., 2014, Resource management for Infrastructure as a Service (IaaS) in Cloud Computing: A survey. Journal of Network and Computer Applications 41(2014) 424–440.

[2]. Rakesh Kumar, Neha Gupta, Shilpi Charu, Kanishk Jain, Sunil Kumar Jangir, 2014, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.5, pg. 89-98.

[3]. Pieter-Jan Maenhaut, Hendrik Moens, Bruno Volckaert, Veerle Ongenae and Filip De Turck, 2017, Resource Allocation in the Cloud: From Simulation to Experimental Validation, proceedings of IEEE 10th International Conference on Cloud Computing.

[4]. Dr. Balamurugan E, Sathishkumar K, Dr. Sangeetha, 2018, A Survey on Software as a Service (SaaS) Cloud for High Level Language Computing. International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing (EECCMC). January 28 & 29, 2018.

[5]. S. Russell and P. Norvig, 2015. Artificial Intelligence: A Modern Approach, Prentice Hall, New York.

[6]. B. Jennings and R. Stadler, Resource management in clouds: Survey and research challenges, 2015 Journal of Network and Systems Management, vol. 23, no. 3, pp. 567 – 619, 2015.

[7]. Richars Layne, 2019. Artificial Intelligence and Cloud Computing, The Future of Scientific Research, https://www.tessella.com.

[8]. https://docs.openstack.org/ceilometer/latest/install/get_started.html.

[9]. https://thenewstack.io/how-openstack-provides-scalable-reusable-infrastructure-for-ai-ml-workloads/.
https://wiki.openstack.org/wiki/Gyan/TFiO.