fitness of the components defined as utility provided by that components versus cost of retaining that component in the cloud. Based on this fitness, it is decided on whether that component needs to stay in the cloud or be moved from one VM to another or be destroyed. In order to accomplish this mechanism, an approach was proposed which has agents that run on physical server. These agents are responsible for handling lifecycle of the components and also perform health checks.

Mathais Bjorkqvist et. al. [11, 12] proposed opportunistic way of managing resources for Web applications. The approach centered on maintaining target system utilization through predictive workload and VM performance. Web applications are based on Service Oriented Architecture where applications are composed of atomic services which are typically Web services and service composition execution engines (middleware for executing composition of atomic services) that reside on Virtual Machines (VMs). The objective of this research was to ensure use of lesser number of faster virtual machines (VMs) to be used at any given time rather than more number of slower VMs for servicing requests from the consumers. Though through put for both options mentioned earlier should be the same however the cost of maintaining former will certainly be lesser than the latter option. The architecture which was considered, took into account of a Virtual Machine Broker which decides on number of VMs where the servers will be running. VM Controller controls the VMs where the services and its replicas are running. Load balancers typically distribute incoming service requests among service replicas. The replication policy is targeted to be implemented on the VM broker. The policy enable VM broker to first decide number of VMs for a service and then to select VMs based on performance of the active VMs and billing periods.

Rodrigo N Calheiros et. al. [13] focused on the problems encountered in the area of virtualization, workload and performance modeling, deployment and monitoring of applications on cloud. These problems are driven by unpredictable behavior of virtualized IT resources and network elements and eventually lead to overprovisioning of underlying resources and negative impact on Quality of Services (QoS). The solution proposed was modeled on analytical performance model and workload information. Both these areas feed in data to Application provisioner that handles application and Virtual Machine provisioning. Analytical performance model was used to predict effectiveness of provisioning schedule on desired Quality of Service (QoS). Workload information provides application provisioner demands for resourcing needs, removing uncertainties and over/under estimation of Cloud resources. The algorithm proposed by them to calculate optimal number of virtual machines is used by the Load predictor and Performance modeler.

Andreas Lodde et al. [15] stressed on service response time as a parameter for measuring and controlling service provisioning. A history of requests and associated response time is maintained where services are hosted. The underlying framework devised, has a key component - request classifier which takes in service requests, collects required information and appends requests with header information. The request scheduler defines the sequence of request processing and a dispatcher acts on the guidelines to send requests to the computing resource. In order to calculate optimal service provisioning resources, dimension evaluator sends number of VM instances needed which a processing simulator verifies using history of processed requests, and SLA evaluator checks on whether SLA is complied with. This process is iterative till SLA evaluator confirms on SLA compliance, once confirmed, dimension evaluator ends its process and this new number is applied to the cloud resources for servicing requests. The key aspect that drives above framework is the history of requests that is stored which forms basis of calculating optimal resourcing. Be it processed requests or requests waiting to be processed, history data is maintained for processing simulator to verify numbers proposed by dimension evaluator. The dimension evaluator uses binary search algorithm to determine minimum possible number of instances needed to service requests.

Kuo-Chan Huang et. al [16] highlight issues that arise from software based distributed workflows and online concurrent user access and addresses important issues pertaining to resource allocation and dynamic provisioning of services in cloud. A service deployment strategy was devised based on application service flows and a resource provisioning technique based on future service requests. The strategy detects queued application service flow instances and calculates number of remaining service type to be executed. Then the resources are adjusted based on the each service type. Two scenarios were taken into consideration, one where number of virtual machines was greater that the number of service types deployed. Therefore each service type can be assigned one or more VMs. In the other computing scenario, multiple service types are assigned to one VM. The provisioning mechanism estimates based on three different policies which are based on the current, short term and long term system workload and nature of remaining service flows.

At times web applications are hosted on cloud. Even though service providers offer SLAs to the consumers, however that is not enough to guarantee response time for web applications. It also becomes complex when there is multitier applications hosted which have Web server tier for consumer requests, application tier for business specific calculations and logic and data tier for storing persistent data. Currently neither the commercial cloud providers nor the existing open source providers support maximum response time guarantees. Waheed Iqbal et. al. [17] researched on this area to provide guarantees on maximizing response time for web applications with minimum resource utilization. The algorithms designed as a part of this research can detect bottleneck in multi-tier Web applications hosted on a cloud. This was done through profiling of the CPU, memory, and I/O resource usage of each tier with a combination of real time monitoring and processing of each tier's log files. Usually web applications need configuration of concurrency levels for Database tier connections, threads to Application tier and worker processers in Web tier. Any inaccuracy on these levels create bottleneck. The algorithm focuses on heuristic analysis done for multi-tier web application and on detection of a bottleneck the solution is either administered through

Licensed Under Creative Commons Attribution CC BY

horizontal scaling the tier with load balancing or by vertical scaling by dynamically increasing underlying resources.

Mahyar Movahed Nejad et. al. [18] focus on problems faced by cloud service providers on virtual machine provisioning, especially in the area of auction based models. The cloud service providers provision resources either on static or dynamic provisioning basis. In static provisioning, the service provider provisions set of virtual machines even before user requests come in. In dynamic provisioning, the service provider considers user demands before provisioning virtual machines. To sell its services, the service providers either use fixed pricing approach or auction based approach. In fixed pricing, the price for each type of VM is fixed and pre-determined by the service provider which hardly changes over time. However, auction based pricing enables service providers to bundle available VMs and use auction to fix price and allocation to the users. This creates a win-win model from service provider and consumer. The service consumer can get resources at lower price than fixed price model and service providers get increased revenue for unutilized VMs. Auction based model was considered in this research and formulated an integer program that considers different types of resources while performing dynamic VM provisioning. This model setup is slightly different from service providers such as Amazon where auction is carried out for short time and for individual VM instances (rather than bundles). Also that wining users pay same price (per unit) for VMs instances. The model that was proposed will bundle available VMs instances and open it for auction to set of users. Each user shall assign private price for bundle and pays if the bid is successful. The underlying uniqueness of this model is that it considers different resources such as processer cores, memory, storage, etc. which is a reality in today's world.

Another interesting area of solving service provisioning problems is application of game theory. G. Wei et. al. [19] proposed game theory to solve resource allocation in cloud services. The proposition had two steps, first each participant tries to solve the problem independently without considering resources getting allocated by other participants. This was achieved through Binary Integer Programming method for individual optimization of resource allocation. Second step is to combine multiplexed strategies of those individual optimal solutions with minimum loss of efficiency. It is also demonstrated that Nash equilibrium is attained during this optimal resource allocation game. In fact generalized Nash equilibria were further extended by Danilo Ardagna et. al. [20] in solving problems faced by SaaS providers who host applications at an IaaS provider. Each SaaS provider has to conform to QoS standards as per agreed SLA which determines net revenues for the services provided, and also at same time it has to ensure minimum cost of using resources provided by IaaS providers. So in order to balance between services rendered by SaaS and service consumed from IaaS provider, this research attempted to devise efficient distributed algorithm based on General Nash Equilibrium for runtime allocation of IaaS resources among competing SaaS providers. The cost model of the algorithm consists of utility functions, which has both revenues received and penalties incurred during the achieved performance of the resource allocation and also includes infrastructural costs associated with IaaS resources. In similar lines, Valerio Di Valerio et. at. [21] also focused on SaaS provider problems and used Stackelberg game approach to devise solution which derives equilibrium price and allocation strategy through Mathematical Program with Equilibrium Constraints(MPEC) problem. The scenario considered here was an IaaS provider offers resources to several SaaS providers, which is on flat pricing basis, is on demand, and for spot VM instances. In turn, SaaS providers offer to end users Web applications with Quality of Service (QoS) guarantees, using the IaaS facilities to host and run the provided applications. For SaaS provider, revenues and penalties depend on provisioning of resource at an adequate performance level, which is specified in a Service Level Agreement (SLA) contract that each SaaS provider conspires with its end users. Therefore, each SaaS provider fasces the problem of determining the optimal number of VMs to satisfy the SLA with his end users while maximizing revenue. A two stage provisioning scheme was proposed. In the first stage, the SaaS providers determine the number of required flat and on demand instances by means of standard optimization techniques. In the second stage the SaaS providers compete, by bidding for the spot instances which are instantiated using the unused IaaS capacity. It was assumed that the SaaS providers want to maximize a suitable utility function which accounts for both the QoS delivered to users and the associated cost. This stage is modelled on Stackelberg game theory which provides way to derive equilibrium solution for maintaining QoS of service delivered and keeping the allocations to optimal.

6. Importance of Service Level Agreement in Cloud

From service design to service rendering, Service Level Agreement (SLA) plays a pivotal role. It is pertinent to have ground rules set in before a service gets consumed. These rules are used by service providers to define nature of service based on the available capacity and vulnerability for hardware layers. Various aspects that need to be considered while designing are network, security, storage, processing power, and database. SLA is legal agreement or a contract which bound service providers on clause pertaining to Quality of Service [15]. It forms basis of how a service is rendered by service provider and consumed by service consumer. The agreement has following important characteristics:

- describes clearly a service so that consumer can understand functionalities of the service
- articulates the performance levels of the service
- defines mode by which the service parameters can be monitored and reported
- imposes penalties in case the service does not meet service requirements

In order to set specific SLAs, service providers need to capture service requirements. The requirements take the form of parameterized set of data which is provided by service consumers. This parameterized set consists of start/stop time of the service, duration of uptime, quantity, throughput/speed, etc. The service provider has to consider

Licensed Under Creative Commons Attribution CC BY

each of these data and render service such that these data is complied with. These SLAs can be categorized based on the following criteria -

- Speed: performance criterion that captures rate at which a service can be consumed
- Availability: service characteristic that encapsulates likelihood of responding to service consumers whenever accessed
- Accuracy: completeness and worthiness of the results from the service

Figure 4 depicts various parameters that are communicated between a service provider and consumer. While service is getting delivered over a period of time, the Quality of Service (QoS) can be measured through verification of SLA based on the criteria mentioned in the following-

- Reliability: ability to perform required function under stated function
- Flexibility: options provided by service provider on this service offered
- Capability: ability to meet demand of a given size under internal conditions
- Usability: effectiveness and efficiency with which a service consumer can accomplish specified tasks

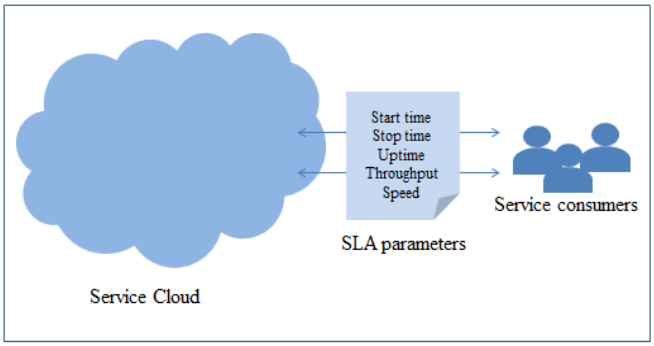

Figure 4: Service Cloud and associated SLAs

Once the service has been designed and deployed based requirements of service consumers, service providers measure SLAs and track in the form of metrics which is shared with service consumers from time to time. However an SLA cannot convert a good service from a bad service, but can mitigate risk from choosing a bad service. There are several tools and framework to manage various aspects of SLA lifecycle such as for SLA specification and modelling, tools such as Cloudscale, Viola can be leveraged, for SLA enforcement, tools such as CloudSOA, Stream can be chosen and for SLA management, Plane, SmartLM can be used.

7. Research Challenges

The research papers reviewed focus primarily on virtual machine provisioning as almost all service providers do not let underlying cloud infrastructure visible to outside world. The algorithms proposed by the researchers on economic fitness of the existing VMs and deciding on whether to abort slower running VMs or to start new instances. From the already completed work, it could be found that service availability is one of primary challenges in current research on this topic. Service consumption is occurring between service providers and consumer over a network. Hence service disruption, network congestion, poor signal or even a node failure are highly intolerable in service provisioning. It has been also noted that with the evolution of cloud infrastructures, supports for complex mathematical models for virtual machine allocation and maintain high Quality of Service (QoS), provides an opportunity to implement sophisticated scheduling algorithms. In addition to enhancements to existing scheduling algorithms, new methodologies could be applied, such as the adaptive virtual machine scheduling and SLA driven models for virtual machine allocation.

Since the complexity of Cloud computing is phenomenal given the size of consumer requests and underlying infrastructure that supports such requests, it will interesting to see how existing algorithms behave in unstable cloud environment where underlying resources are changed. Also another key aspect of service provisioning is Service Levels Agreements (SLAs) which need to take into account while refining service provisioning algorithms. SLA measurement and monitoring has to be considered while purview of service provisioning algorithms to cover 360 degree view of the service provided and consumed.

8. Conclusion

Cloud computing is an emerging paradigm that is revolutionizing the way computing is performed and administered. With virtualization of physical resources, infrastructure, and applications, cloud service provisioning has become a reality. The growing adoption of cloud services by all industries indicates underlying value proposition which cloud computing carries. Due to rise in expectations in service provisioning, providing effective and continuous service in cloud is getting difficult. Therefore it is pertinent to understand service provisioning fundamentals, taxonomy, and several expectations that must be considered to evaluate the provisioned services in terms of user requirements and costs. The Service Provisioning algorithms and techniques reviewed in this paper provide insights on optimizing underlying resources for efficient provisioning of services. Besides, Service Level Agreement which forms basis of service provisioning formulations and consumption has been articulated. Finally, open research challenges are categorized and identified for future research directions.

References

[1] Md Whaiduzzaman, Mohammad Nazmul Haque, Md Rejaul Karim Chowdhury, and Abdullah Gani, “A Study on Strategic Provisioning of Cloud Computing Services”, eScientific World Journal, Volume 2014, Article ID 894362, 16 pages

[2] Ian Foster, Carl Kesselman, Jeffrey M. Nick and Steven Tuecke, “Grid Computing: Making of Global Infrastructure a Reality”, Wiley 2003, pp. 217-249

[3] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia, “Above the Clouds: A Berkeley

Licensed Under Creative Commons Attribution CC BY

View of Cloud Computing", Advanced Computing Machines, pp. 50-58
[4] http://www.nist.gov/itl/cloud/index.cfm
[5] Borko Furht, Armando Escalante , "Handbook of Cloud Computing", Springer, 2010, pp 3-8
[6] Simson L. Garfinkel, "An Evaluation of Amazon's Grid Computing Services: EC2, S3 and SQS", Center for Research on Computation and Society, Harvard University, Technical Report, 2007
[7] Malawski, M. Kuźniar, M. Wójcik, P. Bubak, M. , "How to Use Google App Engine for Free Computing", Internet Computing, IEEE , Volume:17 , Issue: 1 , pp. 50 - 59
[8] Jerry Gao, Xiaoying Bai, W.T. Tsai, Tadahiro Uehara, "SaaS Testing on Clouds-Issues, Challenges and Needs", IEEE 7th International Symposium on Service-Oriented System Engineering, pp.409-415, 2013
[9] Nicolas Bonvin, Thanasis G. Papaioannou and Karl Aberer, "Autonomic SLA-driven Provisioning for Cloud Applications", 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pp.434-443, 2011
[10] Nicolas Bonvin, Thanasis G. Papaioannou and Karl Aberer, "An economic approach for scalable and highly-available distributed applications" IEEE 3rd International Conference on Cloud Computing, Pg. 498-505, 2010
[11] Mathias Bjorkqvist, Lydia Y. Chen, Walter Binder, "Opportunistic Service Provisioning in the Cloud", IEEE 5th International Conference on Cloud Computing, pages 237-244, 2012
[12] Mathias Bj¨orkqvist, Lydia Y. Chen, Walter Binder, "Dynamic Replication in Service-Oriented Systems", 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Pages 531-538, 2012
[13] Rodrigo N. Calheiros, Rajiv Ranjan, and Rajkumar Buyya, "Virtual Machine Provisioning Based on Analytical Performance and QoS in Cloud Computing Environments", International Conference on Parallel Processing, pp.295-304, 2011
[14] Andreas Lodde, Antoine Schlechter, Pascal Bauler, Fernand Feltz, "SLA-Driven Resource Provisioning in the Cloud", 1st International Symposium on Network Cloud Computing and Applications, pp. 28-35, 2011
[15] Mohammed Alhamad, Tharam Dillon, Elizabeth Chang, "SLA-Based Trust Model for Cloud Computing", 13th International Conference on Network-Based Information Systems, pp. 321-324, 2010
[16] Kuo-Chan Huang, Bo-Jyun Shen, Tsung-Ju Lee, Hsi-Ya Chang, Yuan-Hsin Tung, Pin-Zei Shih, "Resource Allocation and Dynamic Provisioning for Service-Oriented Applications in Cloud Environment", IEEE 4th International Conference on Cloud Computing Technology and Science, Pg. 839-84, 2012
[17] Waheed Iqbal, Matthew N. Dailey, David Carrera, "SLA-Driven Dynamic Resource Management for Multi-tier Web Applications in a Cloud", 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, Pg. 37-46, 2010
[18] Mahyar Movahed Nejad, Lena Mashayekhy, Daniel Grosu "Truthful Greedy Mechanisms for Dynamic Virtual Machine Provisioning and Allocation in Clouds", IEEE Transactions on Parallel and Distributed Systems, Pg. 188-195, 2013
[19] G. Wei, A. V. Vasilakos, Y. Zheng, and N. Xiong, "A game theoretic method of fair resource allocation for cloud computing services," The Journal of Supercomputing, pp. 1–18, 2009
[20] Danilo Ardagna, Barbara Panicucci, and Mauro Passacantando, "Generalized Nash Equilibria for the Service Provisioning Problem in Cloud Systems", IEEE Transactions on Services Computing, Vol. 6, No. 4, Oct-Dec 2013
[21] Valerio Di Valerio, Valeria Cardellini , Francesco Lo Presti, "Optimal Pricing and Service Provisioning Strategies in Cloud Systems: A Stackelberg Game Approach", IEEE 6th International Conference on Cloud Computing, Pg. 115-122, 2013

Author Profile

Mridul Paul is currently pursuing PhD in Computer Science and Engineering from Birla Institute of Technology, Mesra, a deemed University. He received B.E. in Computer Science & Engineering from same University and also possesses MBA degree in Finance. He has been associated with Information Technology field for last 15 years and his area of interest is in Cloud Computing. He is currently with Cognizant Technology Solutions as Associate Director-Projects.

Dr. Ajanta Das is working as Associate Professor in the department of Computer Science & Engineering in Birla Institute of Technology, Mesra, a deemed University. She is having altogether eighteen years of experience including six years of industry experience. She had worked for reputed, famous, pioneer companies like, Tata Steel, Jamshedpur, India and Lexis Nexis Inc., Boston, USA. She has proven herself confident and capable of handling real-life projects according to the deadline. Later, she joined academic and started teaching and research simultaneously. She has been awarded PhD in Engineering from Jadavpur University in 2009. Her major interest of teaching includes database, software engineering and distributed computing. Her focused research area includes Grid Computing, Cloud Computing and Wireless Sensor Network. She is life member of Computer Society of India.

Licensed Under Creative Commons Attribution CC BY