**Cloud Resource Management and Scheduling**

**Cloud Resource Management:**

Cloud computing is a new era of remote computing / Internet based computing where one can access their personal resources easily from any computer through Internet. It contains large number of shared resources. So Resource Management is always a major issue in cloud computing like any other computing paradigm.

Resource management is a core function of any man-made system. It affects the three basic criteria for the evaluation of a system: performance, functionality, and cost. An inefficient resource management has a direct negative effect on performance and cost and an indirect effect on the functionality of a system. Indeed, some functions provided by the system may become too expensive or may be avoided due to poor performance.

A cloud is a complex system with a very large number of shared resources subject to unpredictable requests and affected by external events it cannot control. Cloud resource management requires complex policies and decisions for multi-objective optimization. Cloud resource management is extremely challenging because of the complexity of the system, which makes it impossible to have accurate global state information, and because of the unpredictable interactions with the environment. Also, with a wide variety of private, hybrid, and public cloud-based systems and infrastructure already in use, companies surely need to consider resource management in their cloud computing strategy. However, resource management for such a complex system as cloud computing requires different ways of measuring and allocating resources.

It has been argued for some time that in a cloud, where changes are frequent and unpredictable, centralized control is unlikely to provide continuous service and performance guarantees. Indeed, centralized control cannot provide adequate solutions to the host of cloud management policies that have to be enforced. Autonomic policies are of great interest due to the scale of the system, the large number of service requests, the large user population, and the unpredictability of the load. The ratio of the mean to the peak resource needs can be very large.

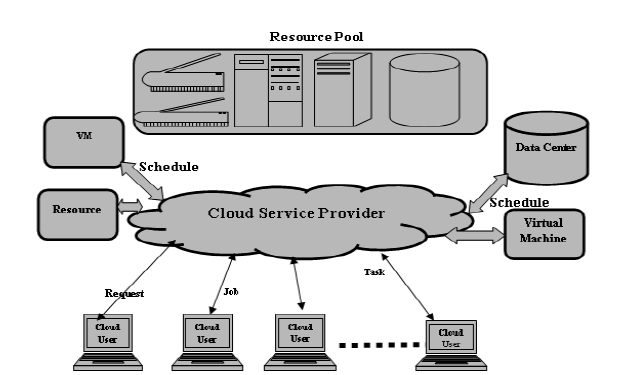


Figure: Basic Resource Management and Scheduling in Cloud Computing

**The resource management strategy:**

Resource management is a core function required for any cloud system, and inefficient resource management has a direct negative effect on performance and cost, while it can also indirectly affect system functionality, becoming too expensive or ineffective due to poor performance.

The strategies for cloud resource management associated with the three cloud delivery models – Iaas, PaaS, SaaS – differ from one another. In some cases, when cloud service providers can predict a spike, they can provision resources in advance (ex. the case of seasonal web services).

Technically speaking, a cloud is a portion of cluster resources capable of growing and shrinking to accommodate the load changes. Also, cloud resources are controlled on three independent levels:

**Cluster level** – The cluster level of power management is represented by cluster resource manager, a software complex that manages resources and tasks in a cluster in order to maintain its efficiency. Basically, a CRM is responsible for creation and deletion of clouds.

**Node level** – The node-level power management is done by an operating system (OS), so an OS controls the high-level state of equipment. For instance, to save energy the OS can put a processor (CPU) into the sleep state or spin-down disks.

**Hardware level** – Modern CPUs consist of many modules, which may not be permanently involved in an operation. Therefore, unused modules can be switched off. This is done by a special circuit responsible for internal power management of the CPU. So, all management is done on a hardware level without involving any OS.

**Controlling the cloud**

Allocation techniques in computer clouds must be based on a disciplined approach, rather than ad hoc methods. So, here are the four basic mechanisms for implementing resource management policies in cloud computing:

**Control theory** – Control theory uses feedback to guarantee system stability and predict transient behavior, but it can only predict local behavior.

**Machine learning** – A major advantage of machine-learning techniques is that they don’t need a performance model of the system. You could apply this technique to coordinating several autonomic system managers.

**Utility-based** – Utility-based approaches require a performance model and a mechanism to correlate user-level performance with cost.

**Market-oriented** – Such mechanisms don’t require a system model, such as combining auctions for bundles of resources.

**Policies and mechanisms for resource management:**

A policy typically refers to the principal guiding decisions, whereas mechanisms represent the means to implement policies. Separation of policies from mechanisms is a guiding principle in computer science.

Cloud resource management policies can be loosely grouped into five classes:

1. Admission control.

2. Capacity allocation.

3. Load balancing.

4. Energy optimization.

5. Quality-of-service (QoS) guarantees.

The explicit goal of an admission control policy is to prevent the system from accepting workloads in violation of high-level system policies; for example, a system may not accept an additional workload that would prevent it from completing work already in progress or contracted. Limiting the workload requires some knowledge of the global state of the system. In a dynamic system such knowledge, when available, is at best obsolete. Capacity allocation means to allocate resources for individual instances; an instance is an activation of a service. Locating resources subject to multiple global optimization constraints requires a search of a very large search space when the state of individual systems changes rapidly.

Load balancing and energy optimization can be done locally, but global load-balancing and energy optimization policies encounter the same difficulties as the one we have already discussed. Load balancing and energy optimization are correlated and affect the cost of providing the services. Indeed, it was predicted that by 2012 up to image of the budget for IT enterprise infrastructure would be spent on energy.

The common meaning of the term *load balancing* is that of evenly distributing the load to a set of servers. For example, consider the case of four identical servers, image, and image, whose relative loads are image, and image, respectively, of their capacity. As a result of perfect load balancing, all servers would end with the same load image of each server’s capacity. In cloud computing a critical goal is minimizing the cost of providing the service and, in particular, minimizing the energy consumption. This leads to a different meaning of the term *load balancing*; instead of having the load evenly distributed among all servers, we want to concentrate it and use the smallest number of servers while switching the others to standby mode, a state in which a server uses less energy. In our example, the load from image will migrate to image and the load from image will migrate to image; thus, image and image will be loaded at full capacity, whereas image and image will be switched to standby mode. Quality of service is that aspect of resource management that is probably the most difficult to address and, at the same time, possibly the most critical to the future of cloud computing.

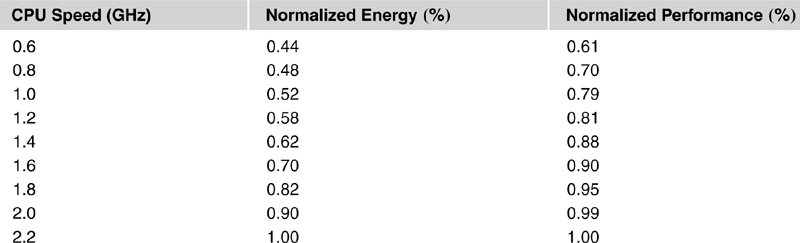


Table – 1 The normalized performance and energy consumption function of the processor speed. The performance decreases at a lower rate than does the energy when the clock rate decreases

As a result of lower voltages and frequencies, the performance of processors decreases, but at a substantially slower rate than the energy consumption. [Table 1](https://learning.oreilly.com/library/view/cloud-computing/9780124046276/xhtml/CHP006.html#T0005) shows the dependence of the normalized performance and the normalized energy consumption of a typical modern processor on clock rate. As we can see, at image GHz we save image of the energy required for maximum performance, whereas the performance is only image lower than the peak performance, achieved at image GHz. This seems a reasonable energy-performance tradeoff!

Virtually all optimal – or near-optimal – mechanisms to address the five classes of policies do not scale up and typically target a single aspect of resource management, e.g., admission control, but ignore energy conservation. Many require complex computations that cannot be done effectively in the time available to respond. The performance models are very complex, analytical solutions are intractable, and the monitoring systems used to gather state information for these models can be too intrusive and unable to provide accurate data. Many techniques are concentrated on system performance in terms of throughput and time in system, but they rarely include energy tradeoffs or QoS guarantees. Some techniques are based on unrealistic assumptions; for example, capacity allocation is viewed as an optimization problem, but under the assumption that servers are protected from overload.

**Resource Scheduling in Cloud Computing:**

Resource scheduling refers to the different algorithms that service providers use to deliver and allocate the different resources in a virtual environment. The premise is that resources are very limited so tenants and users do not actually own or reserve the resources that have been allocated to them, but rather they are allocated what resources they presently require based on the scheduling algorithms and that the advertised amounts of resources they have are only ceiling values. This allows the system to be lean, having no resources pinned down, wasted and unused.

Resource scheduling is a key step of [project management](https://www.visual-planning.com/en/solutions/project-management-software). When resource availability and work capacity are the primary factors that determine a project’s deadline, project managers sometimes speak of resource**-**constrainedscheduling. But resource scheduling is also often used for simple operation management, as it allows managers to outline completion dates for tasks assigned to their teams, which they can report to stakeholders such as customers or a board of directors.

**Resource Scheduling in Cloud Computing: Issues and Challenges**

Dynamic allocation of tasks to computers is complicated in the cloud computing environment due to the complicated process of assigning multiple copies of the same task to different computers. Likewise, Resource scheduling in cloud is a challenging job and the scheduling of appropriate resources to cloud workloads depends on the QoS requirements of cloud applications. In cloud environment, heterogeneity, uncertainty and dispersion of resources encounters problems of allocation of resources, which cannot be addressed with existing resource allocation policies. Researchers still face troubles to select the efficient and appropriate resource scheduling algorithm for a specific workload from the existing literature of resource scheduling algorithms.

There have been various types of scheduling algorithm that exists in distributed computing system. Most of them can be applied in the cloud environment with suitable verifications. The main advantage of job scheduling algorithm is to achieve a high performance computing and excellent system throughput. Traditional job scheduling algorithms are not able to provide scheduling in the cloud environments. According to a simple classification, job scheduling algorithms in cloud computing can be categorized into two main groups are Batch mode heuristic scheduling algorithms (BMHA) and online mode heuristic algorithms. In BMHA, Jobs are queued and collected into a set when they arrive in the system. The scheduling algorithm will start after a fixed period of time.

The main examples of BMHA based algorithms are; First Come First Served scheduling algorithm (FCFS), Round Robin scheduling algorithm (RR), Min–Min algorithm and Max–Min algorithm. By On-line mode heuristic scheduling algorithm, Jobs are scheduled when they arrive in the system. Since the cloud environment is a heterogeneous system and the speed of each processor varies quickly, the on-line mode heuristic scheduling algorithms are more appropriate for a cloud environment.

**Steps in Resource Scheduling:**

Resource management in Cloud includes two stages: i) resource provisioning and ii) resource scheduling. Resource provisioning is defined to be the stage to identify adequate resources for a given workload based on QOS requirements described by cloud consumers whereas resource scheduling is mapping and execution of cloud consumer workloads based on selected resources through resource provisioning as shown in Fig. 2. Firstly, cloud consumer submits request for workload execution in the form of workload details. Based on these details broker (resource provision or) finds the suitable resource(s) for a given workload and determines the feasibility of provisioning of resources based on QOS requirements. Broker sends requests to resource scheduler for scheduling after successful provisioning of resources. Other responsibilities of broker include: release of extra resources to resource pool, contains information of provisioned resources and monitor performance to add or remove resources. After resource provisioning, resource scheduling is done in second stage. All the provisioned resources are kept in resource queue while other remaining resources are in resource pool. Submitted workloads are processed in workload queue. In this stage, scheduling agent maps the provisioned resources to given workload(s), execute the workload(s) and release the resources back to resources pool after successful completion of workload. Based on QOS requirements, scheduling of resources for adequate workloads is a challenging issue. For an efficient scheduling of resources, it is necessary to consider the QOS requirements. There is a need to uncover the research challenges in resource scheduling to execute the workloads without affecting other QOS requirements.

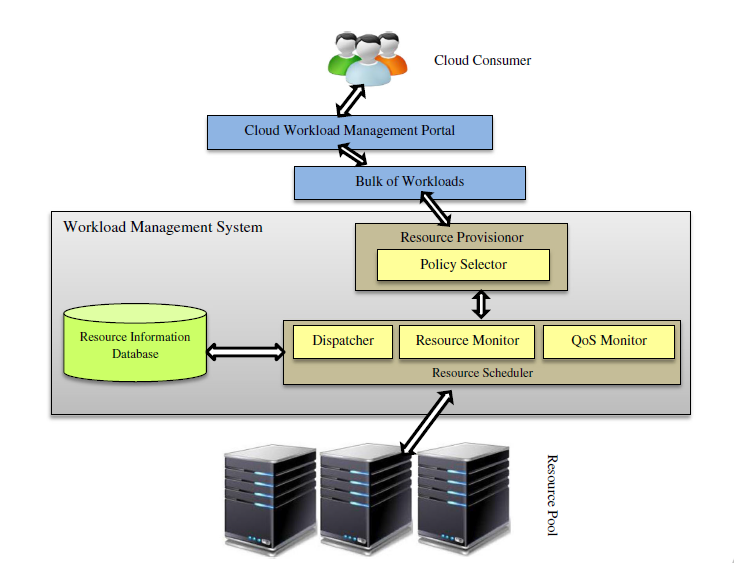


Figure – 2 Resource scheduling in cloud

**How Can I Schedule Resources in Visual Planning?**

In Visual Planning, a resourceschedule will show which resource is assigned to which task. It lists the individual resources of a category (for instance, all employees, or all machines) and shows what each of those resources is assigned to do on a timeline that can be displayed at an hourly, daily, weekly or monthly scale. Individual work items such as tasks or jobs are called events in Visual Planning and shown as blocks of color on the schedule. Events can also be non-working to represent unavailability (vacations for employees, planned maintenance for equipment, etc.). If multiple types of resources are used, users can easily switch from one resource schedule view to another, or display several schedules at once. This is called multi**-**resourcescheduling.

**Applications of control theory to task scheduling on a cloud**

Control theory has been used to design adaptive resource management for many classes of applications, including power management, task scheduling, QoS adaptation in Web servers, and load balancing. The classical feedback control methods are used in all these cases to regulate the key operating parameters of the system based on measurement of the system output; the feedback control in these methods assumes a linear time-invariant system model and a closed-loop controller. This controller is based on an open-loop system transfer function that satisfies stability and sensitivity constraints.

A technique to design self-managing systems based on concepts from control theory is discussed in. The technique allows multiple QoS objectives and operating constraints to be expressed as a cost function and can be applied to stand-alone or distributed Web servers, database servers, high-performance application servers, and even mobile/embedded systems. The following discussion considers a single processor serving a stream of input requests. We attempt to minimize a cost function that reflects the response time and the power consumption. Our goal is to illustrate the methodology for optimal resource management based on control theory concepts. The analysis is intricate and cannot be easily extended to a collection of servers.

**Control Theory Principles.** We start our discussion with a brief overview of control theory principles one could use for optimal resource allocation. Optimal control generates a sequence of control inputs over a look-ahead horizon while estimating changes in operating conditions. A convex cost function has argument  image, the state at step image, and image, the control vector; this cost function is minimized, subject to the constraints imposed by the system dynamics. The discrete-time optimal control problem is to determine the sequence of control variables image to minimize the expression

image(1)

Where  image is the cost function of the final step, image, and image is a time-varying cost function at the intermediate step image over the horizon image. The minimization is subject to the constraints

image(2)

Where  image, the system state at time image, is a function of image, the state at time image, and of image, the input at time image; in general, the function image is time-varying; thus, its superscript.

One of the techniques to solve this problem is based on the *Lagrange multiplier* method of finding the extremes (minima or maxima) of a function subject to constrains. More precisely, if we want to maximize the function image subject to the constraintimage, we introduce a Lagrange multiplier image. Then we study the function

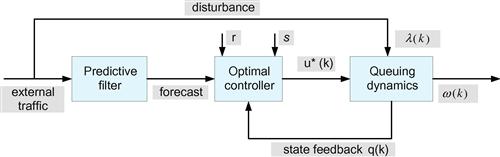
image(3)

A necessary condition for the optimality is that image is a stationary point for image. In other words,

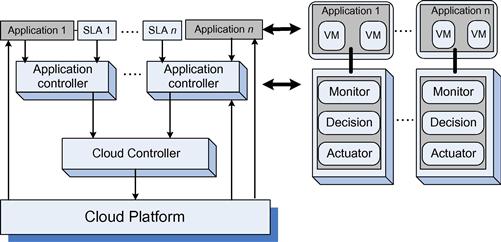
image(4)

The Lagrange multiplier at time step image is image and we solve Eq. [(6.4)](https://learning.oreilly.com/library/view/cloud-computing/9780124046276/xhtml/CHP006.html#E0020) as an unconstrained optimization problem. We define an adjoint cost function that includes the original state constraints as the Hamiltonian functionimage, then we construct the adjoint system consisting of the original state equation and the *costate equation* governing the Lagrange multiplier. Thus, we define a two-point boundary problem[3](https://learning.oreilly.com/library/view/cloud-computing/9780124046276/xhtml/CHP006.html#FN3); the state image develops forward in time whereas the co state occurs backward in time.

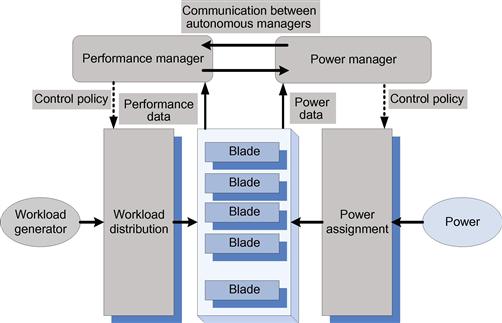
**A Model Capturing Both QoS and Energy Consumption for a Single-Server System.** Now we turn our attention to the case of a single processor serving a stream of input requests. To compute the optimal inputs over a finite horizon, the controller in figure - 3 uses feedback regarding the current state, as well as an estimation of the future disturbance due to the environment. The control task is solved as a state regulation problem updating the initial and final states of the control horizon.



**FIGURE 3** The structure of an optimal controller. The controller uses the feedback regarding the current state as well as the estimation of the future disturbance due to environment to compute the optimal inputs over a finite horizon. The two parameters image and image are the weighting factors of the performance index.



**FIGURE 4** A two-level control architecture. Application controllers and cloud controllers work in concert.



**FIGURE 5** Autonomous performance and power managers cooperate to ensure SLA prescribed performance and energy optimization. They are fed with performance and power data and implement the performance and power management policies, respectively.

We use a simple queuing model to estimate the response time. Requests for service at processor image are processed on a first-come, first-served (FCFS) basis. We do not assume a priori distributions of the arrival process and of the service process; instead, we use the estimate image of the arrival rate image at time image. We also assume that the processor can operate at frequencies image in the range imageand call image the time to process a request at time image when the processor operates at the highest frequency in the range, image. Then we define the scaling factor image and we express an estimate of the processing rate image as image.

The behavior of a single processor is modeled as a nonlinear, time-varying, discrete-time state equation. If image is the sampling period, defined as the time difference between two consecutive observations of the system, e.g., the one at time image and the one at time image, then the size of the queue at time image is

image (5)

The first term, image, is the size of the input queue at time image, and the second one is the difference between the number of requests arriving during the sampling period, image, and those processed during the same interval.

The response time image is the sum of the waiting time and the processing time of the requests

image(6)

Indeed, the total number of requests in the system is image and the departure rate is image.

We want to capture both the QoS and the energy consumption, since both affect the cost of providing the service. A utility function, such as the one depicted in Figure 6, captures the rewards as well as the penalties specified by the service-level agreement for the response time. In our queuing model the utility is a function of the size of the queue; it can be expressed as a quadratic function of the response time

image (7)

with image, the response time set point and image, the initial value of the queue length. The energy consumption is a quadratic function of the frequency

image (8)

The two parameters image and image are weights for the two components of the cost, the one derived from the utility function and the second from the energy consumption. We have to pay a penalty for the requests left in the queue at the end of the control horizon, a quadratic function of the queue length

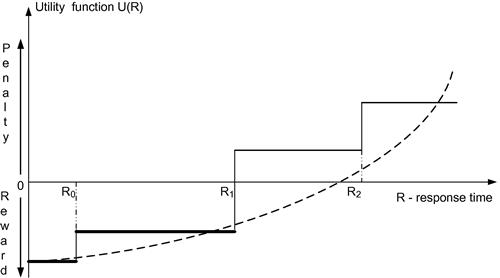
image (9)

The performance measure of interest is a cost expressed as

image (10)

The problem is to find the optimal control image and the finite time horizon image such that the trajectory of the system subject to optimal control is image, and the cost image in Eq. [(10)](https://learning.oreilly.com/library/view/cloud-computing/9780124046276/xhtml/CHP006.html#E0050) is minimized subject to the following constraints

image (11)



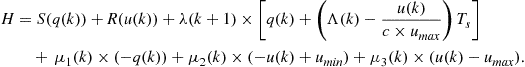
**FIGURE 6** The utility function image is a series of step functions with jumps corresponding to the response time, image, when the reward and the penalty levels change according to the SLA. The dotted line shows a quadratic approximation of the utility function.

When the state trajectory image corresponding to the control image satisfies the constraints

image (12)

then the pair image is called a *feasible state*. If the pair minimizes Eq. [(10)](https://learning.oreilly.com/library/view/cloud-computing/9780124046276/xhtml/CHP006.html#E0050), then the pair is *optimal*.

The Hamiltonian image in our example is

 (13)

According to Pontryagin’s minimum principle,[4](https://learning.oreilly.com/library/view/cloud-computing/9780124046276/xhtml/CHP006.html" \l "FN4) the necessary condition for a sequence of feasible pairs to be optimal pairs is the existence of a sequence of costates image and a Lagrange multiplier image such that

image(14)

where the Lagrange multipliers, image, reflect the sensitivity of the cost function to the queue length at time image and the boundary constraints and satisfy several conditions

image(15)

image(16)

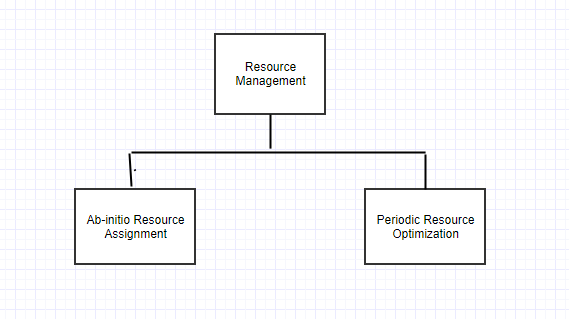
image(17)

The extension of the techniques for optimal resource management from a single system to a cloud with a very large number of servers is a rather challenging area of research. The problem is even harder when, instead of transaction-based processing, the cloud applications require the implementation of a complex workflow.

Taxonomy on Cloud Re-

source Management

**Taxonomy on Cloud Re-source Management**

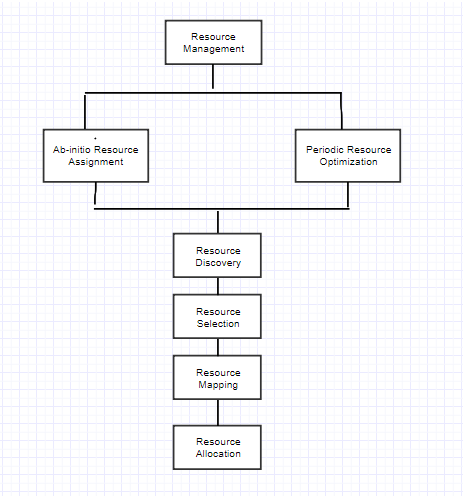
****

**Figure 7:** Taxonomy on Resource management cloud

The goal of resource management in cloud com-putting is to provide high availability of resources, sharing of resources, fulﬁlling time variant service model, providing eﬃciency and reliability on resource usage. From the cloud computing perspective, resource management is a process which eﬀectively and eﬃciently manages above mentioned resources as well as providing QoS guarantees to cloud consumers. This section gives Taxonomy on Cloud Resource Management (Refer Figure 2). The taxonomy is presented as a whole sequential process in two phases.

**Phase 1: Ab-initio Resource Assignment**

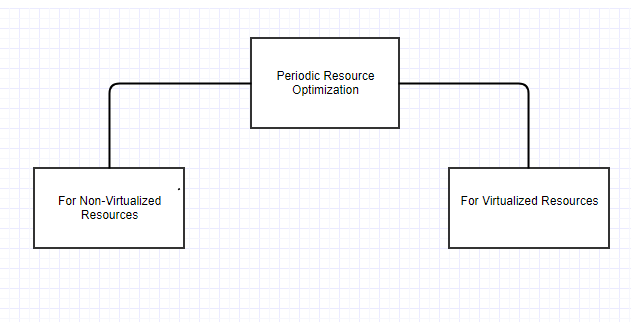
It is initial resource assignment, in a manner that resources are requested by application (on behalf of cloud consumers) ﬁrst time. Figure 3 shows several sequential steps which needs to be followed for completion of this phase



**Figure 8:** Ab-initio Resource Assignment

1. **Request Identiﬁcation**: This is the ﬁrst and foremost step in Ab-initio Resource Assignment. In this step, various resources will be identiﬁed by cloud providers.
2. **Resource Gathering / Resource Formation**: After identiﬁcation of resources in step 1, gathering or formation of resources will take place. This step will identify available resources. This step may also prepare custom resources.
3. **Resource Brokering**: This step is negotiation of resources with cloud consumers to make sure that they are available as per requirement.
4. **Resource Discovery:** This step will logically group various resources as per the requirements of cloud consumers.
5. **Resource Selection:** This step is to choose best resources among available resources for requirements provided by cloud consumers.
6. **Resource Mapping:** This step will map virtual resources with physical resources (like node link etc) provided by cloud providers.
7. **Resource Allocation:** This step will allocate / distribute resources to the cloud consumers. It’s main goal is to satisfy cloud consumers’ need and revenue generation for cloud providers.

**Phase 2: Periodic Resource Optimization**

****

**Figure 8:** Periodic Resource Optimization

As name suggest this is a phase where resource management is done at regular intervals once phase1 is completed. Here periodic resource optimization is presented as a process for two diﬀerent categories of resources which are non-virtualized re-sources and virtualized resources. The non-virtualized re-sources are also called as physical resources. For both categories of resources, periodic resource optimization contains similar steps. The only diﬀerence is that virtualized resources can be assembled together as per the resource requirement and can be disassembled also. So periodic resource optimization for virtualized resources contains two steps more compared to non-virtualized resources which are Re-source Bundling and Resource Fragmentation.

**1. For Non-virtualized Resources**

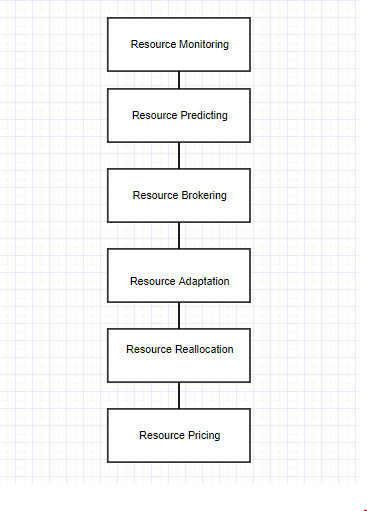
**(a) Resource Monitoring:** Resource Monitoring is the ﬁrst and crucial step in Periodic Resource Optimization. Various non-virtualized cloud resources are monitored to analyze utilization of resources. This step will also monitor availability of free resources for future purpose. The major issue with cloud resource monitoring is to identify and deﬁne metrics/parameters for it.

**(b) Resource Modeling / Resource Prediction:** This step will predict the various non-virtualized resources required by cloud consumer applications. This is one of the complex step as cloud resources are not uniform in nature. Due to this non uniformity, it is very diﬃcult to predict resource requirement for peak periods and as well as for non-peak periods.(c) Resource Brokering: This step is negotiation of non-virtualized resources with cloud consumers to make sure that they are available as per requirement.

**(d) Resource Adaptation:** As per the requirements of cloud consumers, non-virtualized cloud resources can be scaled up or scaled down. This step may increase cost from cloud provider perspective.

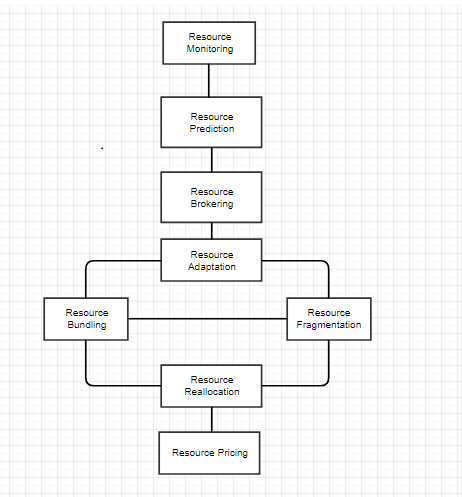
**(e) Resource Reallocation:** This step will real-locate / redistribute resources to the cloud consumers. It’s main goal is to satisfy cloud consumers’ need and revenue generation for cloud providers.

**(f) Resource Pricing:** It is one of the most important step from cloud providers and cloud consumers perspective. Based on cloud resource usage pricing will be done.



**Figure 9:** Periodic Resource Optimization (For Non-Virtualized Resources)

**2. For Virtualized Resources**



**Figure 10:** Periodic Resource Optimization (For Virtualized Resources)

**(a) Resource Monitoring:** Resource Monitoring is the ﬁrst and crucial step in Periodic Resource Optimization. Various virtualized cloud resources are monitored to analyze utilization of resources. This step will also monitor availability of free resources for future purpose. The major issue with cloud resource monitoring is to identify and deﬁne metrics / parameters for it.

**(b) Resource Modeling / Resource Prediction:** This step will predict the various virtualized resources required by cloud consumer applications. This is one of the complex step as resources are not uniform in nature. Due to this non uniformity, it is very diﬃcult to predict resource requirement for peak periods and as well as for non-peak periods.

**(c) Resource Brokering:** This step is negotiation of virtualized resources with cloud consumers to make sure that they are available as per requirement.

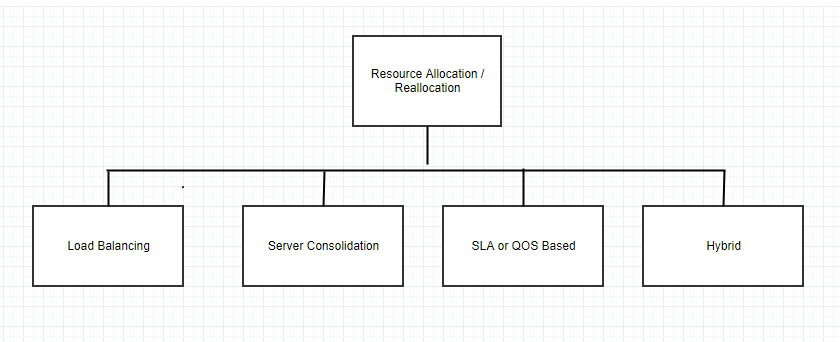
**(d) Resource Adaptation:** As per the requirements of cloud consumers, virtualized cloud resources can be scaled up or scaled down. This step may increase cost from cloud provider perspective.

**(e) Resource Bundling:** As per the requirement various non-virtualized resources can be bundled into virtualized resources.

**(f) Resource Fragmentation:** Various virtualized resources needs to be fragmented to make non virtualized resources free. After this step various non-virtualized resources can be bundled in to virtualized resources as a part of resource bundling.

**(g) Resource Reallocation:** This step will real-locate / redistribute resources to the cloud consumers. It’s main goal is to satisfy cloud consumers’ need and revenue generation for cloud providers.

**(h) Resource Pricing:** It is one of the most important step from cloud providers and cloud consumers perspective. Based on cloud resource usage pricing will be done.



**Figure 11:** Resource Allocation/Reallocation

Resource Allocation and Resource Reallocation can be done based on below mentioned broadly classiﬁed policies as shown in above Figure:

1. Load Balancing

2. Server Consolidation

3. SLA / QoS based

4. Hybrid

**Summary and Comparisons of Resource Management Algorithms:** Cloud resource management process is very complex in nature. In above sections, the whole cloud resource management process had been clearly divided among various steps / techniques which distinguishes all of them from one another. Below is the summary on various resource management techniques.

Table 1: Comparison of Resource Management Algorithm:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Serial No | Publication  (Year) | Techniques /  Algorithms | Tools  and/or  workload  used | Type | Future work  and/or gaps  in existing  technologies |
| 1 | Proceedings  of the World  Congress on  Engineering  and Computer  Science (2011) | Power Aware  Load Balancing  Algorithm | Eucalyptus | Resource  Allocation  / Reallocation (Load  Balancing) | Energy Savings  not considered. |
| 2 | 4th International IEEE  Conference  on Utility  and Cloud  Computing  (2011) | Dynamic Re-  source Allocation for Spot  Instances | Amazon  EC2 | Resource  Allocation /  Reallocation | Customers perspective and  bidding behavior is not  considered. |
| 3 | IEEE Trans-  actions (2012 | Optimal Allocation of Virtual  Resources using  Mixed Integer  Programming  (MIP | Simulator for  Controlling  Virtual Infrastructures  (CVI-Sim) | Resource  Allocation /  Reallocation | Implementation  of proposed  frame work. |
| 4 | ELSEVIER-  Information  Sciences  (2014) | Combinatorial  Double Auction Resource  Allocation  (CDARA) | CloudSim | Resource  Allocation /  Reallocation | Experiments  were done on  simulators,  not on real  environments. |
| 5 | ELSEVIER-  Procedia  Computer  Science (2016) | Power and Load  Aware VM Allocation Policy | CloudSim | Resource  Allocation /  Reallocation | Experiments  were done on  simulators,  not on real  environments |
| 6 | Springer  (2010) | Resource Monitoring Model for  Cloud Computing | Linux  C/C++  and Java | Reliable Re-  source Discovery is future  work. |  |
| 7 | J Grid  computing Springer  (2015 | IaaS Mon | Nagios /  OpenStack | Resource  Monitoring | Integration of  both tools. |
| 8 | IEEE/ACM  (2010) | Dynamic Re-  source Pricing | Planet Lab | Resource  Pricing | Scalability is a  issue |

**References**

[1] M. Armbrust, A. Fox, R. Griﬃth, A. D. Joseph,R. Katz, A. Konwinski, G. Lee, D. Patterson,A. Rabkin, I. Stoica, and M. Zaharia. A view ofcloud computing. Commun. ACM, 53(4):50–58,Apr. 2010.

[2] O. A. Ben-Yehuda, M. Ben-Yehuda, A. Schuster,and D. Tsafrir. The resource-as-a-service (raas)cloud. In Presented as part of the, 2012.

[3] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg,and I. Brandic. Cloud computing and emergingit platforms: Vision, hype, and reality for deliv-ering computing as the 5th utility. Future Gen-eration computer systems, 25(6):599–616, 2009.

[4] D. Ergu, G. Kou, Y. Peng, Y. Shi, and Y. Shi.The analytic hierarchy process: task schedul-ing and resource allocation in cloud computing environment. The Journal of Supercomputing,64(3):835–848, 2013.

[5] I. Foster, Y. Zhao, I. Raicu, and S. Lu. Cloudcomputing and grid computing 360-degree com-pared. In Grid Computing Environments Work-shop, 2008. GCE’08, pages 1–10. Ieee, 2008.

[6] J. M. Galloway, K. L. Smith, and S. S. Vrbsky.Power aware load balancing for cloud comput-ing. In Proceedings of the World Congress onEngineering and Computer Science, volume 1,pages 19–21, 2011.

[7] J. Ge, B. Zhang, and Y. Fang. Research on theresource monitoring model under cloud comput-ing environment. In Web Information Systemsand Mining, pages 111–118. Springer, 2010.

[8] G. E. Gon¸calves, P. T. Endo, T. Cordeiro,A. Palhares, D. Sadok, J. Kelner, B. Melander,and J. Mangs. Resource allocation in clouds:concepts, tools and research challenges. XXIXSBRC-Gramado-RS, 2011.

[9] R. L. Grossman. The case for cloud computing.IT professional, 11(2):23–27, 2009.

[10] J. Gutierrez-Aguado, J. M. A. Calero, and W. D.Villanueva. Iaasmon: Monitoring architecturefor public cloud computing data centers. Jour-nal of Grid Computing, pages 1–15, 2016.

[11] A. Hameed, A. Khoshkbarforoushha, R. Ran-jan, P. P. Jayaraman, J. Kolodziej, P. Bal-aji, S. Zeadally, Q. M. Malluhi, N. Tziritas,A. Vishnu, et al. A survey and taxonomy on en-ergy eﬃcient resource allocation techniques forcloud computing systems. Computing, pages 1–24, 2014.

[12] B. Jennings and R. Stadler. Resource manage-ment in clouds: Survey and research challenges.Journal of Network and Systems Management,23(3):567–619, 2015.[13] R. S. Jha and P. Gupta. Power & load awareresource allocation policy for hybrid cloud. Pro-cedia Computer Science, 78:350–357, 2016.

[14] S. S. Manvi and G. K. Shyam. Resource man-agement for infrastructure as a service (iaas) incloud computing: A survey. Journal of Networkand Computer Applications, 41:424–440, 2014.

[15] P. Mell and T. Grance. The nist deﬁnition ofcloud computing. 2011

[16] M. Mihailescu and Y. M. Teo. Dynamic resourcepricing on federated clouds. In Proceedings of the2010 10th IEEE/ACM International Conferenceon Cluster, Cloud and Grid Computing, pages513–517. IEEE Computer Society, 2010.

[17] M. H. Mohamaddiah, A. Abdullah, S. Subra-maniam, and M. Hussin. A survey on resourceallocation and monitoring in cloud computing.International Journal of Machine Learning andComputing, 4(1):31, 2014.

[18] S. Mustafa, B. Nazir, A. Hayat, S. A. Madani,et al. Resource management in cloud computing:Taxonomy, prospects, and challenges. Comput-ers & Electrical Engineering, 47:186–203, 2015.

[19] G. Pallis. Cloud computing: the new frontier ofinternet computing. IEEE Internet Computing,14(5):70, 2010.

[20] C. Papagianni, A. Leivadeas, S. Papavassiliou,V. Maglaris, C. Cervello-Pastor, and A. Monje.On the optimal allocation of virtual resources incloud computing networks. Computers, IEEETransactions on, 62(6):1060–1071, 2013.

[21] S. M. Parikh. A survey on cloud computingresource allocation techniques. In Engineering(NUiCONE), 2013 Nirma University Interna-tional Conference on, pages 1–5. IEEE, 2013.

[22] V. K. Reddy, B. T. Rao, and L. Reddy. Re-search issues in cloud computing. Global Jour-nal of Computer Science and Technology, 11(11)