CHAPTER 2
LITERATURE REVIEW

2.1 INTRODUCTION

Cloud computing is extensively applied for future generation of computing infrastructure. Cloud mobile services guarantee the flexible services for different ranges of business customer necessities which are isolated in form of infrastructure. Cloud mobile services assure flexible services to the various business customer requirements. Elasticity in cloud is a term which used to analyze cloud that provides different service levels based on changing needs. Elasticity of the application is the one of the most important process in cloud mobile services for effective mapping on the cloud zone. Elasticity in cloud computing is the capacity for the mobile services to adjust its cost requirements in terms of processing in relation to operational requirements.

Cloud computing maintain elastic on demand distribution of server resources, while supporting large number of hardware configurations at different price levels. However, the cost effective ability is the significant task on the different cloud mobile services. Energy efficiency, load balancing and cost factor plays the important role to capacity planning in cloud computing. A significant improvement in cloud computing has resulted in the introduction of new energy and cost efficient techniques with load balanced property. In order to determine the elasticity of cloud computing, their applications have been extensively developed with the help of literature.

2.2 ELASTICITY PROPERTY OF CLOUD

An elastic application model by Xinwen Zhang et al. (2011) allows faultless and cloud resources utilization for improving the ability of resource-constrained mobile devices. The salient features comprise division of one application into many components termed weblets and dynamic adjustment of weblet implementation configuration. When
weblet is platform independent or platform reliant, execution location is apparent that executes on mobile device. An elastic application increases mobile device ability with dynamic execution configuration with device’s status. The motivation increases the elastic applications and structural design with typical elasticity patterns and cost models to establish the elasticity patterns.

The elasticity is capability for measuring the resources up and down to gather the application requirements. An elastic scaling approach by Rui Han (2014) uses cost-aware rule to recognize and observe the bottlenecks in multi-tier cloud-based applications. An adaptive scaling algorithm minimizes the costs acquired by cloud infrastructure services to balance the applications at bottleneck tiers and plan intelligent platform which computerizes the scaling method.

2.3 ENERGY OPTIMIZATION OF MOBILE COMPUTING

Heterogeneous cloud radio access networks (H-CRANs) designed by Mugen Peng et al. (2015) increases the spectral and energy efficiency. The remote radio heads (RRHs) presents high data rates for users with high quality of service (QoS) needs while high-power node (HPN) assures seamless coverage. For reducing the interference and increasing the energy efficiency (EE) results in H-CRANs, user association is categorized with RRH/HPN. A soft fractional frequency reuse (S-FFR) is improved. With RRH/HPN association limitation and S-FFR, energy-efficient optimization issues are addressed with resource assignment and power allocation for orthogonal-frequency-division.

Mobile cloud computing (MCC) emerged as novel model that enables offloading computation-intensive with computing platform in cloud to capacity-limited thin client devices. An enhanced architecture is planned with collaboration of thin clients and desktop or laptop computers termed as the thick clients for improving cloud access. A
new genetic approach is designed by Pham Phuoc Hung and Eui-Nam Huh (2015) for task scheduling where the processing time is reduced.

Cloud service composition comprises many tasks like discovery, compatibility verifying, and operation are difficult methods. The users identify complex process for choosing the best composition. Service composition in Cloud increases the demands by many users with knowledge applications used across geographical locations. The key constrains in choosing the mixture of virtual appliances and infrastructure services that are compatible and guarantee user with indistinct preferences. A framework and algorithms reduces the Cloud service composition for inexperienced users by Amir Vahid Dastjerdi and Rajkumar Buyya (2014). An ontology-based approach examines Cloud service compatibility through reasoning on expert knowledge.

An expert diagnosis system is designed by Kevin C. Tseng et al. (2014) depending on cloud computing. It also categorizes user’s fitness level depending on supervised machine learning methods. The system creates the customized diagnoses consistent with user’s physiological data. An elastic algorithm with Poisson distribution assigns the computation resources. It calculates necessary resources in future consistent with exponential moving average of past examinations.

### 2.3.1 Energy Efficient Cloud Computing

A new framework and principles are designed by Anton Beloglazov et al. (2012) for energy-efficient Cloud computing. Resource provisioning and allocation algorithms are designed for energy-efficient organization of cloud computing situations. Energy-aware distribution heuristics data center resources increase energy efficiency of data center while distributing with Quality of Service (QoS). Many principles are designed for energy-efficient organization of Clouds.
The main aim is to minimize the energy usage of mobile device when increasing the service accessibility for users. Fuzzy vertical handoff algorithm by Anuradha Ravi and Sateesh K. Peddoju (2015) uses the handoff from resource while increasing the energy consumption of device or the connection time reduction with resource. A survey is carried out by Lian Wang et al. (2014) on low-power design of location based applications (LBAs). LBAs and many locating sensing methods are designed. An energy saving technique with locating technologies is studied. Minimization of location updating queries with the trajectory data is revealed. Many cloud-based schemes design energy-efficient locating technology through controlling the cloud.

Cloud data centers utilize the large amount of electrical energy. For green cloud computing, providers reduce the energy consumption while executing QoS. In cloud environments, minimum energy consumption model is designed by A. Horri and Gh. Dastghaibyfard (2015) for time-shared policy in virtualization layer. The cost and energy utilization of time-shared policy were formed from the real system.

A decentralized approach with scalable and energy-efficient organization of virtual machine (VM) instances is provisioned by Michael Pantazoglou et al. (2016) with an enterprise clouds. The computation resources of data center are classified into the hypercube structure. The hypercube faultlessly increases and decreases as resources are inserted or taken away for variations in many provisioned VM instances. Every node operates and controls own workload through relating the distributed load balancing rules and algorithms.

With the growth of mobile cloud computing (MCC), several applications and services are accessible on mobile devices. The constrained battery power of mobile devices is the key problem on user experience. When common applications in mobile cloud environments get increased, LBAs design new inherent constraints with efficient
energy. With severity of the problem, many energy-efficient locating sensing mechanisms are studied by Xiao Ma et al. (2012).

Many IT Executives and CEOs are not concerned in changing the services because of the risks with security and privacy. The security issues are addressed in adaptability of mobile cloud computing model. The essential efforts taken by Abdul Nasir Khan et al. (2013) designed secure mobile cloud computing environments and infrastructure.

To utilize the redundancy level and to increase the download speed, an efficient concurrent method is designed by Nader Mohamed et al. (2013) for downloading the large files from Cloud data servers and traditional FTP servers. A DDFTP use accessibility of replicated files on distributed servers for increasing file download times by simultaneous downloads of file blocks from opposite directions. DDFTP failed to need the coordination between servers and reliability characteristics of TCP. DDFTP presents effective load balancing between many heterogeneous data servers with less overhead.

2.3.2 Applications of Cloud Computing

With large development of mobile applications and cloud computing ideas, mobile cloud computing (MCC) are designed by Shraddha A. Jalan et al. (2014) in mobile services. MCC combines the cloud computing into mobile environment and addresses security issues in mobile computing. The essential problem in mobile cloud computing is end to end delay in planning the request.

The methods designed by Saeid Abolfazlia et al. (2014) examine RMA’s and features. RMAs are differentiated from mobile applications and Rich Internet Applications (RIAs). The variation between Rich User eXperience (RUX) in desktop and mobile computers are considered. The inward resemblance and difference between
RMAs and RIAs increases smartphone-centric applications. Many issues are studied with acceptance of RMAs and developed taxonomy. The majority of issues are addressed from intrinsic characteristics of mobile devices and heterogeneity in environment increases the mobile computing.

Hadoop distributed file system (HDFS) by Bo Dong et al. (2012) maintains the Internet services. Many small files need essential load on NameNode of HDFS. The connections between small files are failed to consider for data placement. In HDFS, threshold point between large and small files is established. With file correlation features, files are divided into three types. They are: structural files, logical files and independent files. An optimized technique increases access effectiveness of small files on HDFS.

In scientific cloud workflows, many application data are collected in distributed data centers. For storing the data, data manager chooses the data centers where the data exist. When any of the tasks demands for many datasets in various data centers, the large volume of data action is a key demand. A matrix based k-means clustering plan is designed by Dong Yuan et al. (2010) for data placement in cloud workflows. The plan has two algorithms which combine the datasets in k data centers in workflow build-time stage. It also groups the generated datasets to suitable data centers in runtime stage.

An autonomic cloud environment collects the health data and sends Cloud depends information repository. The repository uses the analysis on data by software services in Cloud. For calculating the software design, prototype system employs testbed on particular case. The group of electrocardiogram (ECG) data is attained by Suraj Pandeya et al. (2012) at real-time from volunteers for performing the ECG beat analysis.
2.4 LOAD BALANCING ON MOBILE CLOUD SERVERS

Cloud services manage concurrent requests and allow servers to present load balancing capacity with application traffic to attain the information rapidly and exactly. Static load balancing or server response times are used to compute the load balancing ability that causes the server to load randomly. A dynamic annexed balance technique planned by Shang-Liang Chen et al. (2016) addresses the existing issues. Cloud load balancing (CLB) considers the server processing power and computer loading.

A cloud data center has heterogeneous commodity servers hosting virtual machines (VMs) with various specifications and fluctuating resource usages. It resulted in resource utilization imbalance in servers which leads to degradation and violations to service level agreements (SLA). A collaborative agent-based problem solving technique planned by J. Octavio Gutierrez-Garcia and Adrian Ramirez-Nafarrate (2015) balances workloads through VM live migration. The agents are provided with migration heuristics to find out where VMs are migrated.

A new technique is designed by Klaithem Al Nuaimi et al. (2015) addresses the cloud storage problems and design fast load balancing algorithm. The approach is depending on concurrent twofold direction download of files from multiple cloud nodes. The file separation is accumulated on cloud that presents enhanced optimization to cloud storage utilization. The partial replication is employed for guaranteeing the reliability and accessibility of data. The main aim is to increase the results and increase the storage utilization by DaaS on cloud. The algorithm addresses fully replicate large data sets that use precious space on cloud nodes.

A transformation-based optimization framework (ToF) is designed by Amelie Chi Zhou and Bingsheng He (2014) for workflows in cloud. ToF creates six workflow transformation processes. An arbitrary results and cost optimization process are denoted
as transformation plan. All transformations create large optimization space. A cost model
guided planner identifies the optimized transformation for predefined objective.

Two partition-aware fat-tree routing algorithm called pFTree are designed by
Feroz Zahid et al. (2015) for InfiniBand-based HPC systems. pFTree include provider
defined partition-wise policies which manage nodes in many partitions to share the
network resources with other one. A weighted version of pFTree routing algorithm
divides and equalizes the load across network links.

2.4.1 Scheduling of Workloads in Cloud

A new job scheduling plan called Weighted Scheduling Strategy (WSS) by
Najme Mansouria et al. (2013) employs hierarchical scheduling for minimizing the
search time. It takes many jobs waiting in queue, the location of essential data for job
and computing capacity of sites. A dynamic data replication plan termed Enhanced
Dynamic Hierarchical Replication (EDHR) increases the file access time. The plan is an
improved version of Dynamic Hierarchical Replication plan. It also employs economic
model for file removal when there is not sufficient space for replica. The best replica
placement is essential one from replication and minimizes the storage cost and job
execution time.

The results of cloud computing services are analyzed by Alexandru Iosup et al.
(2011) for scientific computing workloads. The computing workloads presence of Many-
Task Computing (MTC) users are quantified who use insecurely joined applications with
many tasks. An evaluation of commercial cloud computing services with Amazon EC2 is
the largest commercial cloud.

The task scheduling issues are addressed by Zhi-Hui Zhan et al. (2014) in cloud
computing through load balance aware genetic algorithm (LAGA) with Minmin and
Max-min techniques. Task scheduling issues in cloud computing is essential and
demanding one. Time load balance (TLB) model identifies the fitness function with makespan that executes best on load balance with equivalent makespan.

A Semi-Elastic Cluster (SEC) computing model is designed by Shuangcheng Niu et al. (2016) for associations to resize virtual cloud-based cluster. A batch scheduling and resource scaling plans are carried out through SEC and online reserved instance provisioning algorithm with job history. An individual user is obtained and controlled the cloud resources resulting in average job wait time. The overhead of preserving the shared cloud cases is exposed in few seconds.

Workflow Scheduling (WFS) aimed on task allocation for attaining the preferred workload balancing through the optimal utilization of existing resources. A significant principle and system distribution structure are taken to address WFS issues in cloud computing through presenting many services to cloud users. The main aim is to choose the suitable cost-aware WFS techniques by Ehab Nabil Alkhanak et al. (2015) from pool of choices. The cost-aware relevant demands of WFS in cloud computing are divided depending on Quality of Service (QoS) results and system use.

Task-level scheduling algorithms are designed by Yang Wang and Wei Shi (2014) with budget and deadline limitations for MapReduce jobs on provisioned heterogeneous machines in cloud platforms. A total monetary budget B is provided by joining in-stage local greedy algorithm and dynamic programming (DP) techniques. A global optimal scheduling algorithm is attained with minimum scheduling length within $O(kB^2)$. For increasing the efficiency level, two greedy algorithms termed Global Greedy Budget (GGB) and Gradual Refinement (GR) are planned with many greedy plans. In GGB, idea of local greedy algorithm is extended to an effective global distribution of budget with lesser scheduling length as goal whilst in GR. DP algorithm is used in minimum budget allocation where the solutions are processed.
A provably-efficient online scheduling algorithm called GreFar by Marco Polverini et al (2014) optimizes the energy cost and equality between many organizations dependent on queuing delay limitations. It also satisfies the maximum server inlet temperature limitations. GreFar failed to require statistical information of workload or electricity prices. It also reduces the cost near the optimal offline algorithm with future information. An interconnection networks is important part in parallel computing system performances. A new interconnection network termed exchanged crossed cube (ECQ) is designed by Keqiu Li et al. (2013). ECQ has better results than hypercube variations with lesser diameter and cost factor.

MCC allows mobile devices to offload their applications and enhances application types on mobile devices as well as on quality of service. MCC architectures partition the designs into four types. Hybrid Local Mobile Cloud Model (HLMCM) is designed by Xianglin Wei et al. (2013) through increasing the Cloudlet architecture. After addressing the application scheduling issues in HLMCM, Hybrid Ant Colony algorithm based Application Scheduling (HACAS) algorithm is introduced.

A heuristic scheduling algorithm termed hyper-heuristic scheduling algorithms (HHSA) designed by Chun-Wei Tsai (2014) identifies the scheduling result for cloud computing systems. The diversity detection and detection operators are used to find out low-level heuristic employed in identifying the candidate keys. The data centers include many machines with capacities and energy consumption features. The workloads running in data centers comprise many applications with dissimilar precedence, ideas and resource needs. Heterogeneity-Aware Resource Management System (HARMONY) is designed by Qi Zhang et al (2014) for dynamic capacity provisioning in cloud computing.
atmosphere. K-means clustering algorithm splits the workload into different task with same features of resource and performance needs.

The optimally redirecting user requests issues in cloud-centric media network (CCMN) are addressed by Jianhua Tang et al. (2014) with multiple destination VMs. The request arrival process controls between the normal and flash crowd modes to form the user requests to CCMN. The trade-offs in flash crowd detection delay and false alarm frequency is calculated at VMs. In every request arrival mode, the optimal redirection policy is identified for each VM. It is task of VM's service cost with requests transmitted to VMs and no redirection to VMs with prices higher than threshold price.

The user needs for cloud computing resources are classified through insecurity and indistinctness. For quality enhancement of cloud computing service, the traditional standards like cost and bandwidth are satisfied. A dynamic resource scheduling technique depending on fuzzy control theory is designed by Zhijia Chen et al. (2015). The connection with resource availability and the resource needs are terminated. A fuzzy control theory identifies the match between user needs and resources availability.

### 2.4.2 Load Balancing Techniques

A geographical load balancing techniques are designed by Muhammad Abdullah Adnan et al. (2012) for data centers hosting cloud computation for minimizing energy cost through using electricity price variations across areas. The flexibility from SLAs distinguishes the workloads in bounded latency needs and cost savings for geographical load balancing. The process of workload performed in all data center are studied and migrated to additional data centers for energy saving. An offline mode creation for geographical load balancing problem is addressed with dynamic delay and online algorithms. The algorithms recognize workload duty to data centers and migration of workload connecting data centers to adjust with electricity price variation.
Software-defined networking (SDN) is the network technologies that recognize the separation between control and data planes. Future wireless networks require control division and data planes. SDN technique manages the large growth of mobile data traffic. A single controller in future wireless network experiences the potential scalability issues. Multiple controllers control the large wide-area wireless network where load balance issues of multicontroller are determined. A multicontroller load balancing approach termed HybridFlow is designed by Haipeng Yao (2015) in software-defined wireless networks. It adopts the distribution and centralization technique and plans double threshold technique to allot the load.

Load balancing algorithms and job allocations are issues in resource organization of future internet. A load balancing model is designed by Shaoyi Song et al. (2014) for future internet. The static load balancing issues are addressed in the model as noncooperative game between users and joint game between processors. Load balancing algorithm is designed for computing center. Load balancing is an optimization issue used by the strategy for different demands. A new ant colony based algorithm is designed by Santanu Dam et al. (2014) balances the load through finding under loaded node.

Load balancing is essential features with sharing of resources where no overloading takes place at any machine. Many algorithms are designed for identifying the answer for load balancing issues. Cloud computing is not same as other one and an individual load balancing algorithm are designed to address the needs. An Autonomous Agent Based Load Balancing Algorithm (A2LB) planned by Aarti Singh et al. (2015) presents dynamic load balancing for cloud computing.

A Taskbased System Load Balancing method with Particle Swarm Optimization (TBSLBPSO) by Fahimeh Ramezani et al. (2013) has better load balancing through transmitting the tasks from overloaded VM. An optimization model is planned to reduce
the additional tasks by Particle Swarm Optimization (PSO). The cloud simulator (Cloudsim) package is expanded and employed PSO as task scheduling model. TBSLB-PSO technique minimizes the time consumption for load balancing process. A new heuristic approach called LB-BC (Load Balancing based on Bayes and Clustering) by Jia Zhao et al. (2016) designs the idea with load balancing in long-term process. LB-BC creates less about physical hosts to attain task deployment approach with global search capability.

A centralized cloud-based multimedia system (CMS) by Chun-Cheng Lin and Der-Jiunn Deng (2014) includes resource manager that allocates clients requests for multimedia service tasks to server clusters. Every cluster head sends the task to servers of server cluster. An efficient load balancing algorithm for CMS sends the multimedia service task load on servers for sending out the multimedia data between server clusters and clients.

A centralized CMS includes the resource manager, cluster heads, and server clusters. The resource manager assigns clients demand for multimedia service to server clusters. Every cluster head sends the allocated task to servers of server cluster. An efficient load balancing algorithm is designed by Chun-Cheng Lin and Der-Jiunn Deng (2013) for CMS. It increases the multimedia service task load on servers with less cost for sending the multimedia data under few limitations. In dynamic multi-service, server cluster manages multimedia tasks and every client requests many multimedia services at different time.

Cloud computing is a new model of service in distributed systems to identify the advantages and disadvantages in implementing the scientific applications. Divisible Load Theory (DLT) planned by Monir Abdullah and Mohamed Othman (2013) reduces the processing time for scheduling jobs in cloud environments. A closed-form solution for
load fractions is allocated to all processor. The jobs are scheduled where the cloud provider increases maximum benefit for Quality of Service (QoS) user's job. Load balancing algorithm is designed by Wenbo Chen et al. (2015) computes complete loading consistent with the CPU use and memory consumption. A load balancing solution depending on SDN technology is used in Open Flow network.

2.4.3 Task Execution and Load Distribution in Cloud

An optimal algorithm by Rubing Duan (2014) reduces the task execution length with resources and payment budget. The upper bound of cloud task length is taken with workload prediction errors and hostload prediction errors. The worst-case task execution result is predictable and raises the Quality of Service. A dynamic version for algorithm changes the load dynamics above the task execution progress and increase the resource utilization.

FastRAQ is a fast approach by Xiaochun Yun et al. (2014) for range-aggregate queries in big data atmosphere. FastRAQ partitions the big data into individual ones partitions with balanced partitioning algorithm and create local estimation sketch for all division. If range-aggregate query request occurs, FastRAQ resulted in better performances in reducing the local estimates from all partitions. FastRAQ includes $O(1)$ time difficulty for data modernizations and $O(N/P \times B)$ time complexity for range-aggregate queries in which $N$ denotes number of distinctive tuples for dimensions, $P$ denotes the division number and $B$ represents the bucket number in histogram.

The power and constrained load distribution techniques are designed by Junwei Cao et al. (2014) for cloud computing in large-scale data centers. An optimal power allocation and load distribution issues are solved using the heterogeneous multicore server processors. The key objective is to plan an optimal power allocation and load distribution for many servers in cloud as optimization issues. The large-scale data centers
is different multivariable optimization issues that search the power-performance tradeoff through fixing one factor and reducing an additional one from optimal load distribution viewpoints.

2.5 COST REDUCTION ON MOBILE CLOUD COMPUTING

A generic mixed-integer linear programming model designed by Xu Gong et al. (2015) permits the job scheduling on single machine for reducing the energy cost beyond due date. A surface grinding machine uses scheduling method that reduces the greenhouse gas emissions in peak time periods through changing the production load to off-peak periods.

Data replication is used to improve the data accessibility and system consistency. Replica placement includes the possible node to duplicate data consistent with the network latency and user request. Replica selection includes the choice of best replica location to use the data for job implementation in data grid. Many replica placement and selection algorithms calculate and examine many parameters. Many replica placement and selection plans are designed by R. Kingsy Grace et al. (2014) with merits and demerits.

A new highly decentralized information accountability structure designed by Smitha Sundareswaran et al. (2012) maintains the actual utilization of user’s data in cloud. An object-centered approach includes the logging method with user’s data and policies. JAR programmable abilities form active and traveling object. It uses the user’s data for trigger verification and logging local to JARs.

For maintaining scalable service and data migration, collaborative integrity verification mechanism design in hybrid clouds is studied. Many cloud service providers to collaboratively store and preserve the clients' data. A collaborative provable data possession scheme by Yan Zhu et al. (2012) implemented homomorphic verifiable
response techniques and hash index hierarchy. The security of scheme is based on multi-prover zero-knowledge proof system that guarantees the knowledge soundness and zero-knowledge.

An optimal algorithm reduces the task execution length with resources and payment budget by Sheng Di et al. (2014). The upper bound of cloud task length is designed with workload prediction errors and host load prediction errors. Worst-case task execution results are predictable for increasing the quality of service. An algorithm is extended to learn the load dynamics in the task execution progress for increasing the resource utilization.

2.5.1 Cost Effective Cloud Computing

Quanlu Zhang et al. (2015) planned cost-efficient multi-cloud data hosting scheme with high availability. Privacy-preserving mechanism enables public auditing on joint data stored in cloud. The ring signatures calculate the verification information to review the reliability of shared data. The identity of signer on every block in shared data maintained private from third party auditor (TPA) to authorize the integrity of shared data lacking the file retrieval.

A cost efficient and VM allocation model is designed by Mohammad Mehedi Hassan and Atif Alamri (2014) depending on Nash bargaining solution. The designed mechanism minimizes the cost of running servers as well as assures QoS demand and increase the resource usage in many dimensions of server resources. Cloud computing is managed by an engine called Internet data center (IDC). The energy cost reduction issues are addressed for IDCs in electricity markets. The geographic and temporal change of energy price reduces the energy cost for IDCs. A new design is planned by Jianying Luo et al. (2015) for spatial and temporal load balancing. The algorithms contain less
computational complexity and need accuracy in electricity price estimation for user requests.

Three power-saving policies are used in cloud systems to reduce the server idle power. A cost function is designed by Yi-Ju Chiang (2015) with the operational costs, system congestion and mode-switching. The function identifies the optimal service rate and mode-switching restriction to minimize the cost in response time with many arrival rates. An efficient green control (EGC) algorithm is planned for addressing the constrained optimization issues and tradeoffs in systems with different power-saving policies.

A new MapReduce cloud service model called Cura is planned by Balaji Palanisamy et al. (2015) for cost-effective MapReduce in cloud. By comparing with existing MapReduce cloud services, Cura contains many distinctive advantages. Cura is planned for presenting cost-effective key to control MapReduce production workloads with interactive jobs. Cura controls MapReduce profile to generate the cluster configuration for jobs. Cura executes effective resource allocation technique that minimizes the resource utilization cost in cloud. Cura controls distinctive optimization opportunity when managing the workloads. Through managing the multiplexing with existing cloud resources between jobs depending on job needs. Cura consumes lesser resource usage costs for jobs.

For progress of user job in cloud computing environment, new resource provisioning scheme is designed by Zhiping Peng (2015) depending on reinforcement learning and queuing theory. With Segmentation Service Level Agreement (SSLA) and Utilization Unit Time Cost (UUTC) ideas, the resource provisioning issue is taken as sequential decision problem.
The cloud database is a new model that maintains many Internet-based applications. Though, the adoption needs the solution of information privacy issues. Luca Ferretti et al. (2014) designed an adaptive encryption of public cloud databases that provides an option to tradeoff between data confidentiality level and flexibility of cloud database formation at planning time. An original cost model is changed for cloud database services calculations and encrypted cases with difference of cloud prices and workloads in medium-term period.

Many companies are responsible to overprovision with IT infrastructures with buffer performance to eliminate the destruction of service level agreements. The energetic scaling in IT infrastructures assures cost-saving potential. The cost-aware operation of on-demand provision system is managed and designed provisioning model by Bendler et al. (2013) to optimally extent the cloud at any time.

Software-as-a-Service (SaaS) is software distribution model in cloud computing and denotes highest software layer in cloud stack. Many cloud services providers charge for resource utilization and needed to form resource efficient applications. The main method used is multi-tenant design of SaaS applications. The tenant-based resource allocation model issues on cost-effectiveness of SaaS systems are studied by Wojciech Stolarz and Marek Woda (2013). The tenant-based resource allocation model is used for under-optimal resource utilization.

2.5.2 Cost Efficient Resource Allocation and Sharing

Many large-scale parallel workflow functions are allocated on heterogeneous computing systems which is NP-complete problem that addresses many QoS needs. The
scheduling issues of large-scale applications classified the homogeneous and coexisting tasks with key sources of bottlenecks.

Cloud computing is an internet based computing that allows the sharing of services. The users without physical control of large size of outsourced data create data integrity protection in cloud computing for users with computing resources. Security in cloud is achieved with the data block sooner than sending to cloud. The signing is carried out through the Boneh–Lynn–Shacham (BLS) algorithm by Jachak K. Bet al. (2012). For guaranteeing the data accuracy, external auditor termed TPA is taken for cloud user to authenticate integrity of data stored in cloud. By public key based homomorphic authenticator and random masking privacy preserving, public auditing is obtained.

The large distributed system enables the users for distributing and trading the resources. The users obtain many resources from one or more cloud providers for inadequate time interval with fixed price. Federated cloud is method for sharing the resources that improves the scalability level. For increasing the resource allocation efficiency, agent based resource allocation technique is employed by Haresh M V et al. (2011). The user failed to identify who is cloud service provider and where resources exist.

An outline of cloud computing problems is addressed by Vrunda J. Patel and Hitesh A. Bheda (2014) for distributed dynamic consolidation of VMs in Cloud data centers. The key objective is to increase utilization of computing resources and reduce energy usage with workload QoS constraints. Dynamic VM consolidation manages the fine-grained fluctuations in application workloads and changes VMs by migration to reduce active physical nodes count. Energy consumption is minimized through disabling and reactivating the physical nodes to address the resource needs. The approach is
distributed and effective in controlling the energy results trade-off. DVFS idea with migration methods improves the efficiency level of energy management and adaption method on real-time services.

An analytical model is designed by Hamzeh Khazaei et al. (2013) and authenticated an independent simulation model. Task service times are formed with probability distribution. The model is measure for deterioration of results because of the workload at every node. The model is designed for computation of result indicators and distribution of many tasks in system. The cloud center results are enhanced when incoming requests are divided with the coefficient change of service time and batch size.

Efficient predicting tasks are essential component of tasks development and resource allocation in Cloud Computing Environment. A framework is designed by Kun Gao et al. (2014) for maintaining knowledge discovery application organization in cloud atmosphere and holistic approach forecast the application. Rough sets theory maintains reduct and calculate execution time prediction. The heuristic reduct algorithm is depending on the frequencies of attributes in discernibility matrix. Dynamic information of many knowledge discovery tools are added with data sources to Cloud Computing Environment.

Cloud computing changes traditional software systems and perform through planning the utility-based model. The consolidation of new model in both enterprises and academia required review where the IT resources are employed. Cloud computing are employed with accessible resources. Cloud usage increases the computing infrastructure capacity is Desktop Grids by Rodrigo N. Calheiros et al. (2012). By increasing the Desktop Grid infrastructures with Cloud resources, QoS is required for the users and encouraging Desktop Grids as possible platform for application execution.
Data centers as architectural and efficient block of cloud computing are essential one to Information and Communication Technology (ICT) sector. A Data Center Network (DCN) comprises the communicational backbone of data center determining the boundary results for cloud infrastructure. DCN are robust to the malfunction and doubts to distribute necessary Quality-of-Service (QoS) level and please SLA. A multilayered graph modeling of many DCNs is designed by Kashif Bilal et al. (2013) and classical robustness parameters have many failures to execute the analysis. The insufficiencies of classical network robustness parameters calculate DCN robustness. A resource allocation technique is designed by Belen Bermejo et al. (2016) increases system efficiency. The technique is depending on taking decisions at two levels, namely physical machine level and overall system level.

2.6 RESEARCH GAP

Cloud computing presents the cloud mobile services with different elasticity application. The role of elasticity in cloud mobile services provides an efficient mapping application on the cloud zone that matches with the resources that can be allocated with which it actually requires. Many research works has been developed for energy-aware resource allocation by using heuristics method which provides substantial amount of cloud services however failed to improve the energy efficient management on elastic cloud computing environments. Secure mobile cloud computing provides better user privacy but failed to include host trusted domain module for other cloud service providers.

Cloud mobile applications guarantee the flexible services to many business customer requirements which are isolated from infrastructure form. But, elastic
application is not effective on the transparent use of cloud mobile resources where the components are rearranged depend on the workload of business logic. When public cloud integrates all the resource strategies on scientific workloads, the CPU load is improved and it is not efficient in solving the optimization issue in different elasticity cloud applications.

In addition, Elasticity of application is the essential processes in cloud mobile services for effective mapping on the cloud zone. Cloud mobile services assure flexible services to the various business customer requirements. However, cost effective ability is the essential task on many cloud mobile services. Many techniques perform Elastic Application Model on cloud mobile devices but failed to run cloud on many device with minimal service cost.

Energy-Efficient locating sensing method provided better understanding on design principles and challenges in mobile cloud computing. However, energy was not efficiently used on large dataset. The energy-aware resource allocation used heuristics method by providing the cloud services for both resource providers and consumers. But, the QoS and power usage characteristic of the devices is unable to develop the energy-efficient organization on elastic cloud computing environments.

Collaborative provable data possession scheme in cloud environment recognizes homomorphic verifiable reaction methods and perform the verification process. The homomorphic verifiable responses integrate the hash index hierarchy with challenging issues. An approach for data placement policy named Matrix based k-means clustering
clusters the data that are suitable with data centers. But, matrix based k-means clustering consumed larger amount of execution time on cloud workflows of mobile services.

SEC offered insight into virtual cloud-based cluster for efficient maintenance of shared cloud instances. SEC failed to predict the total number of instances within the time interval. Performance Analysis for Many Task Computing (PA-MTC) was designed using grids and parallel production infrastructures aiming at minimizing the cost model of cloud. However, elastic application is not effective on the transparent use of cloud mobile resources where the components are rearranged depends on the workload of business logic. When the public cloud combines all resource plans on scientific workloads, CPU load is raised but not effective in solving the optimization problem in many elasticity cloud applications.

Elastic scaling approach in cloud mobile devices (ESA-CMD) identifies and examines the multi-tier cloud-based bottleneck. Workload-adaptive application executes efficient cloud elasticity management. Though, the approach failed to allocate and deallocate resources in cost effective manner. Elastic Application programming Model is designed for increasing the computing capabilities of Mobile Devices. The model eliminates the limitation of particular mobile platforms through presenting the distributed framework that expands the device into cloud. However, it does not attain the higher flexibility rate in cloud mobile devices. In order to overcome such limitations in cloud mobile services, the proposed methodology is designed.
2.7 CONTRIBUTIONS

The main contribution of the thesis is as follows:

Initially, to attain energy efficient system for the cloud mobile services, a method called Machine Flow based Energy-Power Approximation (MFEPA) is presented. MFEPA is executed for each elastic cloud services for efficient energy-power saving in cloud mobile devices. Multi-grid approximation technique reduces the energy consumption. A look-ahead control is presented to minimize the power consumption in mobile cloud services. The mapping decreases the energy during the avoidable computing load and proves to be effective on the terminal mainframe mobile communications. The Look-ahead control in MFEPA assigns weights to all the users and reduces the power usage on the wireless interface.

Next, to reduce the CPU load in the cloud elasticity environment, Dynamic Prioritized Load Balanced Round Robin (DP-LBRR) framework is planned. The Load Balanced Round Robin (LBRR) algorithm in DP-LBRR structure mainly focused on distributing the load to the CPU aiming at balancing the CPU load rate. The CPU in turn executes different types of requests from various cloud mobile environments by allocating each virtual machine in a cyclic order with the objective of decreasing the latency time. The LBRR algorithm added a Dynamic Priority Load Scheduler into the system for reducing the memory consumption and addresses the optimization problem in the elasticity cloud services using priority value.

Finally, to minimize the cost factor on many cloud mobile devices, Dual-Cost Responsive on Cloud Mobile Services (DRCMS) is designed. The two cost function in DRCMS is denoted in elasticity application. For leasing-cost responsive work, the
mobile cloud Virtual Machine (VM) measures the type of virtual machine in active condition during processing. The shift-cost calculates the shift latency in moving elasticity application. Both leasing and shift cost is reduced using Integer Linear Program (ILP). Dual cost responsive algorithm analyzes the two cost criteria for guiding the profits up and down in the elasticity of cloud services.