J Reliable Intell Environ (2016) 2:37–49
DOI 10.1007/s40860-016-0017-7

REVIEW

Taxonomy and issues for antifragile-based multimedia cloud computing

Syed Fawad Haider[1] · Laraib Abbas[1] · Amjad Ali[2] · Muddesar Iqbal[3] · Imran Raza[2] · Syed Asad Hussain[2] · Doug Young Suh[4]

Received: 19 August 2015 / Accepted: 6 February 2016 / Published online: 27 February 2016

Abstract Cloud computing has become one of the most dynamic and adoptable computing paradigms. Multimedia Cloud Computing (MCC) is one of today's hot research topic. MCC is proven to be a most dynamic and efficient platform for managing a large amount of multimedia contents with maximum deployment of computing and processing resources at the service provider instead of users. Resilience and dependability are two key constituents to assure the reliability and availability of any service in the presence of errors and system failures. The heterogeneous environment of MCC has given rise to various challenges related to resource allocation and task management. Antifragility is a key to such environments, to let the disorder drive the strength of these systems. This paper is mainly divided into three parts. The first part discusses in detail the available state-of-the-art related to resource management under MCC. Similarly, the second part presents the comprehensive literature review on the task management in MCC. The third part presents the critical analysis and open research issues in MCC which help the researcher to define their research objectives in the field of MCC.

Keywords Cloud computing · Quality of service · Antifragility · Quality of experience · Resource allocation

✉ Muddesar Iqbal
director.oric@gmail.com

Syed Fawad Haider
fawad.haider@uog.edu.pk

Laraib Abbas
laraib.abbas@uog.edu.pk

Amjad Ali
amjad.ali@ciitlahore.edu.pk

Imran Raza
iraza@ciitlahore.edu.pk

Syed Asad Hussain
asadhussain@ciitlahore.edu.pk

Doug Young Suh
suh@khu.ac.kr

[1] Faculty of Computing and Information Technology, University of Gujrat, Gujrat, Pakistan

[2] Department of Computer Science, Communication and Networks Research Centre, COMSATS Institute of Information Technology, Lahore, Pakistan

[3] Pak-UK Institute of Innovative Technologies for Disaster Management, University of Gujrat, Gujrat, Pakistan

[4] Department of Electronics and Radio Engineering, College of Electronics and Information, Kyung Hee University, Yongin 446-701, South Korea

1 Introduction

In the current era of technology, multimedia devices are evolving exponentially. Such a rapid growth of multimedia devices has introduced a larger number of multimedia applications and services. Thus, multimedia users over the globe are producing a huge amount of multimedia contents which require high processing and storage capabilities. Hence, processing of such a huge amount of multimedia contents in timely and efficient manner by fulfilling their users' Quality of Service (QoS) requirements is a highly challenging [1–3] task. Moreover, due to diverse nature of multimedia data, anomaly management is also one of the big challenges. Thus, Multimedia Cloud Computing (MCC) paradigm, which is motivated by the heterogeneous nature of multimedia application and services, is different than the other general purpose cloud computing paradigms [4,5] and has capabilities to resolve such issues and challenges [6–8].

Cloud computing is a large-scale distributed paradigm where users can access various application softwares and infrastructure through the Internet as shown in Fig. 1. The service-oriented architecture and elasticity of infrastructure provided by cloud computing is one of the major attractions for organizations to opt it for their services. Cloud computing uses a enhanced optimization and focuses on sharing resources among the cloud users to reduce the overall capital investment as well as operating costs [9–12]. Users can benefit from flexible service models of cloud computing which offer Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), according to their demands and requirements [9,13,14]. However, failure is unavoidable in cloud architecture due to internal or external conditions such as human-made faults, unreliable hardware/software, and natural disasters. Therefore, the idea of using these failures to improve the resilience and maintainability of cloud environment is the basic concept of antifragility [15]. Taleb introduced the concept of antifragility stating that "Antifragile: Things that gain from disorder" [16]. This approach can be used in a cloud environment to increase its flexibility and productivity using different techniques. The authors in [17] introduced and discussed a framework for resilience of computational systems through a simple scenario. The authors interpreted resilience as the emerging result of a dynamic process in which process represents the dynamic interplay between the behaviours exercised by a system and those of the environment it is set to operate in. Similarly, in [18] the authors proposed a model of the fidelity of open systems. The authors interpreted fidelity as the compliance between corresponding figures of interest in two separate but communicating domains. In [19] the authors discussed and presented several examples on ongoing research to employ the concept that instead of designing systems to meet known requirements which always lead to fragile systems at some degree, systems should be designed wherever possible to be antifragile.

Cloud computing is the most adoptable field of computing world which is categorized into three different types: Public, Private, and Hybrid clouds. A public cloud is deployed by a third party and provides services to the end users on demands such as Google and Amazon. Private cloud is deployed by a company for it internal use. Similarly, In Hybrid cloud, a company stores some of its data on public cloud and most secret data are stored on its own private cloud [20]. In the next subsection, we present a comparison study over both computing paradigms.

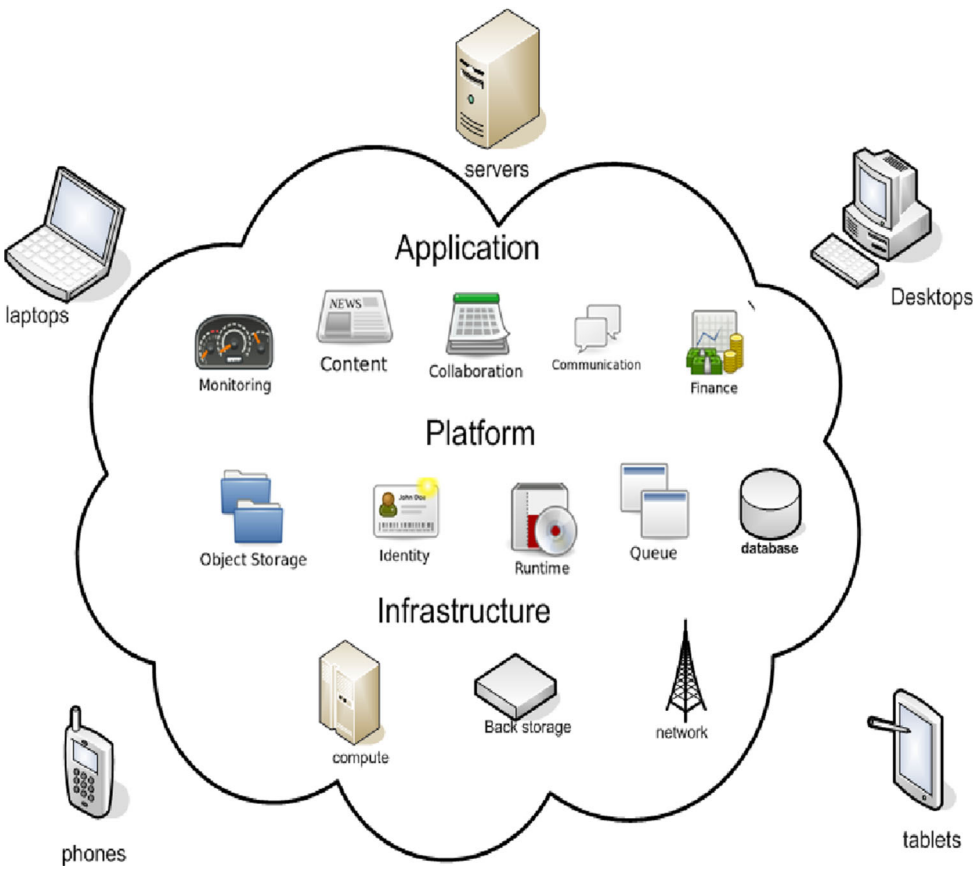


Fig. 1 General cloud computing architecture

1.1 Multimedia cloud vs. conventional cloud computing paradigm

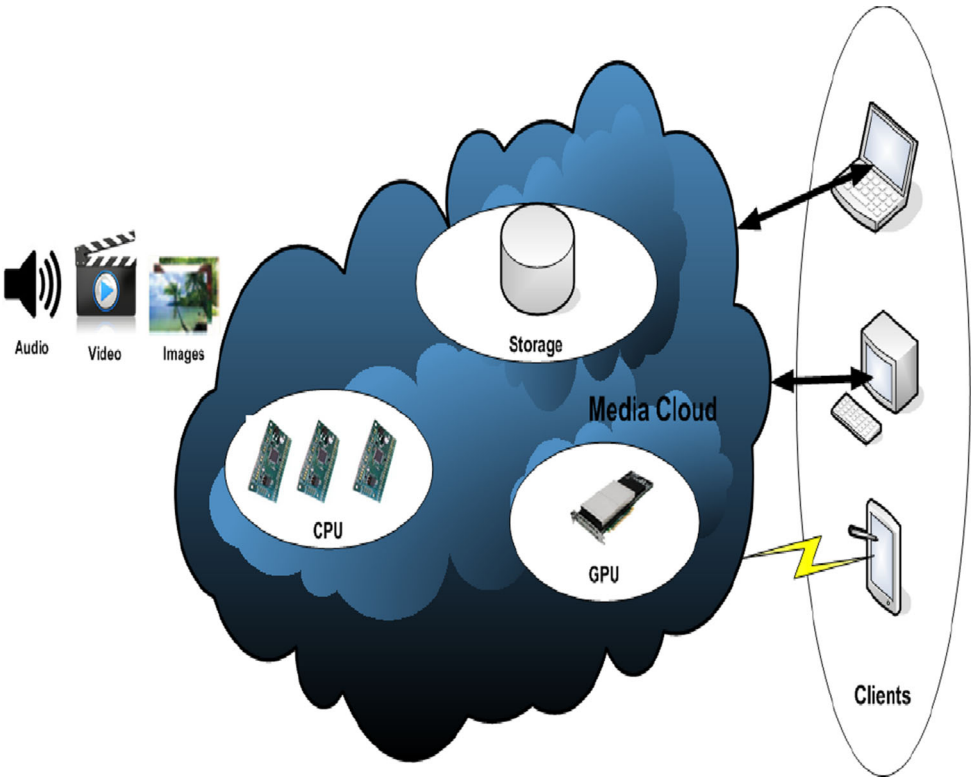
Traditional Cloud Computing (TCC) architecture refers ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources such as servers, networks, storage, applications and services. TCC works as a utility computing where users employ services and pay for what they use. Amazon EC2, Microsoft Azure, and Google App Engine are few examples of Public clouds. TCC handles some key challenges like scalability, QoS, and virtualization which makes the TCC an efficient computing paradigm [21,22].

Due to availability of high computing and storage resources at low cost, cloud is an efficient and optimal option for processing and storing a huge amounts of data [23]. However, due to heterogeneous nature of multimedia devices, application, services, and different QoS requirements of multimedia traffics, the TCC is not an appropriate computing paradigm. Hence, a new efficient and fast computing paradigm, called Multimedia Cloud Computing (MCC), has been introduced, which processes multimedia contents (i.e., images, videos, and graphics, etc.) in a distributed manner and eliminates the installations of media applications on local machines [24]. Therefore, MCC is quite different than TCC due to timely processing and strict QoS requirements of multimedia contents. Multimedia contents require a large amount of computing resources for their efficient and timely processing. Thus, MCC must contain high-performance Graphical Processing Units (GPUs), Central Processing Units (CPUs), and storage capacities a shown in Fig. 2 to perform the efficient and timely processing of multimedia contents [24]. However, TCC and MCC support a few common services and applications but generally their architectures are widely different from each other. Moreover, multimedia applications and services require different types and capacities of processing and storage which motivates to design a cloud that can handle all issues related to multimedia application and service-related issues. The following are two main types of MCC.

Cloud-aware multimedia (Cloud Media) Cloud media [24] is defined as "Multimedia services and applications such as storage, sharing, authoring, mash-up, rendering, retrieval, adaptation, and delivery can efficiently and effectively utilize cloud resources to enhance the Quality of Experience (QoE) of multimedia users". The main characteristics are listed below:

1. Storage is always made available to the cloud users so they can easily share their data everywhere.
2. Cloud can be used efficiently to edit different segments of multimedia contents and combine them.
3. Cloud system can be used efficiently and speedily for rendering and retrieval of multimedia contents.
4. Cloud can be used to transform different media contents and deliver them to the users.

Fig. 2 MCC architecture



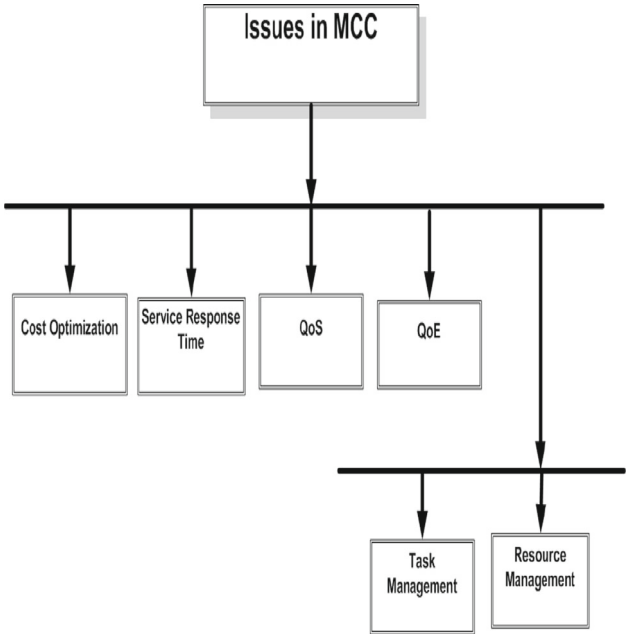


Fig. 3 Taxonomy of MCC

Multimedia-aware cloud (Media Cloud) Media cloud is a new multimedia computing paradigm [24] that ensures multimedia features such as QoS and supports various types of multimedia contents and heterogeneous devices. Many solutions are discussed for task, resource, and QoS management in Media Cloud [24]. To enhance the QoS, storage and processing of multimedia contents are performed by cloud.

In the following sections, we present a comprehensive review of the state-of-the-art related to MCC. This paper mainly focuses on two important issues: (1) task and (2) resource management in MCC. The complete taxonomy of MCC is presented in Fig. 3.

2 Resource management in multimedia cloud computing

In this section, we discuss a comprehensive overview of different techniques for multimedia cloud computing under the category of resource management in MCC. MCC is newly emerged computing paradigm and it's highly challenging because the MCC users required highly efficient and timely computation of multimedia data which makes it different from conventional cloud computing. Resource Management in terms of processing and storage is a key issue in MCC as different multimedia contents require different computing resources.

Queuing model Quality of Experience (QoE) is one of the main objectives of a media cloud provider. Nan et al. [25] introduced an optimal resource allocation scheme for MCC

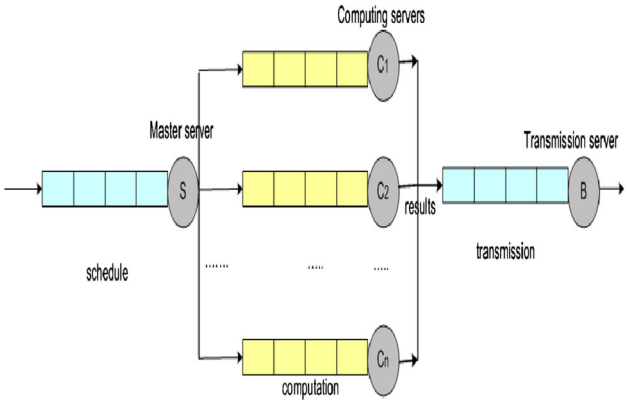


Fig. 4 Queuing model

based on queuing model. There are certain issues such as cost optimization, service response time, QoE, resource allocation cost that severely affect the performance of MCC. Thus, the service response time, that is a time between arrival of user request and its departure, should be minimized to enhance the QoE of the multimedia user.

Single-class/multi-class service Nan et al. [26] also introduced an optimal resource allocation scheme for MCC based on priority services scheme in both single-class and multi-class services case. In their proposed scheme, cloud is designed in the form of a data centre which consists of a master server and group of computing servers. The master server acts like a controller node that receives, schedules, and forwards user requests to the other computing servers which further process these user requests. The master server and the computing servers form a logical tree, in which master server acts like a root node and computing servers are the leaf nodes connected with the root. Schedule rate and computation rates are used to assign the weights to the links between the root and leaf nodes, respectively.

Priority services scheme mainly contains three types of queues; (1) schedule queues, (2) computation queues, and (3) transmission queues as shown in Fig. 4. The schedule queue, installed at master server is responsible for receiving and scheduling user's requests for further processing. After scheduling, the user tasks reach the computation queues of their respective server and wait for execution. When the requests/tasks are processed in computation server they are forwarded to the transmission server, where they wait in transmission queue before they are transmitted to their respective users as shown in Fig. 5. Media cloud can optimally process multimedia applications and services while ensuring their respective users QoS requirements. A number of multimedia applications, i.e. images/videos processing and 3D rendering cannot efficiently process client machines as they need intensive storage and processing capacity, so these applications efficiently process in MCC. There are two major challenges; (1) service response time and (2) resource

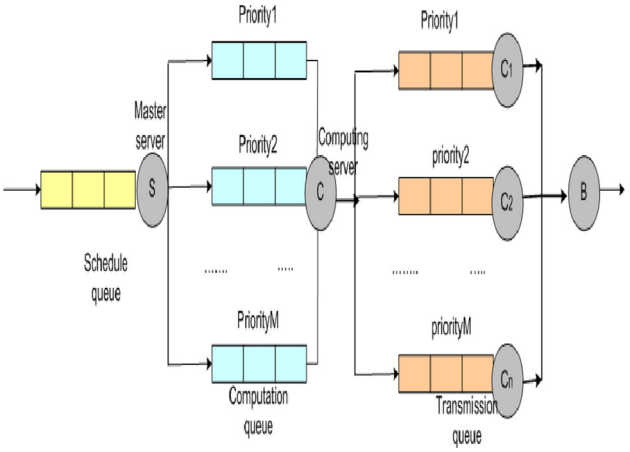


Fig. 5 Priority services-based queuing model

cost optimization that can severely affect the performance of MCC.

Dynamic resource allocation Media cloud enhances the performance of media processing as it worked in dynamic, distributed, parallel, and shared environment [27]. The resource allocation for media tasks in such a dynamic and shared environment is a big challenge to achieve system efficiency and QoS for the media users. Many previously presented designs for media cloud have strived to improve system efficiency but could not enhance the QoS for media tasks. Authors in [27] investigated the properties of media tasks and required resources to process them in a media cloud.

The main research focus of work presented in [27] is to improve the system efficiency and QoS by finding and assigning exact amount of resources for any media task. For this purpose, a dynamic allocation scheme is introduced based on machine learning and previous task history. Survival function controls the resource allocation and grants the QoS to media tasks. This dynamic scheme is compared with the static resource allocation schemes and is found to be the most appropriate scheme for multimedia cloud computing. Moreover, their proposed scheme also addresses the resource allocation problem in cloud based media platform. Media tasks that arrive at multimedia cloud are heterogeneous and have different QoS requirements. They initially proposed a static resource allocation scheme, in which resources are stored in a pool and are assigned to tasks as they arrive. Media task requirements in terms of processing and storage vary quite often, so a machine learning based dynamic resource allocation scheme has been proposed that predicts the exact amount of resources and enhances the performance of media processing.

QoS-based resource allocation Yirui et al. [28] discussed the resource allocation problem in MCC and proposed a QoS-based efficient resource allocation scheme to enhance the performance of MCC. Their proposed scheme considers both the capacity enhancement of media servers and parameters required to enhance the QoS in multimedia cloud. Energy consumption, resource allocation, cost, and completion time should be considered before allocating resources to the media tasks in MCC. Their proposed resource allocation scheme is divided in to three sections; (1) task analysis, (2) cloud broker, and (3) resource manager. When a user request arrives at the cloud, it is analysed to figure out its QoS parameters. The cloud broker then compares the QoS requirements of user's request with the cloud resources and estimates the required resources for the given job. Finally the resource servers assign these tasks to the actual virtual machines which process user's requests.

Authors introduced a utility function in their proposed scheme that is based on the game theory in which they efficiently distribute the cloud resources among all cloud users. The utility function covers the objectives of both the users and cloud providers. The users are concerned with the multimedia task completion time and cost of required resources and the cloud provider's concerns are about energy consumption in the cloud platform. The main objective of utility function is to fulfil user and cloud provider concerns and ultimately gain user's satisfaction and minimize the energy consumption in the cloud platform. This utility function allocates resources to the media task in two phases. In the first phase, arrived media tasks occupy the cloud resources without considering the shared nature of these resources. It eventually creates a conflict among other media tasks, which is eventually resolved by reallocating the cloud resources to optimize the performance of MCC.

Media task QoS based resource allocation Bohai et al. [29] discussed the overall QoS for media task and resource utilization in MCC and proposed an efficient resource allocation scheme based on media task QoS for MCC. The computational weight of the task which is found to fulfils the QoS weight of media task is measured in terms of QoS weight vector and an expected resource vector. Later, the resource similarity vector from expected resources is calculated along with an alternative resource vector by using linear normalization. At the end, service satisfaction by Euclidean distance is formulized and resources are allocated according to the service satisfaction. Their proposed architecture is divided into three components; (1) *task pool, (2) media service manager, and (3) cloud platform*.

The task pool is the component where media tasks, i.e. video, images, and graphics are processed. Media Service Manager (MSM) is responsible for task analysis, scheduling, and resource allocation and assigns appropriate virtual machine to the media task. The MSM is a coordinator between the task pool and cloud platform. The cloud platform is the actual cloud resource that process media tasks. As the allocation of resources to the media task is based on QoS weight vectors so the task analysis module must be

aware of all the physical resources. It can be easily computed based on the expected resource vector. Furthermore, there is heterogeneity in media tasks and cloud platform so allocation and scheduling of resources is based on expected resources vector. MSM selects appropriate resources from resource pool and sends resource information to the virtual machine (VM) module. A VM is created which processes the media task and return the resources when finished.

Cost effective resource allocation MCC computing paradigm is also very effective and efficient solution for E-health systems [30]. E-health systems require high capacity of computing resources, i.e. high processing power, storage capacity, and network bandwidth. Keeping in view the dynamic and delay-sensitive nature of E-health applications, the cloud computing is an optimal option because it can dynamically adjusts the resources for such systems. However, issues like cost effectiveness of resource allocation and energy consumption can still affect the performance for E-health systems. The authors in [30] focus on these issues and propose a cost effective QoS-aware resource allocation scheme for MCC based E-health systems. Their proposed resource allocation scheme overcomes two important goals. The first is to minimize the overhead in cloud platform and guarantees the required resources for VM and given media tasks to finish timely. The second is to minimize the cost and energy consumption to handle all VMs with less number of servers. The servers that are not in use should be powered off. To minimize task completion time and overhead on servers, there should be a trade off between the cost and energy consumption via increasing or decreasing the utilization of servers.

Their proposed idea is based on the Nash Bargaining Game that contains commodities and bargaining as a game theoretic approach. In their proposed cloud resource allocation scheme, the entire datacentre acts as bargaining market and the VMs which process the media tasks act as commodities in the bargaining market. All the servers participate in the market and bargain for their desired commodities. The Nash bargaining game is responsible for assigning the desired commodities to the servers in the market in such a way that the social welfare is achieved and any VMs do not exceed their capacities.

Resource allocation for media streaming applications Amr et al. [31] proposed a prediction based resource allocation scheme for media streaming applications. Their proposed scheme mainly focuses on optimizing the tariff cost of media cloud by minimizing the media cloud reserved resources. The whole media cloud is divided into three modules; (1) *demand forecasting* that predicts the future streaming needs from users based on previous usage patterns, (2) *cloud broker* that allocates the required resources and reserves a few others for specific time and implements the prediction scheme obtained in forecasting module and (3) *media provider* that provides resources and streaming services to the users. Their proposed solution considers the time-discount and non-linear tariff that is changed by a cloud provider for the purpose of resource reservation in cloud. On the basis of prediction of streaming capacity's future demand, the financial cost of Media Cloud Provider (MCP) is minimized by using an algorithm that accumulates the reservation time and amount of cloud resources. Their proposed scheme also ensures that enough resources are being reserved in cloud without any wastage.

Resource allocation for cloud-based video surveillance platform Authors in [32] presented a resource allocation scheme for cloud-based video surveillance environment. Their proposed scheme focuses on optimizing the VM resources to fulfil various types of user services which are provided by MCC. Single service is not effective for user requests but composite services could be efficient and optimal. Their proposed design is the composition of two mechanisms: (1) *Linear Programming Model* and (2) *Heuristic Approach* that can dynamically allocate resources in media cloud. Cost optimization and service response time are being improved by this approach. They tested their proposed prototype initially inside the Amazon cloud under a limited type of extent. The VSS directory is a composition for the users such as fire fighters and security personnel that provide the payment services, analysis, sharing, streaming, and transcoding services through an interface of web browser.

Media-edge cloud Wenwuet et al. in [23] introduced a media-edge cloud (MEC) design for MCC. Their proposed design provides a parallel and distributed processing and Quality of Service schemes. The media cloud is divided into three clusters; *(1) Central Processing Unit (CPU)*, *(2) Storage*, and *(3) Graphics Processing Unit (GPU)*. Media graphic processes are performed on GPU clusters. However, the CPU clusters performs the general media processes and storage is managed by storage clusters. A media cloud proxy is also used to handle the heterogeneous media contents of mobile users.

Resource allocation controller for cloud-based adaptive video streaming Luca et al. [33] presented a resource allocation controller for video streaming in MCC. The aim was to provide high-quality video streaming to end users at minimal distribution cost. A controller is responsible for resource allocation that dynamically archives the resource management. This resource allocation controller contained three modules: (1) *Load Balancer, (2) Resource allocation, and (3) Stream switching adaptation controller (SSAC)*. First two modules are synchronized such that the user sends a request which is assigned to the active server by a load balancer and then adaptive video streaming session is started. SSAC instance is also started with this process. In the mean while, active server sends feedback to RAC through monitor that decides the behaviour of the machine to be ON or OFF. Figure 6 describes this design. To evaluate the performance of the proposed solution, authors used CDNSim after some mod-

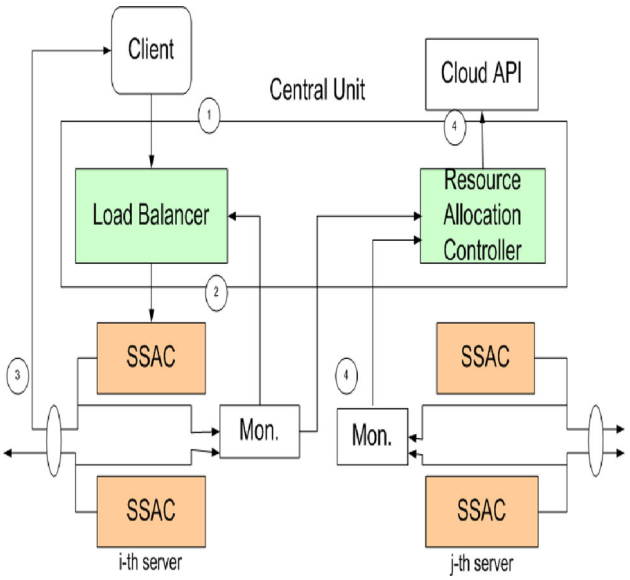


Fig. 6 System overview

ifications on the bases of INET and Omnet++ framework. The following metrics are being measured for the purpose of performance evaluation of resource allocator controller:

- The number of active machines.
- The cost of CPU usage.
- The fraction of streams that obtain the maximum video level.

The simulation is done with the help of realistic work load. The results are taken considering the delivery cost so that their proposed architecture can provide high-quality video streaming to the end users while saving the distribution costs.
Power saving scheme for multimedia streaming services Yi-Wei et al. [34] introduced a power saving based scheme for multimedia streaming services in MCC. The authors proposed a multimedia service architecture for streaming services over embedded system that uses the digital signal processor. The main focus is on delivering high-quality media service in cloud and save the power for mobile devices. The whole design is divided into three parts; (1) *Android OS on hardware platform*, *(2) Power saving scheme*, and (3) *Digital signal processing (DSP) module*. When the user requests for video streaming, the server loads the desired video into memory buffer using HTTP. The multimedia framework is called by application layer to analyse video format and eventually the packets are sent to the DSP for decoding. Finally, the desired video packets are sent to the memory buffer and users, respectively, through advanced ARM Linux hardware platform.

For achieving an efficient embedded system development and to minimize the kernel burning time, the Network File System (NFS) is used to mount the Android File System (AFS). To mount the catalog of the AFS on the local server, TCP/IP protocol stack is being used by NFS. Application files can be removed, added, or modified by the user through a single remote control. Video files are used to test the performance of the hardware platform. The performance of DSP chip, ARM chip, and mixed ARM is compared on system operation state.
QoS-aware data forwarding for multimedia streaming service Seokhoon et al. [23] presented a QoS-aware scheme for data forwarding for Multimedia Streaming Services (MSS). In the proposed scheme, the QDFA synchronization process is used to improve the transmission efficiency with the application of more accurate and precise measures of transmission time under IEEE 1588 standard. To provide better quality, an improved and novel scheme, based on QDFA, uses the techniques of GMPLS, DiffServ, POSIA, and IntServ. The characteristics of background traffic like SMTP and interactive traffic like HTTP is used to burst the QDFA. To evaluate the performance of proposed algorithm, different schemes such IntServ, MPLS, and DiffServ are compared with QDFA. The proposed scheme used 250 destinations for each side and multimedia streaming traffic is used in simulation as a traffic type for P2P hybrid network.
Two-stage approach for task and resource management Biao et al. [35] presented a two-stage scheme for task and resource management in MCC. The proposed approach mainly focuses on resource management, in which it defines the way of assigning VMs to the actual servers and task management where VMs were assigned the tasks. Their proposed heuristic scheme for resource management and task management has been done by queuing mechanism via adding a deadline approach in user requests. Then, task manager and resource allocator optimally minimize the cost and enhance the QoS for multimedia services. The following Fig. 7 describes the design of this scheme.

3 Task management in multimedia cloud computing

In this section, we discuss a comprehensive overview of different techniques for multimedia cloud computing under the category of task management in MCC.
Effective load balancing for cloud based multimedia system Hui et al. [36] highlighted a key challenge that can severely affect the performance of the cloud computing (i.e., assigning accurate amount of cloud resources in a short period of time to the multimedia applications). To overcome this issue an efficient load balancing scheme for cloud-based processing is introduced, which is called a cloud-based multimedia load balancing scheme. This scheme manages the capacity

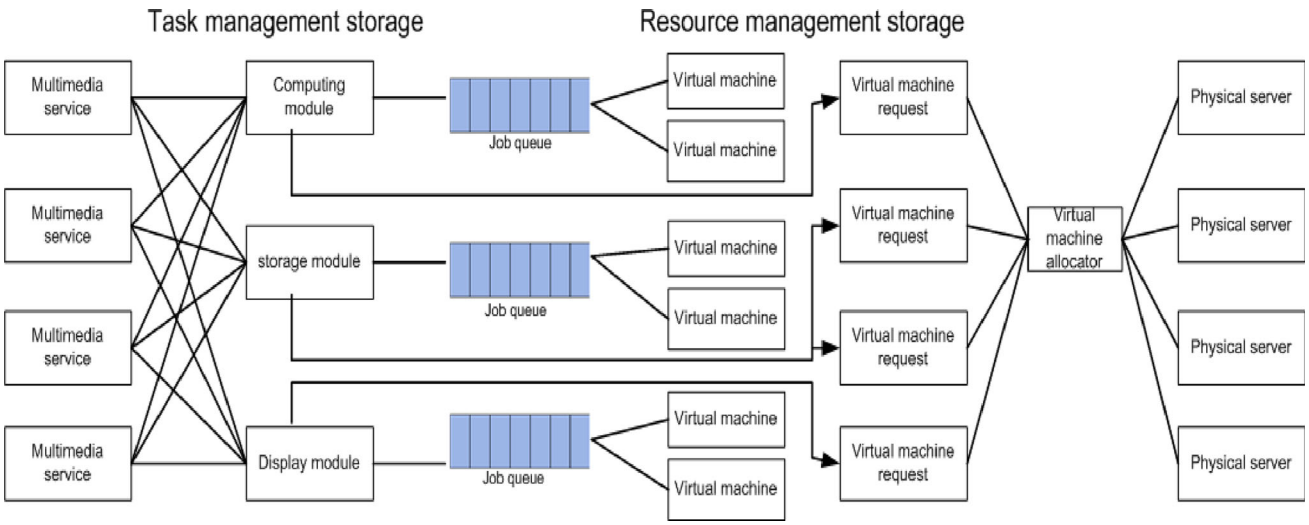


Fig. 7 System overview

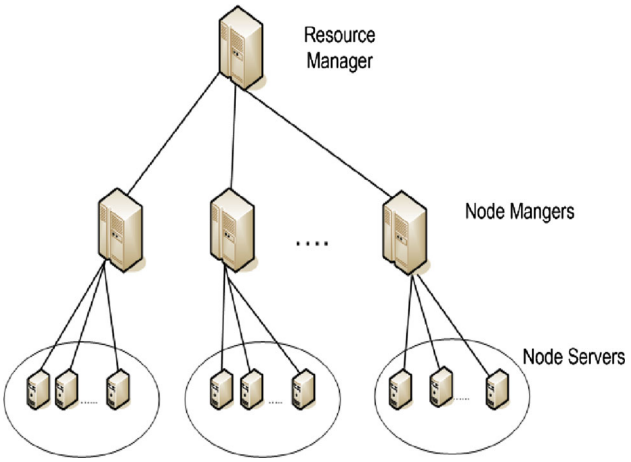


Fig. 8 Workload scheduling for MCC architecture

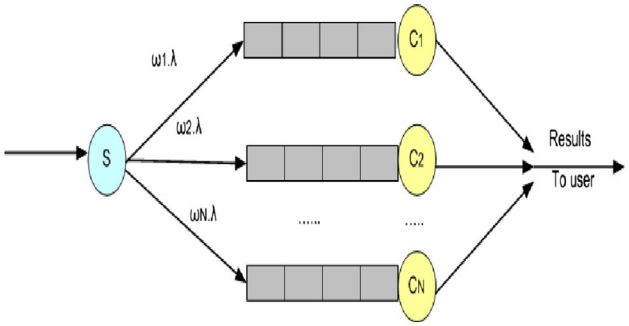


Fig. 9 Workload scheduling for multimedia cloud computing

of individual node in the cloud as well as assigns the exact capacity to the coming traffic over the parallel links.

The system consists of multiple scattered service nodes which process multimedia applications and services over the Internet. The proposed approach is divided into three different layers: (1) *resource manager* that is responsible for assigning appropriate resources; (2) *node manager*, which is responsible for managing node in the system and allocates the tasks, and (3) *node servers* that are responsible for executing multimedia tasks as presented in Fig. 8. In order to achieve required objective the multimedia task is processed in this three-layer system. In the proposed cloud-based multimedia load balancing scheme, the resource manager selects required and appropriate node for media tasks that depend on the need of user media tasks. Finally, the node manager assigns that services node to the media tasks.

Workload scheduling for multimedia cloud computing Multimedia applications and services are time oriented and produce large-scale workload. So, there is need of an efficient workload scheme that can enhance the performance of cloud computing. Different workload scheduling schemes are discussed in [37]. There are two main challenges which can affect the performance of multimedia applications in a cloud platform; (1) service response time and (2) resource cost minimization. The authors in [37] addressed these two important issues and proposed a greedy algorithm for workload scheduling in MCC.

The proposed approach dynamically adjusts the workload in MCC. As the workload is varying with the time, the time domain is divided into two different slots and distributed between these two time slots. The VMs are divided into different classes and each class is dedicated for a specific task, also known as virtual cluster, as presented is Fig. 9. Their proposed workload scheme addressed the service response time and resource cost optimization by formulating their optimized solutions. Service response time is used as important factor to minimize the mean response time by optimizing the workload scheduling weights for different virtual clusters, subject to the queuing stability constraint at

each cluster. The workload conserving constraint, scheduling weight constraints, and resource cost optimization are all about minimizing the total resource cost by jointly optimizing the workload assignments and the allocated VMs in the on-demand reservation schemes. The optimization is carried out with subject to the application response time constraint, queuing stability constraint, VM reservation constraint, workload conserving constraint, and the workload scheduling weight constraint.

Dynamic multiservice load balancing Chun-Cheng et al. [37] presented a hierarchical cloud-based multimedia system that consisted of a resource manager, cluster heads, and server clusters. User requests arrive at resource manager which transfers them to the contended server cluster. Cluster head assigns the appropriate resources to the user requests. The main issue addressed in this cloud-based multimedia system is how to optimally distribute workload in the system without affecting the performance of the multimedia system. To enhance the workload distribution in this system, a Genetic algorithm-based dynamic multi-service load balancing algorithm is proposed.

Authors in [36] proposed a scheme that is based on scheme presented in [38]. Their proposed scheme handles the similar types of multimedia tasks by formulating the dynamic cloud multimedia system (CMS) according to time. Time is further divided into multiple time steps where in each time step, CMS is modeled as a complete weighted graph. CMS components are represented in the form of weighted graph U, in which a set of vertices represents a server clusters. V is a set of vertices which represents the users requests and similarly, E presents the set of edges between U and V. The proposed functions limit the media servers for a specific media task by providing the QoS to the system and minimizing the link costs between client and cloud server clusters. Furthermore, the functionality of Genetic Algorithm (GA) works is based on evolutional theory of population. GA finds the optimal solution from the given solution set. GA also finds an efficient load balancing technique for CMS by dividing the problem into different steps. First, it randomly generates a set of all the possible solutions. Second, it selects the optimal pair of solutions. Third, it compares the selected pair of optimal solution set with the optimal solutions. Based on this idea the proposed dynamic load balancing scheme is divided into four components; (1) initialization, (2) selection crossover, (3) mutation, and (4) repair. These components work in the same way as GA performs its actions.

Cloud-based service architecture for multimedia streaming using Hadoop Myoungjin et al. [39] discussed the multimedia applications and services that require very high processing and computing technologies. There are certain issues like load balancing, fault tolerance, and task management that could affect the performance of media applications. To overcome these issues and enhance the performance of multimedia application, a distributed multimedia streaming system is proposed in [39]. The prime objective of proposed system is to convert the variety of videos in MPEG-4 format which is used to further distribute them among a variety of heterogeneous multimedia users. Their proposed scheme reduces the transcoding time by using Hadoop file system and minimizes the content delivery delays by using streaming distribution algorithm. Hadoop clustering techniques are used to enhance the performance and efficiency of content delivery using Hadoop policies and strategies.

Their proposed system architecture is divided into three modules; (1) the *Hadoop-based distributed multimedia transcoding module (HadoopDMT)*, (2) *Hadoop-based distributed multimedia streaming module* (HadoopDMS), and (3) *cloud multimedia management module (CMM)*. The HadoopDMT transcodes the video into standard format such as MPEG-4 which is compatible with most of the media devices. When a video was transcoded, it moves to the Hadoop distributed file system (HDFS) which is a part of HadoopDMS. Then, HadoopDMS divides the media contents into sub parts and distributes them in the system. These sub parts of media contents are saved on three nodes, so, the data could be made available in case of any failure. The prime purpose of CMM system is to manage tasks, schedule tasks, and load balancing.

Adaptive multimedia cloud computing centre applied on H.264/SVC streaming Multimedia applications have now become highly efficient and mature with the advancement in network bandwidth and internet technologies [40]. Media streaming has become a challenging issue in cloud-based media system. The authors in [40] addressed these issues and proposed an adoptive multimedia system based on H.264/SVC. Their proposed scheme addressed the communication links between the client side and cloud provider as well as the load balancing in the cloud platform. Algorithm in [39] determines the streaming path between the client and cloud provider based on the client side bandwidth and processing power to provide the high-quality video streaming to the clients.

The media cloud proposed architecture is comprised of three different types of nodes: *(1) index node, (2) content node, and (3) streaming nodes.* The working of this media cloud architecture is described as follows:

- When a user request for a video arrives then this request is sent to the index node. As the request reaches the index node, the load balancing mechanisms are executed and asked for the require content holder node.
- Videos are divided into multiple segments and stored on different content nodes. Content nodes are checked to find the required contents containing nodes.

- At client side the index node inquires the streaming nodes for bandwidth requirement and forwards this bandwidth report to required content containing nodes.
- In order to find the load balancing capacities streaming nodes are examined for their hardware resources. After doing this results are handed over to the load balancing model for content capturing, analysis, separation, and streaming.
- Index node asks for setting the quality level to the streaming node and gives information about the required content. So, the streaming node can download the required video.
- The downloaded content is analysed and separated according to the H.264/SVC coding scheme. Finally, the address of selected streaming node is sent to index node and it forwards that address to client side.

QoS/QoE mapping and adjustment model in the cloud-based multimedia infrastructure Cloud-based multimedia system interacts with current IP-based network where it faces QoS and multi-cast service support challenges [41]. The cloud providers ensure the QoS in the system. They neglect user perspective which is an important factor in the design of cloud-based multimedia systems. QoE describes that how cloud-based multimedia system would enhance the usability of the users. QoE is categorized into two further subcategories named objective and subjective. When system is monitored for technical parameters, i.e. throughput, delay, and packet loss, it is called objective approach and when it monitors the system on user opinion, then it is called subjective approach.

In multimedia systems, user perspective has always been lacking. Thus, Wei-Ting et al. [41] proposed a QoE mapping and adjustment model that worked by translating the network QoS parameters into the user QoE under a cloud-based multimedia infrastructure. The proposed model consists of three major sectors and the simulation results showed that the user's QoE and network QoS are consistent with each other. The service provider could use the proposed QoE function to monitor the users' QoE perception and to respond quickly to rectify problems that degraded QoE in the multimedia cloud.
Quality-assured cloud bandwidth auto-scaling for video-on-demand applications The Video on Demand (VoD) providers used cloud computing bandwidth resources to guarantee the availability of VoD to the clients. Authors in [42] proposed a predictive resource auto-scaling system which dynamically booked the minimum bandwidth resources from multiple data centres for the VoD provider to match its short-term demand projections. The proposed architecture enhances the performance of video streaming using cloud-based auto-scaling bandwidth mechanism.

Table 1 In this section, we present the critical analysis of the presented state-of-the-art related to resource allocation and task management and also present the corresponding open research issues

S. no.	Issue	Description
1	QoE	It is the extent to which user accepts the application
2	QoS	It is the efficiency of the system to perform user tasks
3	Cost optimization	It is the process, in which cloud provider minimizes the user's cost to rent the cloud platform
4	Service response time	It is time, in which user request arrive at cloud manager, processed by the cloud and then send result back to user
5	Task management	It is the scheduling of user's tasks in the cloud platform
6	Resource allocation	It is the allocation of resources to incoming user's requests

Their proposed scheme contains certain features as follows: it is predictive, it records the history of bandwidth utilization for each video channel, and estimates the coming requirements of clients. Moreover, a channel interleaving scheme is used for video contents. This scheme provides high quality videos to the clients by ensuring the demanded video quality and guarantees the less utilization of network bandwidth.

The following Table 1 presents the important performance metrics used to evaluate the performance of MCC and their corresponding short description.

4 Future directions and open issues

(1) *Queuing model*: Queuing model is an efficient mechanism but there is a possibility of wastage of resources. As every queue has a its own separate computing unit, it is not possible to keep the computing resource busy all the time as well as queues work-loaded. Hence, more accurate and efficient mechanisms are required which minimize the resource wastage and improve the system performance.
(2) *Priority services-based queuing model*: Priority queuing scheme is an efficient mechanism but there is a high possibility of increased service response time. As every queue is processed according to predefined priority and if any urgent task needs to be processed then it will not be processed until its priority arrives. There is a need to enhance this priority queuing scheme for the same reason.

Table 2 Comparison among presented works

S. no.	QoE	QoS	Cost optimization	Service response time	Task management	Resource allocation
[24]	Yes	Yes	No	No	No	Yes
[25]	No	No	Yes	Yes	No	Yes
[26]	No	No	Yes	Yes	No	Yes
[27]	No	Yes	No	No	No	Yes
[28]	No	Yes	Yes	Yes	No	Yes
[29]	No	Yes	No	No	No	Yes
[30]	No	Yes	No	No	No	Yes
[31]	No	No	Yes	No	No	Yes
[32]	No	Yes	No	No	No	Yes
[33]	No	No	Yes	No	No	Yes
[34]	No	Yes	No	No	No	Yes
[23]	No	Yes	No	No	No	Yes
[35]	No	No	No	No	Yes	Yes
[36]	No	No	No	Yes	Yes	Yes
[37]	No	No	Yes	Yes	Yes	No
[38]	No	No	No	Yes	Yes	No
[39]	No	No	No	No	Yes	No
[40]	No	Yes	No	No	Yes	No
[41]	No	Yes	No	No	Yes	Yes
[42]	No	Yes	No	No	Yes	Yes

(3) *Dynamic resource allocation:* Dynamic resource allocation mechanism is working well in homogeneous environment. However, there is a need to enhance this mechanism for heterogeneous scenarios as well.

(4) *QoS-based resource allocation:* QoS-based resource allocation scheme is an efficient mechanism that can greatly improve the usage and cost of resources. However, service response time increases in resource allocation and re-allocation processes. Therefore, some more efficient ways are required that can improve the service response time under QoS-based resource allocation techniques.

(5) *Media task QoS-based resource allocation:* Media task QoS-based resource allocation scheme can be further improved by adding some other important performance metrics.

(6) *Cost effective QoS-based resource allocation:* Cost effective QoS-based resource allocation scheme jointly deals with the resource cost minimization and QoS-based resource allocation. This scheme has a great potential to greatly minimize the resource. However, there is overhead in migration of virtual machines and physical servers which can be further improved as an enhancement in this architecture.

(7) *Resource allocation for media streaming:* Resource allocation for media streaming services scheme is based on future forecasting and is an efficient mechanism which can improve the resource usage and cost of resources jointly. However, the future forecasting is not much matured. Therefore, prediction function could be improved more by applying some efficient artificial intelligence techniques to obtain more accurate results and better performance.

(8) *Media-edge cloud:* Media-edge cloud computing is an efficient resource allocation mechanism that can improve the QoS. However, there is need to enhance its resource allocation mechanism for better cost and resource utilization.

(9) *Two-stage resource allocations:* Two-stage resource allocation is an efficient mechanism that supports QoS features in MCC. However, it cannot support QoE for the users that can be added as a future work in this scheme to obtain more efficient results.

(10) *Effective load balancing:* Effective load balancing scheme is an efficient scheme for MCC. However, it still has some scalability issues especially under the increased work load scenarios. Therefore, more efficient ways need to be explored to enhance the system overall performance.

(11) *Workload scheduling:* Workload scheduling technique which is mainly based on a greedy algorithm is also an efficient way for load balancing. However, under heavy workload scenarios it does not produce efficient results. Therefore, more improvements and efficient techniques are needed to enhance the system performance.

Table 3 Cloud infrastructures

S. no.	Cloud infrastructure	Multimedia service
[24]	Media edge cloud	Distributed image processing
[25]	Windows Azure	General multimedia services
[26]	Windows Azure	General multimedia services
[27]	MediaPaaS	Video streaming/
[28]	CloudSim	General multimedia services
[29]	CloudSim	General multimedia services
[30]	CloudSim	E-health services
[31]	Numerically/simulation	Multimedia streaming
[32]	Amazon cloud platform	Video surveillance
[33]	CDNsim	Video streaming
[34]	Android multimedia platform	Multimedia streaming
[23]	Hybrid peer-to-peer networks	Multimedia streaming
[35]	Amazon EC2	General multimedia services
[36]	Deployment of cloud infrastructure	General multimedia services
[37]	Amazon EC2	General multimedia services
[38]	Deployment of cloud infrastructure	General multimedia services
[39]	Deployment of Hadoop cloud infrastructure	Multimedia streaming
[40]	H.264/SVC based cloud platform	Multimedia streaming
[41]	NS-2 simulation	Video streaming
[42]	Simulations	Video-on-demand

(12) *Dynamic multi-service load balancing:* Dynamic multi-service load balancing scheme manages the load balancing using genetic algorithms that further can be improved for achieving more accurate results and better system performance.
(13) *Cloud-based service architecture for multimedia streaming using Hadoop:* Kim et al. have evaluated this design on Linux-based self-configured system. Therefore, for better performance analysis and getting more interesting results, this system should also be tested on some other commonly used architecture/systems such as Amazon, and Rackspace compute.

The below Table 2 presents the comprehensive analysis of above discussed state-of-the-art. Similarly, Table 3 presents cloud infrastructures and their corresponding multimedia services and applications.

5 Conclusion

This paper discussed the state-of-the-art under two important performance metrics; (1) resource allocation and (2) task management in newly emerged computing paradigm, known as multimedia cloud computing, which turned the whole computing pattern into service-oriented environment. This newly emerged computing paradigm has high processing capabilities and large storage capacities to process quality of service demanding multimedia applications and services with very strict quality of service requirement. This paper has also highlighted some future directions and open research issues which can help the researchers to define their future research directions in the field of multimedia cloud computing.

References

1. Yuan H, Kuo CCJ, Ahmad I (2010) Energy efficiency in data centers and cloud-based multimedia services: an overview and future directions. In: 2010 international green computing conference. IEEE, New York, pp 375–382
2. Koavchev D, Cao Y, Klamma R (2011) Mobile multimedia cloud computing and the web. In: 2011 workshop on multimedia on the web (MMWeb). IEEE, New York, pp 21–26
3. Chen JL, Wuy SL, Larosa YT, Yang PJ, Li YF (2011) IMS cloud computing architecture for high-quality multimedia applications. In: 2011 7th international wireless communications and mobile computing conference (IWCMC). IEEE, New York, pp 1463–1468
4. Li ZN, Drew MS, Liu J (2014) Cloud computing for multimedia services. In: Fundamentals of multimedia. Springer, New York, pp 645–674
5. Lai CF, Huang YM, Chao HC (2010) DLNA-based multimedia sharing system for OSGI framework with extension to P2P network. Syst J (IEEE) 4(2):262–270
6. Wu Y, Wu C, Li B, Qiu X, Lau F (2011) Cloudmedia: when cloud on demand meets video on demand. In: 2011 31st international conference on distributed computing systems (ICDCS). IEEE, New York, pp 268–277

7. Glitho RH (2011) Cloud-based multimedia conferencing: business model, research agenda, state-of-the-art. In: 2011 IEEE 13th conference on commerce and enterprise computing (CEC). IEEE, New York, pp 226–230
8. Gadea C, Solomon B, Ionescu B, Ionescu D (2011) A collaborative cloud-based multimedia sharing platform for social networking environments. In: 2011 Proceedings of 20th international conference on computer communications and networks (ICCCN). IEEE, New York, pp 1–6
9. Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, Zaharia M (2010) A view of cloud computing. Commun ACM 53(4):50–58
10. Zhang S, Zhang S, Chen X, Huo X (2010) Cloud computing research and development trend. In: Second international conference on future networks, 2010. ICFN'10. IEEE, New York, pp. 93–97
11. Foster I, Zhao Y, Raicu I, Lu S (2008) Cloud computing and grid computing 360-degree compared. In: Grid computing environments workshop, 2008. GCE'08. IEEE, New York, pp 1–10
12. Dikaiakos MD, Katsaros D, Mehra P, Pallis G, Vakali A (2009) Cloud computing: distributed internet computing for IT and scientific research. Internet Comput IEEE 13(5):10–13
13. Patidar S, Rane D, Jain P (2012) A survey paper on cloud computing. In: 2012 second international conference on advanced computing and communication technologies (ACCT). IEEE, New York, pp 394–398
14. Creeger M (2009) Cloud computing: an overview. ACM Queue 7(5):2
15. Arney C (2013) Antifragile: things that gain from disorder. Math Comput Educ 47(3):238
16. Antifragile-Wikipedia (2015). https://en.wikipedia.org/wiki/Antifragile. Accessed 02 October 2015
17. De Florio V (2015) On resilient behaviors in computational systems and environments. J Reliab Intell Environ 1–14
18. De Florio V (2014) Antifragility = elasticity + resilience + machine learning models and algorithms for open system fidelity. Proc Comput Sci 32:834–841
19. Jones KH (2014) Engineering antifragile systems: a change in design philosophy. Proc Comput Sci 32:870–875
20. Qian L, Luo Z, Du Y, Guo L (2009) Cloud computing: an overview. In: Cloud computing. Springer, Berlin, pp 626–631
21. Fernando N, Loke SW, Rahayu W (2013) Mobile cloud computing: a survey. Future Gener Comput Syst 29(1):84–106
22. Rimal BP, Choi E, Lumb I (2009) A taxonomy and survey of cloud computing systems. In: Fifth international joint conference on INC, IMS and IDC, 2009. NCM'09. IEEE, New York, pp 44–51
23. Kim S (2014) QoS-aware data forwarding architecture for multimedia streaming services in hybrid peer-to-peer networks. In: Peer-to-peer networking and applications, pp 1–10
24. Zhu W, Luo C, Wang J, Li S (2011) Multimedia cloud computing. Signal Process Mag IEEE 28(3):59–69
25. Nan X, He Y, Guan L (2011) Optimal resource allocation for multimedia cloud based on queuing model. In: 2011 IEEE 13th international workshop on multimedia signal processing (MMSP). IEEE, New York, pp 1–6
26. Nan X, He Y, Guan L (2012) Optimal resource allocation for multimedia cloud in priority service scheme. In: 2012 IEEE international symposium on circuits and systems (ISCAS). IEEE, New York, pp 1111–1114
27. Sembiring K, Beyer A (2013) Dynamic resource allocation for cloud-based media processing. In: Proceeding of the 23rd ACM workshop on network and operating systems support for digital audio and video. ACM, New York, pp 49–54
28. Li Y, Zhuo L, Shen H (2013) An efficient resource allocation method for multimedia cloud computing. In: Intelligence science and big data engineering. Springer, Berlin, pp 246–254
29. Hong B, Tang R, Zhai Y, Feng Y (2013) A resources allocation algorithm based on media task QOS in cloud computing. In: 2013 4th IEEE international conference on software engineering and service science (ICSESS). IEEE, New York, pp 841–844
30. Hassan MM (2014) Cost-effective resource provisioning for multimedia cloud-based e-health systems. Multimed Tools Appl 1–17
31. Alasaad A, Shafiee K, Behairy HM, Leung V (2015) Innovative schemes for resource allocation in the cloud for media streaming applications. IEEE Trans Parallel Distrib Syst 26(4):1021–1033
32. Hossain MS, Hassan MM, Qurishi MA, Alghamdi A (2012) Resource allocation for service composition in cloud-based video surveillance platform. In: 2012 IEEE international conference on multimedia and expo workshops (ICMEW). IEEE, New York, pp 408–412
33. De Cicco L, Mascolo S, Calamita D (2013) A resource allocation controller for cloud-based adaptive video streaming. In: 2013 IEEE international conference on communications workshops (ICC). IEEE, New York, pp 723–727
34. Ma YW, Chen JL, Chou CH, Lu SK (2014) A power saving mechanism for multimedia streaming services in cloud computing. Syst J IEEE 8(1):219–224
35. Song B, Hassan MM, Alamri A, Alelaiwi A, Tian Y, Pathan M, Almogren A (2014) A two-stage approach for task and resource management in multimedia cloud environment. Computing 1–27
36. Wen H, Hai-ying Z, Chuang L, Yang Y (2011) Effective load balancing for cloud-based multimedia system. In: 2011 international conference on electronic and mechanical engineering and information technology (EMEIT), vol 1. IEEE, New York, pp 165–168
37. Nan X, He Y, Guan L (2013) Optimization of workload scheduling for multimedia cloud computing. In 2013 IEEE international symposium on circuits and systems (ISCAS). IEEE, New York, pp 2872–2875
38. Lin CC, Chin HH, Deng DJ (2014) Dynamic multiservice load balancing in cloud-based multimedia system. Syst J IEEE 8(1):225–234
39. Kim M, Han SH, Jung JJ, Lee H, Choi O (2014) A robust cloud-based service architecture for multimedia streaming using Hadoop. In: Mobile, ubiquitous, and intelligent computing. Springer, Berlin, pp 365–370
40. Cho WT, Lai CF (2014) Adaptive multimedia cloud computing center applied on H. 264/SVC streaming. In: Cloud computing. Springer, New York, pp 14–26
41. Hsu WH, Lo CH (2014) QoS/QoE mapping and adjustment model in the cloud-based multimedia infrastructure. Syst J IEEE 8(1):247–255
42. Niu D, Xu H, Li B, Zhao S (2012) Quality-assured cloud bandwidth auto-scaling for video-on-demand applications. In: 2012 Proceedings IEEE INFOCOM. IEEE, New York, pp 460–468