J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

A Novel SDN-based System for Provisioning of Smart Hybrid Media Services☆

Myunghoon Jeon[1] Byoung-dai Lee[1*]

ABSTRACT

In recent years, technology is rapidly changing to support new service consumption and distribution models in multimedia service systems and hybrid delivery of media services is a key factor for enabling next generation multimedia services. This phenomenon can lead to rapidly increasing network traffic and ultimately has a direct and aggravating effect on the user's quality of service (QOS). To address the issue, we propose a novel system architecture to provide smart hybrid media services efficiently. The architecture is designed to apply the software-defined networking (SDN) method, detect changes in traffic, and combine the data, including user data, service features, and computation node status, to provide a service schedule that is suitable for the current state. To this end, the proposed architecture is based on 2-level scheduling, where Level-1 scheduling is responsible for the best network path and a computation node for processing the user request, whereas Level-2 scheduling deals with individual service requests that arrived at the computation node. This paper describes the overall concept of the architecture, as well as the functions of each component. In addition, this paper describes potential scenarios that demonstrate how this architecture could provide services more efficiently than current media-service architectures.

☞ keyword : Cloud Computing, Smart hybrid media services, SDN, Quality of Service

1. Introduction

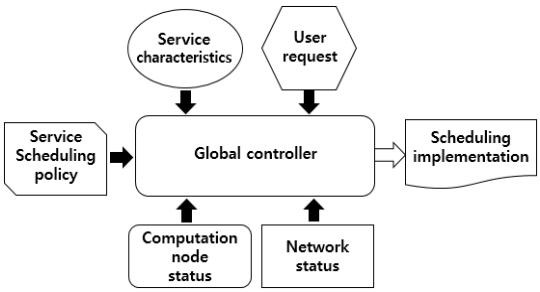
In recent years, due to increasing popularity of smart TVs with integrated Internet capability and widespread use of personal multimedia devices such as smartphones and tablets, the environment surrounding multimedia service systems are now moving into a new phase and hybrid media services are at the core of the changes. In keeping with such trends, numerous services have appeared that use the cloud-computing environment; its type and domain continue to expand. Early cloud-based services largely focused on storage; they had not expanded to other areas, e.g., big data analysis, games, and media processing. According to the International Data Corporation (IDC), the public cloud-service market will grow to 203.4 billion dollars by 2020, posting an annualized growth rate of 21.5% from 2015 to 2020 [1]. In particular, cloud-based media services are gaining attention as one of the areas expected to grow the most over the next five years.

The growth in the media service sector is being accelerated by the continuous improvements in media-related technology. A case in point is the coding technology for ultra-high definition (UHD)[2], and devices that support the technology. These devices have made it possible for users to enjoy high-quality media services, regardless of where or when they are available, raising demands for the services and facilitating their market growth.

However, with the growth, the data volume required to perform such services has also increased, resulting in a surge of network traffic [3]. This is a critical component that directly affects users' quality of service (QOS) and needs to be addressed. Traditional media services are based on an architecture that does not consider the network. This makes the traffic problem more important for a service provider; it needs to be solved to ensure service competitiveness.

This paper proposes a smart media-service architecture as a solution. Figure 1 describes the overall concept of the proposed architecture. The underlying concept is to analyze

[1] Department of Computer Science, Kyonggi University, Suwon, 16227, Korea.
* Corresponding author (blee@kgu.ac.kr)
[Received 16 October 2017, Reviewed 6 November 2017(R2 11 January 2018), Accepted 31 January 2018]
☆ A preliminary version of this paper was presented at APIC-IST 2017.
☆ This work was supported by Kyonggi University Research Grant 2016



(Figure 1) Comcept of the Proposed Architecture

a combination of data, e.g., the network and computation node status, service features, service access patterns, and user information, based on a service-scheduling policy defined from the perspective of a service provider. Thus, the architecture can determine an efficient service policy suitable for the current state, and ensure the user's QOS from the service provider's point of view.

This paper is composed as follows. In Section II, we will look into related work. In Section III, we will discuss potential issues that may arise from the traditional media service to provide grounds for this paper. In Section IV, we will describe the concept and components of the smart media-service architecture, as well as the functions of each component. In Section V, we will describe service scenarios to illustrate the performance of the proposed architecture. Section VI concludes the paper.

2. Related Work

As cloud-based media services are growing, studies designed to improve service efficiency are ongoing in a wide range of areas. Resource management is one such area [5][6][7][8]. Ref. [5] discussed a dynamic resource-provisioning method to ensure QOS when transcoding online video-sharing services. They designed an online algorithm to determine the volume of resources needed to ensure QOS and applied this algorithm to improve the resource-consumption efficiency. Ref. [6] examined resource utilization and task complexity when providing cloud-based transcoding services. They proposed a task-scheduling scheme based on the data to efficiently provide resource slot allocation. Ref. [7] proposed a dynamic scheduler for DASH (Dynamic Adaptive Streaming over HTTP) video transcoding designed for cloud environments. This study forecasted the transcoding time and proposed a model to apply priorities based on the forecast. Further, the resource functionality was classified and allocated to reduce the overall processing delay. Ref. [8] proposed the content placement strategy for user-generated contents based on media cloud. The first step of the strategy is to find the optimal placement for a single content and in the second step, the analytical results obtained in the first step is used to solve the cost optimization problem.

In addition to resource-management studies, one field of study focuses on the computation and seeks to improve the performance of the service processing itself [9][10][11]. Ref. [9]developed a distributed transcoding scheme that considers how the segments' dependency can improve the bit rate and time efficiency. Ref. [10] proposed a two-level scheduling scheme to improve the efficiency in distributed encoding. They proposed an algorithm that identified a video content's complexity and dependency, created segments based on the data, and defined the encoding order. Further, they demonstrated their proposed algorithm's improved efficiency in terms of the bit rate and encoding time. Ref. [11] proposed a motion estimation algorithm for GPU-based cloud encoding. In order to improve the most time-consuming part of the H.264 encoding job, the proposed algorithm applied CUDA-based thread optimization. In addition, a parallelized

(Table 1) Summary of related work

References	Proposition
Ref.[5]	A dynamic resource management for online video-sharing services
Ref.[6]	Resource management for cloud-based media transcoding service
Ref.[7]	A priority-based dynamic scheduler for DASH-based video transcoding
Ref.[8]	Content placement algorithm based on two-step strategy
Ref.[9]	A cloud-based distributed transcoding algorithm in consideration of segments' dependency
Ref.10]	MapReduce-based two-level scheduling algorithm for UHD video encoding
Ref.[11]	GPU-based parallel video encoding in cloud environment

motion estimation algorithm in consideration of motion tendency was also proposed. Table 1 summarized the abovementioned studies.

However, since these studies only focused on media services' resources and computation, they did not consider the network, which directly affects the user's QOS. Therefore, it may be difficult to effectively control the variety of data traffic arising from recent media services. Our paper proposes a smart media-service architecture that considers both the computation and the network to solve the aforementioned problems.

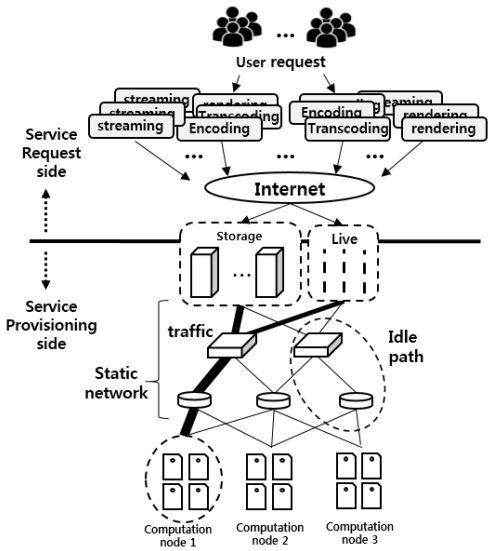
3. Problem Descripotion

Media services basically identify the internal status of the computation nodes to consider a user's requested service, select the right node, and process the service. In this process, the input data and results are transmitted via network. Since traditional media services do not consider the network, services are provided based on the static network structure. This may lead to problems, as the type and volume of media services grow and data-access patterns diversify.

Figure 2 shows examples of such issues. It displays a scenario where the service requests to be processed in computation node 1 surge, and the network traffic increases. In this case, the network path to computation node 1 deteriorates in transmission; however, since the path was fixed in the initial setting, the data is transmitted through the same path for subsequent requests. If the data can be transmitted to computation node 1, using another path with less traffic, the transmission efficiency will improve, along with the entire service performance.

In the hybrid multimedia service environment, a single media content can be composed of multiple videos or data contents, and is provided to users through logical display such as TV or other various smart devices. At this point, multiple networks may be used depending on the physical location of the server that actually transmits the contents. Here, an integrated control technology is required for individual networks.

Unfortunately, traditional media services only focus on the computation side with no regard to the network, making it difficult to assure users' QOS when providing diversified and high quality media services. We believe our proposed architecture will solve these problems.



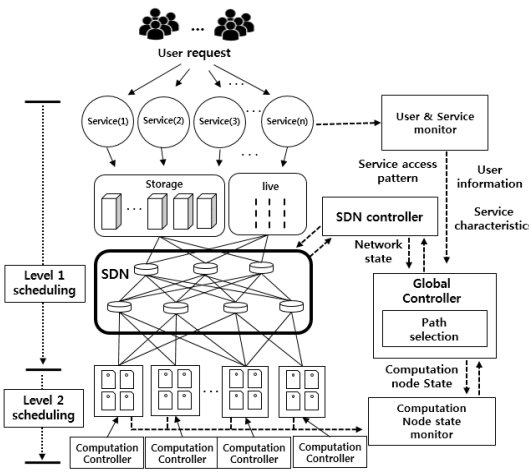
(Figure 2) Media-Service Architecture that Does Not Consider the Network Side

4. Proposed Architecture

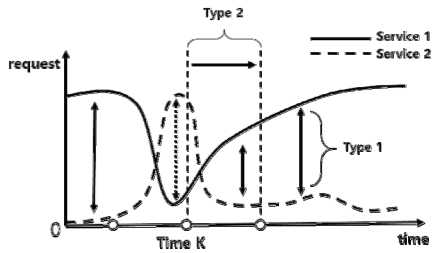
The underlying concept of the proposed architecture is to define the factors that can impact the QOS when providing media services, and effectively control them. Figure 3 illustrates the overall concept of the proposed architecture. We classified the impact factors into a data-transmission side and a data-processing side, and defined their scheduling as Level 1 and Level 2, respectively. Level 1 schedules the data transmission; it seeks to determine the most optimal node and create a network path to the node. Level 2 schedules the essential service processing and seeks to improve performance by applying a scheduling algorithm within the computation node.

4.1 Level 1 Scheduling

Level 1 scheduling considers the transmission efficiency and the network state and provides services that suit the current state. The following components are used during Level 1 scheduling.



(Figure 3) Proposed Smart Hybrid Media Service Architecture



(Figure 4) Examples of changes of access patterns

4.1.1 User and Service Monitor

The user and service monitor watches the users and services. Service monitoring can be classified into two sides: the service characteristics, including the service type and the data volume, and the service access pattern. The monitor determines whether the access pattern of a requested service changes as time passes. All of the monitored data is sent to the global controller. The user and service monitor utilizes two time-wise parameters, Type 1 and Type 2, to realize the access pattern. The Type 1 parameter denotes the number of client requests at a time point, whereas the Type 2 parameter represents the pattern of client requests (e.g., increasing or decreasing pattern). Figure 4 shows examples of changes of access pattern of two services and how these two parameters are used to determine the characteristics of the access patterns of two services. According to the Type 1 parameters, it can be seen that service 1 is being requested more than service 2 in most cases. However, at the time point K, client requests for service 2 grows rapidly, whereas the number of client requests for service 1 decrease. However, according to Type 2, the monitor is able to capture that the sudden increase in the client request was transient and, as a result, it does not respond to the change of the access pattern of the service 1.

4.1.2 Computation-node state monitor

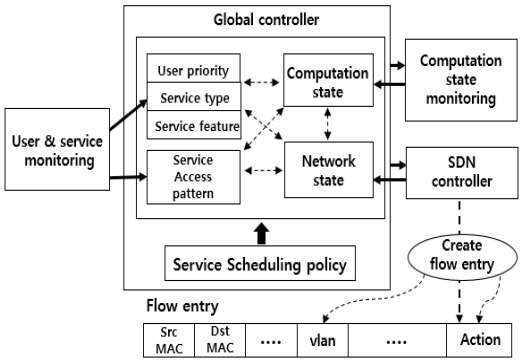
The computation monitor serves to identify the status of an available computation node when a service is requested. It investigates the volume and type of ongoing work, identifies computation nodes that suit the current request, and sends a list of available nodes to the global controller. The computation monitor receives periodically the status information of computation nodes from corresponding computation controllers. Some of important information delivered to the computation monitor include the current load status of individual computation node, the complexity of the currently executing service based on the input and the output of the service. Using the information, the computation monitor determines the best computation node for executing the client request and sends the target node to the global controller.

4.1.3 SDN controller

The SDN controller identifies the network status and creates a flow entry for a requested service. In identifying the network status, each switch sends counter data for the flow entry to identify the traffic level; the data is sent to the global controller. The SDN controller receives the results calculated by the global controller and creates a flow entry for the particular service.

4.1.4 Global controller

The global controller comprehensively looks through the data collected from the above three components, and determines the best network path for the current state. Figure 5 shows the entire flow of the global controller. As seen in the figure, when a user submits a request, information



(Figure 5) The Workflow of The Global Controller

regarding the user and the service is first sent to the global controller. This information can be grouped into user priority, service type, service feature, and service access pattern.

Different weights can be given to users and service-related data to reflect the service provider's perspective. The global controller can then apply a higher priority to a streaming service, among the other service types, on a real-time basis, or may apply a higher priority to a task that demands more calculation complexity, e.g., UHD encoding. Thus, the global controller can define priorities for the users and service-related data based on the service provider's perspective.

The global controller requests the states of the computation nodes that are fit for the requested services, based on the priorities, and receives a list accordingly. Then, it examines the network state for each node and determines the most optimal node. In doing so, the global controller runs a network-path check that can access the decided node and determine a path. This course is run based on the service-scheduling policy of the service provider.

For example, if a computation node has been selected and two paths access the node from the start point, the SDN controller may choose whether to apply the priorities from the service provider's perspective or to ignore them. The global controller sends the network-path data, determined based on the monitored data, to create a flow entry. In sum, the global controller takes the service provider's point of view and looks into the user information, service information, and network and computation statuses, to determine a network path that can efficiently provide the service under the current circumstances.

4.2 Level 2 Scheduling

Once a computation node is determined through Level 1 scheduling, it is given the data and processes the service. This computation node will apply an algorithm to more effectively process the service; this course constitutes Level 2 scheduling. This means, the efficiency of the Level 2 scheduling may differ based on the type and performance of the algorithm applied.

For example, when processing a video-encoding service, the video segmentation and encoding order may be applied differently according to the algorithm, and that may change the performance of the encoding time, bit rate, or peak signal-to-noise ratio (PSNR). Examples of Level 2 scheduling can be seen in [9][10][12] The algorithm types can differ based on the user's requirements (time priority, bit-rate priority, time and bit rate to be considered equally, etc.).

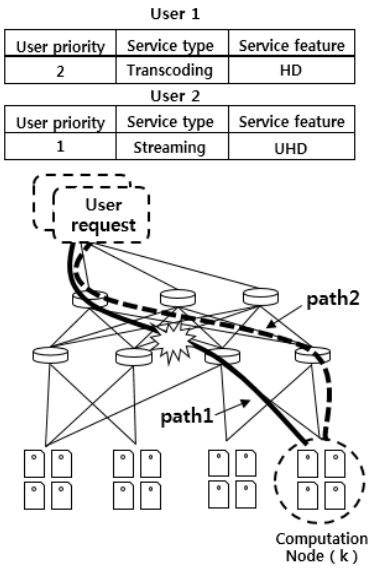
5. Expected Scenarios

Media services basically identify the internal status of the computation nodes to

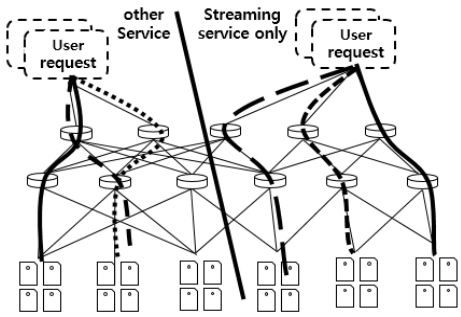
In this section, we describe a possible service scenario to illustrate how the proposed media-service architecture works. A major scenario for this architecture is responding to rapidly rising network traffic.

Figure 6 shows a situation where a high volume of traffic comes to path 1, among the network paths connected to computation node (k). If services requested from two users are to be processed in the computation node, the global controller may select different paths for the services, based on the service provider's perspective, as described earlier. In Figure 5, if the service provider focuses on the user priority, the global controller opens path 2 to user 1 and gives path 1 to user 2 to ensure the QOS, focusing on the user's priority.

Figure 7 illustrates a service scenario that can appear on the other side of the scenario in Figures 6 and 7 shows what happens when a rapid change is made in the access patterns of a service requested from a user. The service provider may give a higher priority to streaming services and focus on the QOS for the service as requests for streaming services



(Figure 6) Path Allocation based on User Priority as Traffic Increases

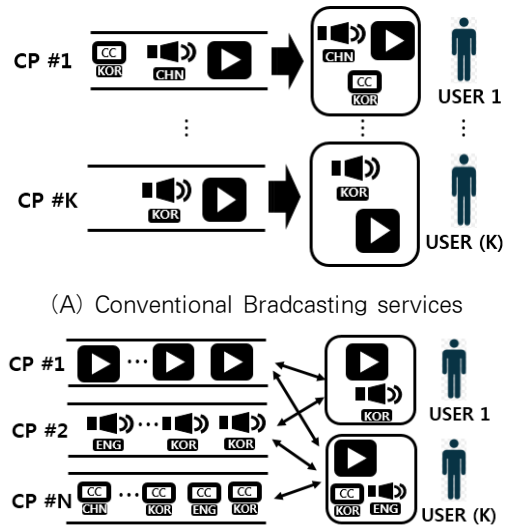


(Figure 7) Path Allocation based on Changing Service Access Patterns

increase as time passes.

Streaming services, by nature, should be provided with a seamless and stable network path. However, if the paths are offered uniformly to all services with no regard to the network, a stable path may not be available. However, in the proposed architecture, as shown in Figure 7, any changes in the service access pattern are recognized from the service provider's perspective, and a relatively stable network path can be guaranteed, if the paths can be separated for subsequent services by the global controller.

Hybrid broadcasting service is drawing much attention as

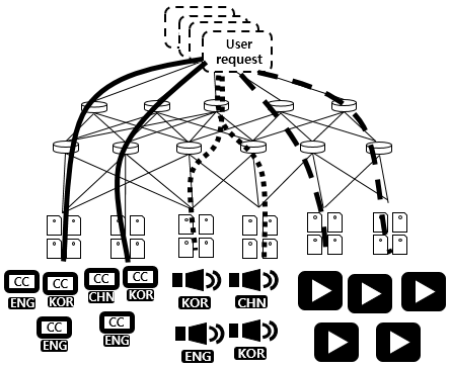


(A) Conventional Bradcasting services

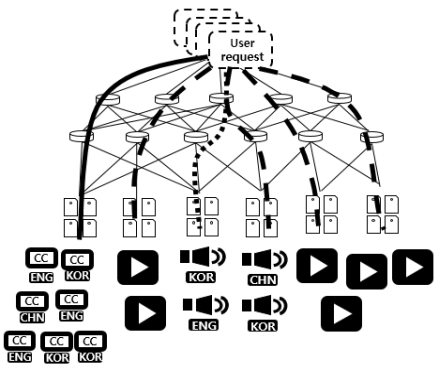
(B) Hybrid broadcasting services

(Figure 8) Media transmission environment of conventional and hybrid broadcasting services

a next generation broadcasting service that can provide not only existing unidirectional broadcasting service but also multimedia service [12]. A representative feature of the hybrid broadcasting service is transmission of multimedia contents through a bidirectional IP network. Figure 8 shows the differences in the transmission environment between the hybrid broadcasting service and the traditional broadcasting service. In the case of the transmission environment of the conventional broadcasting service, the user selects and receives the unidirectional contents from the contents provider. However, the users receiving the hybrid broadcasting can determine the image quality of the content they want from multiple providers, and can select the language of the subtitles and sound. It is also possible to chat over the Internet, search for and play related videos, and share them with other people close to them. The transmission environment of such a hybrid broadcasting service requires a network capable of transmitting multimedia data efficiently. However, the current Internet architecture, which is vertical and closed, has a technical limitation that it is difficult to flexibly cope with the maintenance and expansion of the network, which is directly connected with the quality



(A) Initial path allocation



(B) After chaning network path in cosideration of relationship among media components

(Figure 9) Path Allocation based on relationship among media components

degradation problem of the hybrid broadcasting service.

The problem described above can be solved by applying the SDN-based media service architecture proposed in this study. Figure 9 shows an example of a solution to this problem and it illustrates a process of preventing performance degradation of the overall services by considering the characteristics of contents for the requested service in the transmission environment of the hybrid broadcasting service and efficiently allocating the network based on the characteristics of the contents. Figure 9 (A) shows the initial state before applying the network change. The network allocation information at this time shows that the multimedia data of the service are relatively uniformly allocated to the network. However, if users request more videos or higher image quality, the existing allocated network path may be burdened, resulting in deterioration of the overall service quality. In this case, the network load can be lowered by reducing the existing network path for transmitting subtitles and extending the transmission path for videos as shown in Figure 9 (B) through the proposed architecture. What is notable in this process is to extend the network path for videos, but reduce the network for subtitle language. In other words, considering the characteristics of the multimedia provided in the current state, the total network load is decreased by reducing the network path for the subtitle language having a relatively low computational complexity and network requirement in changing the path for the video content. In conclusion, as shown in the above example, effective network management considering the characteristics of contents is possible in providing a hybrid broadcasting service by applying the SDN-based media service architecture proposed in this study because it can flexibly manage the allocation information of the entire network.

The three service scenarios described above are partial examples that show how the proposed media-service architecture can use the service provider's perspective to consider the user and service information, the network side, and the computation side, and thereby provide efficient services.

6. Conclusion

This paper focused on the service provider's perspective to comprehensively look at the user and service information, and network and computation elements, and proposed a smart media-service architecture to improve the service efficiency. To effectively describe the proposed architecture, we pointed out problems with the traditional media-service architecture. In addition, we examined the components of the proposed architecture and their roles and functions. Finally, we described some possible scenarios to highlight our architecture's potential efficiency.

We suggest that future studies implement and test the proposed architecture to verify its QOS efficiency. In addition, numerous experiments should be conducted based on clear definitions of algorithms for each component and from different service-provider perspectives.

Reference

[1] IDC, “Worldwide Semiannual Public Cloud Services Spending Guide,” Feb 2017. https://www.idc.com/getdoc.jsp?containerId=prUS42321417

[2] G.J. Sullivan, J. Ohm, W.J. Han and T. Wiegand, “Overview of the High Efficiency Video Coding (HEVC) Standard,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, Sept. 2012, pp. 1649-1668. https://dx.doi.org/10.1109/TCSVT.2012.2221191

[3] Cisco Visual Networking Index, “Forecast and Methodology. 2015-2020 White Paper,” Technical Report. Cisco, 2015. https://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.html

[4] Open Networking Foundation, “Software-Defined Networking: The New Norm for Networks,” April 2012. http://www.bigswitch.com/sites/default/files/sdn_resources/onf-whitepaper.pdf

[5] G. gao, Y. Wen, and C. Westpahl, “Dynamic Resource Provisioning with QoS Guarantee for Video Transcoding in Online Video Sharing Service,” Proceedings of the ACM on Multimedia Conference, PP. 868-877, 2016. https://dx.doi.org/10.1145/2964284.2964296

[6] C.C. Huang, J.J. Chen, and Y.H. Tsai, ″A Dynamic and Complexity Aware Cloud Scheduling Algorithm for Video Transcoding,″ Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2016. https://dx.doi.org/10.1109/ICMEW.2016.7574743

[7] H. Ma, B. Seo, and R. Zimmermann “Dynamic scheduling on video transcoding for MPEG DASH in the cloud environment,” Proceedings of the ACM Multimedia Systems Conference, PP. 283-294, 2014. https://dx.doi.org/10.1145/2557642.2557656

[8] Y. Jin, Y. Wen, and K. Guan, “Toward cost -efficient content placement in media cloud: modeling and analysis,” IEEE Transactions on Multimedia, vol. 18, no.5, pp. 807-819, Mar. 2016. https://dx.doi.org/10.1109/TMM.2016.2537199

[9] M.R. Zakerinasab, and M. Wang, “ Dependency- Aware Distributed Video Transcoding in the Cloud,” Proceedings of the IEEE 40th Conference on Local Computer Networks (LCN), 2015. https://dx.doi.org/10.1109/LCN.2015.7366317

[10] M.H. Jeon, N.G. Kim, and B.D. Lee, ″MapReduce -Based Distributed Video Encoding Using Content-Aware Video Segmentation and Scheduling,″ IEEE Access, vol. 4, pp. 6802-6815, Oct. 2016.

[11] W. Jiang, P. Wang, M. Long, and H. Jin, “A Novel Parallelized Motion Estimation Algorithm for GPU Based Video Encoding,” IEEE International Symposium on A World of Wireless, Mobile and Multimedia Networks, PP. 1-8, 2016. https://dx.doi.org/10.1109/WoWMoM.2016.7523558

[12] K.M. Park, N.G. Kim, and B.D. Lee, “Performance Evaluation of the Emerging Media-Transport Technologies for the Next-Generation Digital Broadcasting Systems,” IEEE Access, vol. 5, PP. 17597-17606, Aug. 2017. https://dx.doi.org/10.1109/ACCESS.2017.2737557

◐ 저 자 소 개 ◑

Myunghoon Jeon

2011 B.S. Dep. of Computer Science, Kyonggi Univ., Korea
2014 M.S. Dept. of Computer Science, Kyonggi Univ., Korea
2014~present: Ph.D Candidate, Dept. of Computer Science, Kyonggi Univ., Korea
Research Interests: distributed video coding, video compression, Software Defined Networking, resource scheduling and cloud computing.
E-mail : jmh@kgu.ac.kr

Byoung-Dai Lee

2003 Ph.D. Dept. of Computer Science and Engineering, University of Minnesota, U.S.A.
2003~2010 Samsung Electronics, Co., Ltd.
2010~present : Department of Computer science, Kyonggi Univ., Korea
Research Interests: clolud computing, mobile multimedia platform and mobile multimedia broadcasting
E-mail : blee@kgu.ac.kr