Quality of Service and Resource Provisioning Utilization in Media Delivery Datacenters

Minh N.H. Nguyen, Chuan Pham, S.M. Ahsan Kazmi and Choong Seon Hong
Department of Computer Science and Engineering, Kyung Hee University
E-mail: {minhnhn, pchuan, ahsankazmi, cshong}@khu.ac.kr

Abstract: Media delivery datacenters provides large-scale media content can get benefit from resource virtualization of cloud infrastructure. In the media service, on demand content popularity yields to extreme imbalance experience during peak hours and social events. Therefore, media datacenters can efficiently improve system utilization by using virtual machine to allocate dynamically resource for content access. To address this problem, we propose optimization model for maximizing global resource provisioning utilization with three resource dimensions: CPU, bandwidth and storage as well as Quality of Service (QoS) reasoning. In term of QoS provision, priority of content uses calculated weight factor from the number of content requests.

1. Introduction

Consecutive online video's growth belong with large-scale Internet viewers from all over the world. According to Akamai report [1], around 87% of Internet U.S. users watch video online, consuming nearly 40 billion content videos and 23 hours of video per viewer each month. Especially, during social events, online media content such as live streaming, on demand videos, songs, sharing photos get extraordinary user access. World Cup 2014 was recorded as the largest live sporting event ever delivered in the history of Akamai, the leading content delivery network provider [2]. Fig.1 illustrates a huge traffic rate divergence from peak to ordinary hours on Akamai network during the tournament.

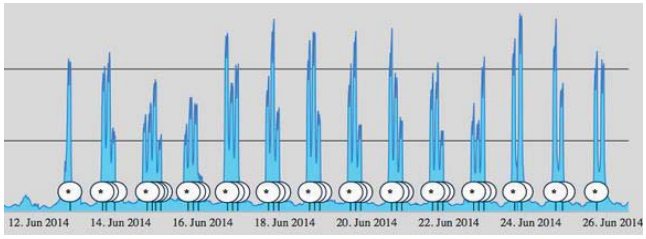

Figure 1: Akamai's traffic levels during World Cup 2014 [2]

Consequently, leverage virtualized resource allocation which could provide an ideal shared pool environment for media delivery service. From [6], author recommends virtual machine (VM) usage with scalability of computation, storage and bandwidth can provide access service for each on demand content. For example, different quality level of videos in adaptive bit rate live streaming require different transcoding computation, bandwidth and storage cost. Moreover, live VM migration feature provides a huge benefit from achieving better resource allocation utilization. When contents get peak access, datacenter can move content VMs to get available space, enabling greater user access. After peak hours, in order to save energy and operation cost, VMs can be gathered into lower number of servers. Therefore, resource virtualization in cloud computing can efficiently and costly handle on demand media content.

2. Preliminaries

Research in [3] uses constraint on the deadline of requests, while dynamic resource provisioning with virtual appliances queueing model [4] uses service level agreement (SLA) and one simple resource capacity parameter. In [5], VM placement strategy considers servers resource capacity is divided by slots of CPU/memory and focuses on communication cost. Different from previous research, maximizing utilization formulation across multiple resource dimensions constraints was introduces in [6] but hasn't regarded QoS.

The Quality of Service in media delivery service is a significant factor from customer satisfaction. Especially in peak hours, user requests might be denied or get long delay time. In that case, QoS can be defined as user access priority by differentiated service classes or weigh content by the number of requests. User access pattern or prediction can be useful for getting better priority of contents.

This paper focus on optimize global utilization with multiple resource dimensions and QoS concern, using weight factor for each content. In addition, by using mathematic analysis, we propose reference lower bounds for the number of servers in homogenous datacenters during peak hours or ordinary usage.

3. The problem formulation

Resource allocation problem was inspired from Generalized Assignment Problem (GAP) and related research on video streaming Datacenter [6]. In system model, we define set of contents $\mathcal{K}$ and set of servers $\mathcal{N}$. The objective function of optimization problem shows global resource utilization with storage, bandwidth, CPU usage ratios. This function was constructed from fairness between three resources and weighted sum of contents [7].

Maximize $\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\left(\frac{s_k}{S_i}+\frac{b_k R_k}{B_i}+\frac{c_k R_k}{C_i}\right)x_i^k\omega_k$

Subject to:

$$\sum_{k\in\mathcal{K}} x_i^k s_k \le S_i \qquad \forall i\in\mathcal{N} \qquad (1)$$

$$\sum_{k\in\mathcal{K}} x_i^k R_k b_k \le B_i \qquad \forall i\in\mathcal{N} \qquad (2)$$

$$\sum_{k\in\mathcal{K}} x_i^k R_k c_k \le C_i \qquad \forall i\in\mathcal{N} \qquad (3)$$

$$\sum_{i\in\mathcal{N}} x_i^k = 1 \qquad \forall k\in\mathcal{K} \qquad (4)$$

$$\omega_k\hat{R}_k \le \sum_{i\in\mathcal{N}} x_i^k R_k \le \hat{R}_k \qquad \forall k\in\mathcal{K} \qquad (5)$$

Binary variable $x_i^k = \{0,1\}$ indicates whether content *k* is stored on server *i*. Each request of content k needs s_k storage, b_k bandwidth and c_k CPU resource unit. Constraint (4) shows content *k* will be stored in one of servers in datacenter. Variable R_k specifies the number of requests to content *k* can be served and $\hat{R}_k$ is the required number of requests to content *k*.

The capacity of each server *i* was denoted by S_i, B_i, C_i and resrouce limiations are implied in contraint (1), (2), (3). Due to storage of video content will not be affected by the number of user requests within constraint (1). In fact, datacenter should provide enough storage for store all of contents (video or sharing user data). Whereas, the more user access the more CPU and bandwidth are consumed. In general, storage resource in constraint (1) and objective function can be smilarly considered as on demand CPU and bandwidth.

Additionally, in order to perform the priority of the content *k*, weight ω_k is attached into objective function and lower bound in constraint (5) as QoS factor. In this paper, we assign weight values from the number of requests for contents at one time slot. Moreover, weight factor can be produced base on pattern of user access history or prediction.

$$\omega_k := \frac{\hat{R}_k}{\sum_{k\in\mathcal{K}}\hat{R}_k}$$

4. Reference lower bound of system resources:

From constraint (5) we get lower bound of storage, bandwidth, CPU resource using mathematic analysis.

Bandwidth constraint of each content *k*:

$$\omega_k\hat{R}_k b_k \le \sum_{i\in\mathcal{N}} x_i^k R_k b_k \le \hat{R}_k b_k \qquad \forall k\in\mathcal{K}$$

Sum of all contents:

$$\sum_{k\in\mathcal{K}}\omega_k\hat{R}_k b_k \le \sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{N}} x_i^k R_k b_k \le \sum_{k\in\mathcal{K}}\hat{R}_k b_k$$

Combine with condition (2):

$$\sum_{k\in\mathcal{K}}\omega_k\hat{R}_k b_k \le \sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{N}} x_i^k R_k b_k \le \sum_{i\in\mathcal{N}} B_i$$

A homogeneous datacenter includes $|\mathcal{N}|$ servers with same resource capacity $\bar{S}_i, \bar{B}_i, \bar{C}_i$.

Therefore, $\sum_{k\in\mathcal{K}}\omega_k\hat{R}_k b_k \le |\mathcal{N}|\bar{B}_i \Leftrightarrow \frac{\sum_{k\in\mathcal{K}}\omega_k\hat{R}_k b_k}{\bar{B}_i} \le |\mathcal{N}|$

Similarly, for CPU constraint:

$$\frac{\sum_{k\in\mathcal{K}}\omega_k\hat{R}_k c_k}{\bar{C}_i} \le |\mathcal{N}|$$

Storage constraint is derived from (1) and (4):

$$\sum_{k\in\mathcal{K}} s_k \le |\mathcal{N}|\bar{S}_i \Leftrightarrow \frac{\sum_{k\in\mathcal{K}} s_k}{\bar{S}_i} \le |\mathcal{N}|$$

Due to the limited resource in peak hours, system manager can use reference lower bound for number of servers to avoid high QoS violation rate:

$$|\mathcal{N}| \ge \left\lceil max\left\{\frac{\sum_{k\in\mathcal{K}} s_k}{\bar{S}_i}, \frac{\sum_{k\in\mathcal{K}}\omega_k\hat{R}_k b_k}{\bar{B}_i}, \frac{\sum_{k\in\mathcal{K}}\omega_k\hat{R}_k c_k}{\bar{C}_i}\right\}\right\rceil$$

However, after peak hours the some idle servers can be turned off for saving energy and resource operation cost. System manager can reduce the number of servers to threshold that datacenter still provides enough resource for user requests:

$$|\mathcal{N}| \ge \left\lceil max\left\{\frac{\sum_{k\in\mathcal{K}} s_k}{\bar{S}_i}, \frac{\sum_{k\in\mathcal{K}}\hat{R}_k b_k}{\bar{B}_i}, \frac{\sum_{k\in\mathcal{K}}\hat{R}_k c_k}{\bar{C}_i}\right\}\right\rceil$$

5. Numerical results

For solving mixed integer nonlinear programs (MINLP) optimization problem [8], we are using Julia JuMP package in JuliaOpt [9] with CbcSolver. Firstly, we provide resource allocation in two homogenous servers for three contents: A, B and C. Resource parameters are shown in Table 1.

Resources	Storage (GB)	Bandwidth (MBps)	CPU (MIPS)
Server	4	10	8
Content A	1	2	2
Content B	2	2	1
Content C	1	1	1

Table 1: Server capability and Content service requirement

Initially, each content A, B, C receives one request and the optimize resource allocation in (Fig.2). Server 1 will contain content A and C VMs while Server 2 will serve for content B VM. Consequently, due to small number of requests, system utility is very low, around 0.7.

When content A becomes popular with five requests, content B receives three and content C still has one. Content B was moved to Server 2 to save space for content A resource extension. As a result, Server 1 reaches maximum bandwidth resource on Server 1 and system utility was achieved 1.61 (Fig.3). If the number of requests for content A keeps increasing, datacenter will get

overload and requests cannot be served. In this scenario, content B still has a chance to get one more request and content C can gets three more.

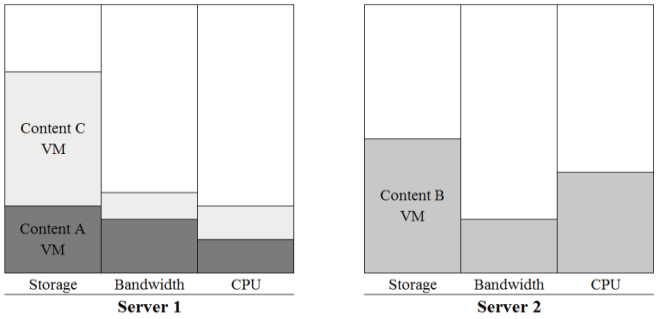

Figure 2: Initial VM allocation for each content

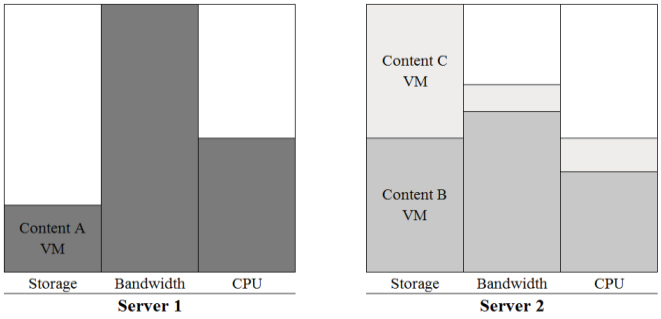

Figure 3: VM allocation after increasing number of requests

Additionally, we observe system utility from objective function between fairness allocation strategy for each content and using attached weight priority strategy. Fairness allocation strategy is configured with equal weight of each content by $1/|K|$. Two servers also have the resource capacity in Table 1 and provide VMs for two and three contents with the same resource requirements (2GB storage, 2MBps bandwidth and 1MIPS CPU). After keep increasing number of requests on one content from one to seven, global utility increasing because of more requests are served (Fig.4). In two content VMs scenario, from the 7 requests datacenter will not provide resource using priority strategy and 9 requests for fairness strategy. In three content VMs scenario, the limitation of priority strategy is 8. This limitation could refer back to the lower bound of QoS constraint violation. Therefore, overload system status is sooner reached using priority strategy.

6. Conclusion

In order for global utilization improvement in media delivery datacenters with variety of resource dimensions such as storage, bandwidth, CPU that needs QoS analysis for better service provision especially in peak hours. Attached weight factor are used to set priority for each content and quickly notify QoS violation by lower denied access rate on popular contents.

In practical, moving content VM cost depends on live migration technique that might affects the global utilization and needs to be evaluated in the future work. The better media delivery service could be provided by investigating and multiplexing more QoS criteria from user's point of view.

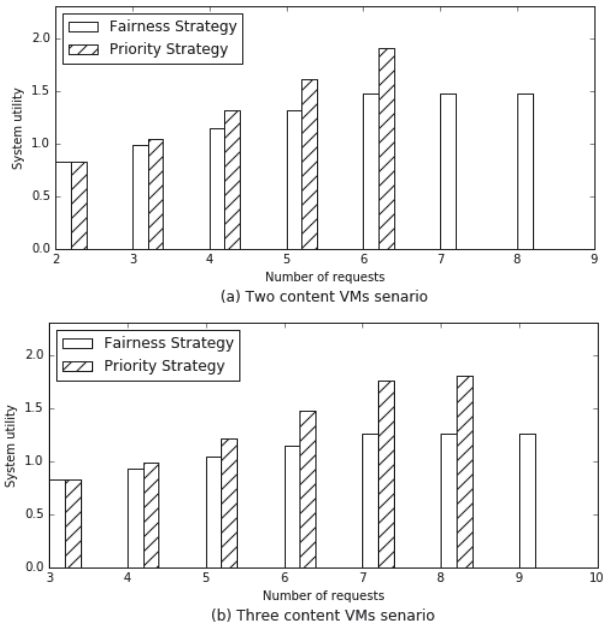

Figure 4: System utilization of fairness and priority strategy

7. Acknowledgment

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (B0190-15-2017, Resilient/Fault-Tolerant Autonomic Networking Based on Physicality, Relationship and Service Semantic of IoT Devices).

This research was one of KOREN projects supported by National Information Society Agency (15-951-00-001). *Dr. CS Hong corresponding author.

References:

[1] Akamai Tech. "Maximizing Audience Engagement: How online video performance impacts viewer behavior." White paper (2015)
[2] Akamai Tech. "The Media Delivery Cookbook" (2015)
[3] Aggarwal, Vaneet, et al. "Exploiting virtualization for delivering cloud-based IPTV services." Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on. IEEE, 2011.
[4] Wang, Xiao Ying, et al. "Appliance-based autonomic provisioning framework for virtualized outsourcing data center." Autonomic Computing, 2007. ICAC'07. Fourth International Conference on. IEEE, 2007.
[5] Meng, Xiaoqiao, Vasileios Pappas, and Li Zhang. "Improving the scalability of data center networks with traffic-aware virtual machine placement." INFOCOM, 2010 Proceedings IEEE. IEEE, 2010.
[6] Feng, Yuan, Baochun Li, and Bo Li. "Bargaining towards maximized resource utilization in video streaming datacenters." INFOCOM, 2012 Proceedings IEEE. IEEE, 2012.
[7] Chen Lee and Dan Siewiorek "An Approach for Quality of Service Management." TechReport (1998).
[8] Bussieck, Michael R., and Stefan Vigerske. "MINLP solver software." Wiley Encyclopedia of Operations Research and Management Science (2010).
[9] Lubin, Miles, and Iain Dunning. "Computing in operations research using Julia." INFORMS Journal on Computing 27.2 (2015): 238-248.