ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 1, January 2019

A Survey on an Effective Video Provisioning Scheme for Content Delivery Network

Suraj B. Patil, Prof. Santosh T. Waghmode

M. E Student, Department of Computer Engineering, Imperial College of Engineering and Research, Pune, Savitribai Phule Pune University, Pune, India

Department of Computer Engineering, Imperial College of Engineering and Research, Pune, Savitribai Phule Pune University, Pune, India

ABSTRACT: Content delivery networks (CDNs) have been widely implemented to provide innovative cloud services. Such networks encourage resource pooling by granting virtual machines or physical servers to be dynamically activated and deactivated as stated in current user requirement. This paper examines online video replication and placement problems in CDNs. An effective video provisioning scheme must concurrently (i) Utilize system resources to diminish total energy consumption and (ii) Decrease replication overhead. We propose a scheme called adaptive data placement (ADP) that can dynamically place and reorganize video replicas among cache servers on subscribers' arrival and departure. Both the analyses and simulation results display that ADP can reduce the number of activated cache servers with limited replication overhead. Additionally, ADP's performance is approximate to the optimal solution.

KEYWORDS: Content Delivery Networks, Resource Management, Video Streaming, Replication, Video Provisioning.

I. INTRODUCTION

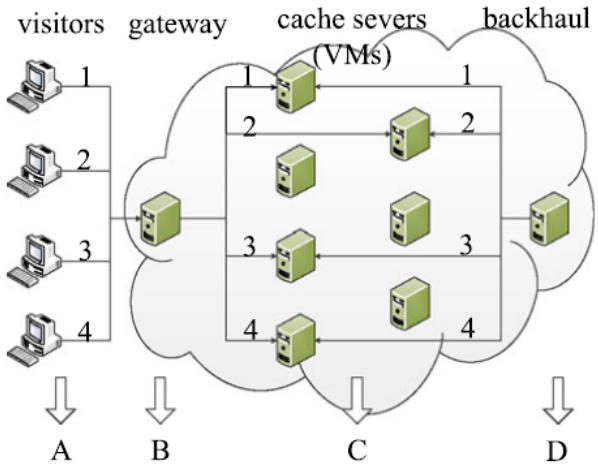

Fig1 Illustration of a local CDN

In recent years, content delivery networks (CDNs) have been widely implemented to provide scalable cloud services. Fig. 1 shows a typical local CDN whose servers are located in the same place. The request of each visitor (A) is first processed and identified by a gateway server (B) and then directed to a cache server (CS) (C) in the server farm.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 1, January 2019

CSs are typically virtual machines and can dynamically provide various services by executing different contents loaded from a backhaul database (D). This cloud-based architecture can provide a high degree of scalability and flexibility for service provisioning because it adaptively utilizes storage space, computing power, and network bandwidth by activating different numbers of CSs. As mentioned in Refs. The use of CSs is critical because the average loading of a single CS is typically substantially lower than its maximum capability. Therefore, minimizing the number of activated CSs is correlated to, if not equal to, minimizing the total energy consumption because of two reasons. First, supporting man activated virtual/physical machine requires considerable power compared with the dynamic workload of visitors. Second, the system resources (e.g., network bandwidth and CPU time) and power consumption required by each visitor are almost identical. Many related studies, such as focused on analyzing or reducing the number of activated CSs in a CDN.

CDNs are effective platforms for providing various types of services. Among them, on-demand video provisioning is a popular application, allowing numerous users to arbitrarily request videos from a massive database. Video websites such as YouTube, Vimeo, and DailyMotion are examples. When a visitor arrives and requests a video clip, the system must assign a serving CS and copy a replica of the clip from the backhaul database if the CS does not cache the clip for other visitors. Because each CS has limited capability, the total number of video clips it stores and the total outgoing bandwidth of subscribers it bears are limited by its space and bandwidth constraints. Fig. 2 illustrates an example of this two-dimensional resource allocation problem, where the length and width respectively express the space and bandwidth capacity of a CS, and u and p respectively illustrate users and their requested programs.

Here, Visitors 1, 2, and 3 require Video Programs A, B, and C respectively. When User 4 is later directed to this server, an additional unit of bandwidth is required because each video is independently transmitted played by each user. However, Visitor 4 does not occupy an additional unit of storage space because Program A has already been requested by Visitor 1.

Because of the many time-variant requirements of video clips, intelligently placing videos among CSs and determine their serving subscribers without violating capacity and bandwidth limits is challenging. Typical CDN management schemes in data centers fail to address this video provisioning problem for three reasons. First, rather than being separately required by a specific user/subscriber, a video clip may be simultaneously accessed by numerous online subscribers.

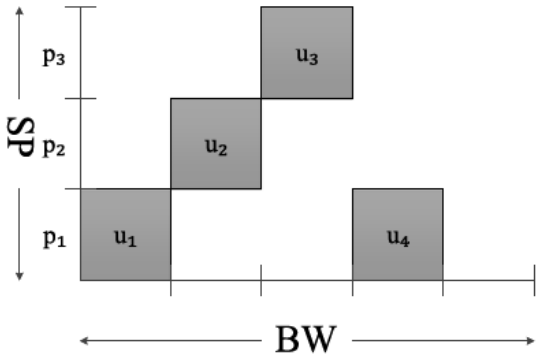

Fig2 Example of a CS in a CDN.

Second, the two resource requirements (i.e., bandwidth and storage size) exhibit different characteristics (i.e., when different visitors request the same clip, their storage requirements can be combined, whereas their bandwidth requirements are independent). Third, because video placements and server selections are conducted online as visitors arrive and depart; the migration/management of clip overhead should be limited to provide real-time responsiveness.

Therefore, a new resource management scheme that is specifically designed for provisioning online videos in CDNs is required.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 1, January 2019

II. RELATED WORK

iAware, a lightweight interference-aware VM live migration strategy. It analytically captures the essential relationships between VM performance interference and key factors that are practically accessible through realistic experiments of benchmark workloads on a Xen virtualized cluster platform. iAware jointly estimates and minimizes both migration and co-location interference among VMs, by designing a simple multi-resource demand-supply model. Major experiments and complementary large-scale simulations are conducted to validate the performance gain and runtime overhead of iAware in terms of I/O and network throughput, CPU consumption, and scalability, compared to the traditional interference-unaware VM migration approaches [1].

In this review the state-of-the-art research on managing the performance overhead of VMs, and summarize them under diverse scenarios of the IaaS cloud, ranging from the single-server virtualization, a single mega datacenter, to multiple geodistributed datacenters. Specifically, we unveil the causes of VM performance overhead by illustrating representative scenarios, discuss the performance modeling methods with a particular focus on their accuracy and cost, and compare the overhead mitigation techniques by identifying their effectiveness and implementation complexity [2].

Video conferencing, regardless of itsstringent delay constraints, must also be furnished as a cloud service, taking complete gain of theinter-datacenter network in the cloud. We design Airlift, a new protocol designed for the inter-datacenter network, tailored to the needs of a cloud-based video conferencing service. Airlift delivers packets in live video conferences to their respective destination datacenters, with the objective of maximizing the total throughput across all conferences, yet without violating end-to-end delay constraints. In order to simplify our protocol design in Airlift, we use intersession network coding and the concept of conceptual flows, such that the optimization problem that can be conveniently formulated as a linear program. Our real-world implementation of Airlift has been deployed over the Amazon EC2 cloud [3].

By optimizing the position of VMs on host machines, traffic patterns among VMs can be higher aligned with the communication distance between them, e.g. VMs with large mutual bandwidth usage are assigned to host machines in close proximity. We formulate the VM placement as an optimization hassle and prove its hardness. We design a two-tier approximate algorithm that effectively solves the VM placement problem for very large problem sizes. Given the significant difference in the traffic patterns seen in present data centers and the structural variations of the currently proposed data center architectures, we further behavior a comparative analysis on the impact of the traffic patterns and the network architectures on the capability overall performance advantage of traffic-aware VM placement [4].

Consequently, an application's environmental impact can vary accordingly depending on the geographical distribution of end-users, as electricity cost and carbon footprint per watt is location specific. In this work, we describe FORTE: Flow Optimization based framework for request-Routing and Traffic Engineering. FORTE dynamically controls the portion of user traffic directed to each datacenter in response to changes in both request workload and carbon footprint. It allows an operator to navigate the three-way tradeoff between access latency, carbon footprint, and electricity costs and to determine an optimal datacenter upgrade plan in response to increases in traffic load [5].

Previous works are limited on cutting down the power consumption of the datacenters to defuse such a concern. In this work, we show how the spatial and temporal variabilities of the electricity carbon footprint can be fully exploited to further green the cloud running on top of geographically distributed datacenters. We together consider the electricity cost, service level agreement (SLA) requirement, and emission reduction budget. To navigate such a three-way tradeoff, we take advantage of Lyapunov optimization techniques to design and analyze a carbon-aware control framework, which makes online decisions on geographical load balancing, capacity right-sizing, and server speed scaling [6].

With the presence of fluctuating workloads in datacenters, the lifetime and reliability of servers underneath dynamic power-aware consolidation could be adversely impacted by using repeated on-off thermal cycles, wear-and-tear and temperature upward push. In this paper, we suggest a Reliability-Aware server Consolidation strategy, named RACE, to deal with when and a way to perform energy-efficient server consolidation in a reliability-friendly and worthwhile way. The focus is on the characterization and analysis of this hassle as a multi-objective optimization, by means of developing a software model that unifies a couple of constraints on overall performance SLAs, reliability factors, and strength costs in a holistic manner.An improved grouping genetic algorithm is proposed to search the global optimal solution, which takes advantage of a collection of reliability-aware resource buffering, and virtual machines-to-servers re-mapping heuristics for generating good initial solutions and improving the convergence rate [7].

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 1, January 2019

An analytical scheme for characterizing and optimizing the power-performance tradeoff in Software-as-a-Service (SaaS) cloud platforms. Our objectives are two-fold: (1) We maximize the operating profit when serving heterogeneous SaaS applications with unpredictable user requests, and (2) we minimize the power consumption when processing user requests. To achieve these objectives, we take advantage of Lyapunov Optimization techniques to design and analysis an optimal control framework to make online decisions on request admission control, routing, and virtual machine (VMs) scheduling. In particular, our control framework can be flexibly extended to incorporate various design choices and practical requirements of a datacenter in the cloud, such as enforcing a certain power budget for improving the performance (dollar) per watt [8].

Introduce a dynamic capacity management policy, Auto Scale that greatly reduces the number of servers needed in data centers driven by unpredictable, time-varying load, while meeting response time SLAs. Auto Scale scales the data center capacity, adding or removing servers as needed. Auto Scale has two key features: (i) it autonomically maintains just the right amount of spare capacity to handle bursts in the request rate; and (ii) it is robust not just to modifications in the request rate of real-world traces, but also request size and server efficiency. We evaluate our dynamic capacity management approach via implementation on a 38-server multi-tier data center, serving a web site of the type seen in Face book or Amazon, with a key-value store workload. We demonstrate that Auto Scale vastly improves upon existing dynamic capacity management policies with respect to meeting SLAs and robustness [9].

As virtual machines dynamically enter and leave a cloud system, it becomes necessary to relocate virtual machines among servers. However, relocation of virtual machines introduces run-time overheads and consumes extra energy, thus a careful planning for relocation is necessary. We model the relocation problem as a modified bin packing problem and propose a new server consolidation algorithm that guarantees server consolidation with bounded relocation costs. We also conduct a detailed analysis on the complexity of the server consolidation problem, and give an upper bound on the cost of relocation. Finally, we handling simulations and compare our server consolidation algorithm with other relocation methods, like First Fit and Best Fit method. The experiment results suggest an interesting trade-off between server consolidation quality and relocation cost [10].

III. OPEN ISSUES

Content Delivery Networks (CDNs) are effective platforms for providing various types of services. Among them, on-demand video provisioning is a popular application, allowing numerous users to arbitrarily request videos from a massive database. Video websites such as YouTube, Vimeo, and DailyMotion are examples. When a visitor arrives and requests a video clip, the system must assign a serving CS and copy a replica of the clip from the backhaul database if the CS does not cache the clip for other visitors. Because each CS has limited capability, the total number of video clips it stores and the total outgoing bandwidth of subscribers it bears are limited by its space and bandwidth constraints.

Disadvantages are:

1. Because of the many time-variant requirements of video clips, intelligently placing videos among CSs and determine their serving subscribers without violating capacity and bandwidth limits is challenging. Typical CDN management schemes in data centers fail to address this video provisioning problem

IV. SYSTEM OVERVIEW

This paper introduces a new problem called resource-saving video placement (RSVP) and proposes a scheme called adaptive data placement (ADP). Through analysis and simulations, we demonstrate the two main advantages of ADP: (i) the worst case performance difference between ADP and the optimal solution can be guaranteed, and (ii) the replication overhead on each arrival or departure of a visitor is limited. Because ADP is based on common assumptions, it can be applied to various types of CDNs to improve their resource and power efficiency.

Advantages of Proposed System:

1. To achieve high resource utilization, our proposed scheme, ADP, follows three principles: (i) it maintains only one OPS server in a system to enable most CSs to achieve at least one aspect (i.e., bandwidth or space) of full utilization; (ii) it maintains the exclusiveness of video clips (i.e., allows at most one replica for each clip) among the

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 1, January 2019

OPS and SPF servers to improve space efficiency, which we demonstrate in the next section; and (iii) it conducts less physical replication to limit overhead.

Proposed System Architecture:

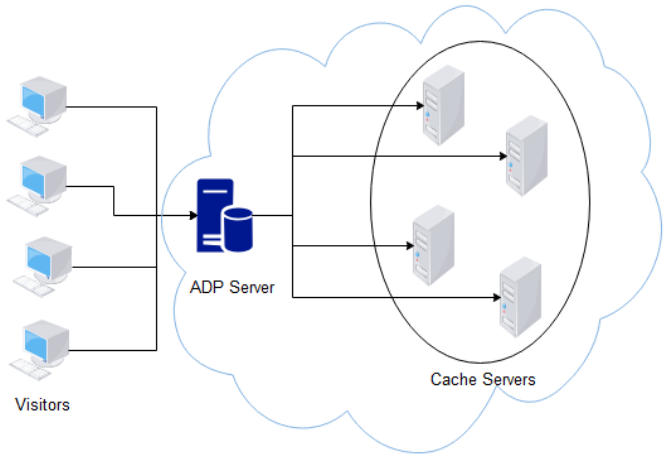

Fig. 3 Proposed Content Delivery Network Architecture

V. CONCLUSION

In this work, we examine an online video placement scheme for superior utilization and energy-saving in cloud delivery networks. We introduce a new problem that dynamically places incoming video subscribers to CSs to limit the number of active machines as well as the replication overhead. This problem considers both transmissions bandwidth and storage space constraints and is modeled. It can therefore be applied effectively to various types and scales of CDNs. By classifying servers into different types, our proposed ADP scheme places and reorganizes video subscriptions on their arrival and departure.

Through analysis, we demonstrate the effectiveness of ADP regarding performance and overhead. The worst-case overhead of ADP is limited, and the performance difference to the optimum is bounded. The outstanding performance of ADP is also evidenced by the simulations. The results show that ADP significantly outperforms the compared scheme under various conditions and maintains performance approximate to the optimal solution. In addition, the replication overhead of the system is also limited. To the best of our knowledge, ADP is the only scheme that addresses this placement problem and provides all the mentioned advantages.

REFERENCES

[1] F. Xu, F. Liu, L. Liu, H. Jin, B. Li, and B. Li, "iAware: Making livemigration of virtual machines interference-aware in the cloud,"IEEE Trans. Comput., vol. 63, no. 12, pp. 3012–3025, Nov. 2014.
[2] F. Xu, F. Liu, H. Jin, and A. V. Vasilakos, "Managing performanceoverhead of virtual machines in cloud computing: A survey, stateof the art, and future directions," in Proc. IEEE, 2014, vol. 102,no. 1, pp. 11–31.
[3] Y. Feng and B. Li, "Airlift: Video conferencing as a cloud serviceusing inter-datacenter networks," in Proc. IEEE Int. Conf. Netw.Protocols, 2012, pp. 1–11.
[4] X. Meng, V. Pappas, and L. Zhang, "Improving the scalability ofdata center networks with traffic-aware virtual machineplacement," in Proc. IEEE INFOCOM, 2010, pp. 1–9.
[5] P. X. Gao, A. R. Curtis, B. Wong, and S. Keshav, "It's not easybeing green," in Proc. ACM SIGCOMM, 2012, pp. 211–222.
[6] Z. Zhou, F. Liu, Y. Xu, R. Zou, H. Xu, J. C. S. Liu, and H. Jin,"Carbon-aware load balancing for geo-distributed cloud services,"in Proc. IEEE Modelling, Anal. Simul. Comput. TelecommunicationSyst., 2013, pp. 232–241.
[7] W. Deng, F. Liu, H. Jin, X. Liao H. Liu, and L. Chen, "Lifetimeor energy: Consolidating servers with reliability control invirtualized cloud datacenters," in Proc. IEEE CloudCom, 2012,pp 18–25.
[8] Z. Zhou, F. Liu, H. Jin, B. Li, B. Li, and H. Jiang, "On arbitratingthe power-performance tradeoff in SaaS clouds," in Proc. IEEEINFOCOM, 2013, pp. 872–880.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 1, January 2019

[9] A. Gandhi, M. H.-BALTER, and R. Raghunathan, "AutoScale: Dynamic, robust capacity management for multi-tier data centers,"ACM Trans. Comput. Syst., vol. 30, no. 4, p. 14, Nov. 2012.

[10] Y. Ho, P. Liu, and J. Wu, "Server consolidation algorithmswith bounded migration cost and performance guarantees incloud computing," in Proc. IEEE Utility Cloud Comput., 2011,pp. 154–161.