Cost-Minimizing Dynamic Migration of Content Sharing Services into Hybrid Clouds

[1]Seeni Askar Ali. S, [2]Bala Murugan. G, [3]Bala Subramanian. N, [4]Mohamed Rafi. M

[1]Final year MCA, Mohamed Sathak Engineering College, Kilakarai

[2]Assistant Professor, Dept of MCA, Mohamed Sathak Engineering College, Kilakarai

[3]Associate Professor, Dept of MCA, Mohamed Sathak Engineering College, Kilakarai

[4]Professor, Dept of MCA, Mohamed Sathak Engineering College, Kilakarai

ABSTRACT: Cloud computing technologies, an upward number of content distribution applications are contemplating a control to cloud-based services, for better scalability and lower cost. Two input tasks are concerned for such a move: to migrate the contents to cloud storage and to share out the web service load to cloud-based web services. The main issue is to best use the cloud as well as the application provider's obtainable private cloud, to serve volatile requirements with service return time guarantee at all times, while incurring the minimum operational cost. While it may not be too difficult to design an effortless heuristic, proposing one with assured cost optimality over a long run of the system constitutes an intimidating challenge. Employing optimization techniques, we design a dynamic control algorithm to optimally place contents and send off requests in a hybrid cloud infrastructure spanning geo distributed data centers, which minimizes overall operational cost over time, subject to service response time constraints. Rigorous analysis shows that the algorithm well bounds the response times within the preset QoS target, and guarantees that the overall cost is within a small constant gap from the optimum achieved by a T-slot look ahead mechanism with known future in sequence. To verify the performance of our dynamic algorithm with prototype-based evaluation.

KEYWORDS: Dynamic Migration, Optimal Migration, Lyapunov Optimization Techniques.

1. INTRODUCTION

Cloud computing technologies have enabled rapid provisioning and release of server utilities (CPU, storage, and bandwidth) to users anywhere, anytime. To exploit the diversity of electricity costs and to provide service Proximity to users in different geographic regions, a cloud service often spans multiple data centers over the globe, e.g., Amazon Cloud Front, Microsoft Azure, Google App Engine. The elastic and on-demand nature of resource provisioning has made cloud computing attractive to providers of various applications. More and more new applications are being created on the cloud platform, while many existing applications are also considering the cloud ward move, including content distribution applications. As an important category of popular Internet services, content distribution applications,e.g., video streaming, web hosting and file sharing, feature large volumes of contents and demands that are highly dynamic in the temporal domain. A cloud platform with multiple, distributed data centers is ideal to host such a service, with substantial advantages over a traditional private or public content distribution network (CDN) based solution, in terms of more agility and significant cost reduction with respect to machines, bandwidth, and management. In this way, the application providers can focus their business more on content provisioning, rather than IT infrastructure maintenance. Two major components exist in a typical content distribution application, namely backend storage for keeping the contents, and front-end web services to serve the requests. Both can be migrated to the cloud: contents can be stored in storage servers in the cloud, and requests can be distributed to cloud-based web services. Therefore,

the key challenge for cloud-ward move of a content distribution application is how to efficiently replicate contents and dispatch requests across multiple cloud data centers, as well as the provider's existing private cloud, such that good service response time is guaranteed and only modest operational expenditure is incurred. It may not be too hard to design a simple heuristic for dynamic content placement and load distribution in the hybrid cloud; however, proposing one with guarantee of cost optimality over a long run of the system, is an intriguing yet intimidating challenge, especially when arbitrary arrival rates of requests are considered. Some existing work has advocated optimal application migration into clouds, but none focus on guaranteeing over-time cost minimization with a dynamic algorithm.

In this paper, we present a generic optimization framework for dynamic, cost minimizing migration of content distribution services into a hybrid cloud (i.e., private and public clouds combined), and design a joint content placement and load distribution algorithm that minimizes overall operational cost over time, subject to service response time constraints. Our design is rooted in Lyapunov optimization theory, where cost minimization and response time guarantee are achieved simultaneously by efficient scheduling of content migration and request dispatching among data centers. Lyapunov optimization provides a framework for designing algorithms with performance arbitrarily close to the optimal performance over a long run of the system, without the need for any future information. It has been extensively used in routing and channel allocation in wireless networks, and has only recently been introduced to address resource allocation problems in a few other types of networks. We tailor Lyapunov optimization techniques in the setting of a hybrid cloud, to dynamically and jointly resolve the optimal content replication and load distribution problems.

2. PROBLEM DESCRIPTION

Migration of Application into Clouds: A number of research projects have emerged in recent years that explore the migration of services into a cloud platform. Hajj at et al. develop an optimization model for migrating enterprise IT applications onto a hybrid cloud. Their model takes into account enterprise-specific constraints, such as transaction delays and security policies. One-time optimal service deployment is considered, while our work investigates optimal dynamic migration overtime, to achieve the long-term optimality. Propose an intelligent algorithm to factor workload and dynamically determine the service placement across the public cloud and the private cloud. Their focus is on designing an algorithm for distinguishing base workload and trespassing workload.

Migration of Content Delivery Services into Clouds: Some research efforts have been put into migrating generic content delivery services onto clouds. Meta CDN by Path is a proof-of-concept test bed, experiments on which show that deploying content delivery based on storage clouds can improve utility, based on primitive content placement and request routing mechanisms. Chen et al. propose to build CDNs in the cloud in order to minimize cost under the constraints of QoS requirement, but they only propose greedy-strategy based heuristics without provable properties. In contrast, we target an optimization framework which renders optimal migration solutions for long run of the system. Some work focuses on migrating specific types of content delivery services onto clouds, e.g., social networking service, or video streaming service. Change study the partition of social data and their storage onto a number of cloud servers, to migrate a social networking application into the cloud. It focuses on balancing the data access load, by considering social relationships and user access patterns in the data storage. Cost saving by partial migration of a VoD service to a content cloud. Heuristic strategies are proposed to decide the update of cloud contents, which are verified by trace-driven evaluations. Our work focuses on cost minimization in migration of a generic content distribution application, based on differentiated charging models of different data centers.

Application of Lyapunov Optimization Theory: Lyapunov optimization was developed from the stochastic network optimization theory and has been applied in routing and channel allocation in wireless networks as well as in aother types of networks including peer-to-peer networks propose various VM configuration scheduling algorithms for cloud computing platforms, that achieve arbitrary fraction of the capacity region of the cloud. But their model does not take into consideration delay guarantee, which is an important component in our optimization framework. The work of Rant al. also considers an online scheduler that dispatches workloads across multiple geographically distributed data centers subject to delay requirements. It assumes where each job's data is stored is fixed and known. However, in our work we further incorporate the decision on data migration into the scheduling. The work of close to ours in that it also utilizes Lyapunov function to study request routing and content caching, but in the setting of CDNs with capacitated caches and links. They investigate the

optimality of different caching policies. Given a workload within the capacity region, they prove that several types of caching and content eviction methods can each provide a throughput equal to the workload. Instead, our study focuses on optimal migration of content distribution services onto a hybrid cloud, such that the operational cost is minimized while.

3. PLANNED DESIGN

To consider a typical content distribution application, which provides a collection of contents (files), denoted as set M, to users spreading over multiple geographical regions. There is a private cloud owned by the provider of the content distribution application, which stores the original copies of all the contents. The private cloud has an overall upload bandwidth of b units for serving contents to users. There is a public cloud consisting of data centers located in multiple geographical regions, denoted as set N. One data center resides in each region. There are two types of inter-connected servers in each data center: storage servers for data storage, and computing servers that support the running and provisioning of virtual machines (VMs). Servers inside the same data center can access each other via a certain CN (Data Center Network). The provider of the content distribution application (application provider) wishes to provision its service by exploiting a hybrid cloud architecture, which includes the geo-distributed public cloud and its private cloud. The major components of the content distribution application include: (i) back-end storage of the contents and (ii) frontend web service that serves usersrequests for contents. The application provider may migrate both service components into the public cloud: contents can be replicated in storage servers in the cloud, while quests can be dispatched to web services installed on VMs on the computing servers. An illustration of the system architecture is given in. Our objective in this paper is to design a dynamic, optimal algorithm for the application provider to strategically make the following decisions for service migration into the hybrid cloud architecture: (i) content replication: which content should be replicated in which data center at each time? (ii) Request distribution: How many requests for content should be directed to the private cloud and to each of the data centers that store this content at the time? The goal is to pursue the minimum operational the system architecture. Cost for the application provider over time, while ensuring the service quality of content distribution Cost-Minimizing Service Migration Problem We supposes that the system runs in a time-slotted fashion. Each time slot is a unit time which is

enough for uploading any file m 2 M with size v (m) (bytes) at the unit bandwidth. In time slot t, a (m)j (t) requests are generated for downloading file m 2 M, from users in region j. We assume that the request arrival is an arbitrary process over time, and the number of requests arising from one region for a file in each time slot is upper-bounded by Amax. The cost of uploading a byte from the private cloud is. The charge for storage at data center i is pi per byte per unit time. gi and oi per byte are charged for uploading from and downloading into data center. The cost for renting a VM instance in data center is fi per unit time.

4. SYSTEM ARCHITECTURE

Fig:3.1 Dynamic Migration

To assume that the storage capacity in each data center is sufficient for storing contents from this content distribution application. We also assume that each request is served at one-unit bandwidth, and the number of requests that a VM in data center i can server unit time is r_i. The file can be carried out in parallel: after receiving a small portion of the file, a data center can already start to serve the received chunks of the file to users. To assume that upload bandwidth is reserved for replicating files to data centers from the private cloud, and this bandwidth is not counted in b, the maximum units of bandwidth that the private cloud can use to upload contents to users. Not all requests arising in one time slot are dispatched in the same time slot, subject to capacity constraints.

4.1. SERVICE FEATURE

The service quality experienced by users is evaluated by request response delay, consisting of two major components: queueing delay in the request queue, and round-trip delay from when the request is dispatched from the queue to the time the first byte of their quested file is received. We ignore the processing delay inside a data

center, due to the high inter-connection bandwidth and CPU capacities inside a data center. Let dj and ej i denote the round-trip delay between region j and the private cloud and between region j and datacenter i, respectively. Let be the upper-bound of the average round-trip delay per request, which the application provider wishes to enforce in this content distribution application. We reasonably assume this bound is larger than the round-trip delay between a user and the data center in the same region. I will show that our dynamic optimal service migration algorithm can bound both the average round-trip delay and queueing delay experienced by users.

4.2 OPERATIONAL EXPENDITURE

Our algorithm focuses on minimizing recurring operational cost of the content distribution system, not one-time costs such as the purchase of machines in the private cloud and contents.

5. CONCLUSION

In this paper, to present a generic optimization framework for dynamic, cost-minimizing migration of content distribution services into a hybrid cloud (i.e., private and public clouds combined), and design a joint content placement and load distribution algorithm that minimizes overall operational cost over time, subject to service response time constraints. Our design is rooted in Lyapunov optimization theory, where cost minimization and response time guarantee are achieved simultaneously by efficient scheduling of content migration and request dispatching among data centers. Lyapunov optimization provides a framework for designing algorithms with performance arbitrarily close to the optimal performance over a long run of the system, without the need for any future information. We propose a generic optimization framework for dynamic, optimal migration of a content distribution service to a hybrid cloud consisting of a private cloud and public geo-distributed cloud services. To design a joint content placement and load distribution algorithm for dynamic content distribution service deployment in the hybrid cloud. I tailor Lyapunov optimization techniques in the setting of a hybrid cloud, to dynamically and jointly resolve the optimal content replication and load distribution problems. To demonstrate optimality of our algorithm with rigorous theoretical

analysis and prototype-based evaluation. The algorithm nicely bounds the response times (including queueing and round-trip delays) within the preset QoS target in cases of arbitrary request arrivals, and guarantees that the overall cost is within a small constant gap from the optimum achieved by a T-slot look ahead mechanism with information into the future.

REFERENCES

[1] Amazon cloud front
http://aws.amazon.com/cloudfront/.

[2] Microsoft Azure,
http://www.microsoft.com/windowsazure/.

[3] Google App Engine,
http://code.google.com/appengine/.

[4] Dropbox, http://www.dropbox.com/.

[5] Microsoft Office Web Apps,
http://office.microsoft.com/enus/web-apps/.

[6] Google docs, http://docs.google.com/.

[7] M. Hajji, X. Sun, Y. E. Sung, D. Maltz, and S. Rao, ―Cloudward Bound: Planning for Beneficial Migration of Enterprise Applications to the Cloud, ‖ in Proc. of IEEE SIGCOMM, August 2010.

[8] H. Zhang, G. Jiang, K. Yoshihira, H. Chen, and A. Saxena, ―Intelligent Workload Factoring for a Hybrid Cloud Computing Model, ‖ in Proc. of the International Workshop on Cloud Services (IWCS 2009), June 2009.

[9] H. Li, L. Zhong, J. Liu, B. Li, and K. Xu, ―Cost-effective Partial Migration of VoD Services to Content Clouds, ‖ in Proc. of IEEE CLOUD, July 2011.

[10] X. Cheng and J. Liu, ―Load-Balanced Migration of Social Media to Content Clouds, ‖ in Proc. of NOSSDAV, June 2011.

[11] L. Georgiadis, M. J. Neely, and L. Tassiulas, ―Resource allocation and cross-layer control in wireless networks, ‖ Foundations and Trends in Networking, vol. 1, no. 1, pp. 1– 149, 2006.

[12] M. J. Neely, Stochastic Network Optimization with Application to Communication and Queueing Systems. Morgan & Claypool, 2010.