COST-MINIMIZING DYNAMIC MIGRATION OF CONTENT DISTRIBUTION SERVICES INTO HYBRID CLOUDS

M.Angel Jasmine Shirley[1], Dr.Suneel Kumar[2]

[1]Research Scholar, [2]Asst. Professor, Maharishi University Lucknow

ABSTRACT: With the recent advent of cloud computing technologies, a growing number of content distribution applications are contemplating a switch to cloud-based services, for better scalability and lower cost. Two key tasks are involved for such a move: to migrate the contents to cloud storage, and to distribute the web service load to cloud-based web services. The main issue is to best utilize the cloud as well as the application provider's existing private cloud, to serve volatile requests with service response time guarantee at all times, while incurring the minimum operational cost. While it may not be too difficult to design a simple heuristic, proposing one with guaranteed cost optimality over a long run of the system constitutes an intimidating challenge. Employing Lyapunov optimization techniques, we design a dynamic control algorithm to optimally place contents and dispatch 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 nicely 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 lookahead mechanism with known future information. We verify the performance of our dynamic algorithm with prototype-based evaluation.

1. INTRODUCTION

BACKGROUND OF THE STUDY

Cloud computing technologies have enabled rapid provisioning and release of server utilities (CPU, storage, 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 back-end 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 have 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 DEFINITION

Migration of applications into clouds: A number of research projects have emerged in recent years that explore the migration of services into a cloud platform. Hajjat 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. Zhang et al. 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 an et al. 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.

Cheng etal. 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. Li et al. advocate 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 a few other types of networks including peer-to-peer networks Magulurietal 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 Ren et 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 Amble etal. is 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. PROPOSED DESIGN

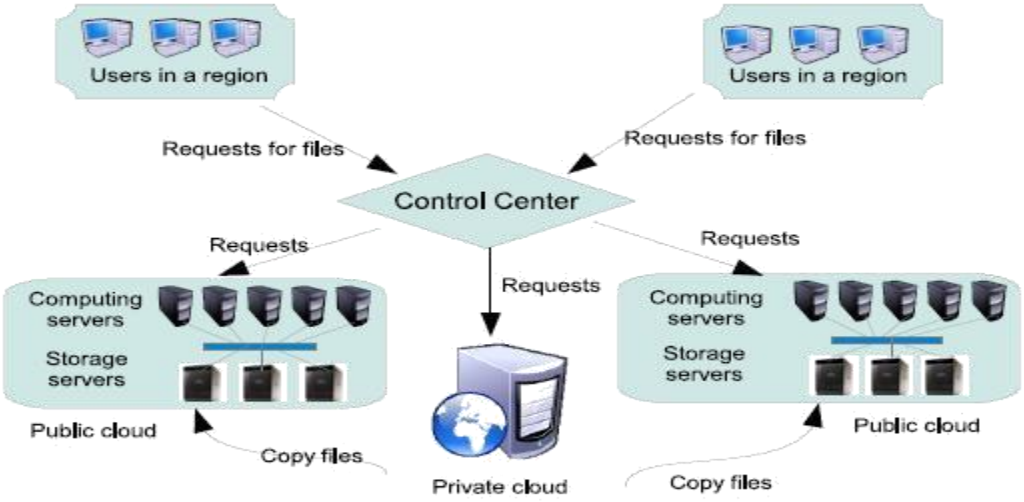
We consider a typical content distribution application, which provides a collection of contents (files), denoted as set M, to users spreading over multiple geographical regions. There is a private cloud owned by the provider of the content distribution application, which stores the original copies of 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)front-end web service that serves users' requests for contents. The application provider may migrate both service components into the public cloud: contents can be replicated in storage servers in the cloud, whilere quests can be dispatched to web services installed on VMs on the computing servers. An illustration of the system architecture is given in Fig. 1.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 a 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 Fig. 1. The system architecture. cost for the application provider over time, while ensuring the service quality of content distribution

Cost-Minimizing Service Migration Problem We suppose 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 ish. The charge for storage at data center i is pi per byte perunit time. gi and oi per byte are charged for uploading from and downloading into data center i, respectively .The cost for renting a VM instance in data center i is fi per unit time. These charges follow the charging model of leading commercial cloud providers, such as AmazonEC2 [25] and S3[26]. We 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 ri. 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. We 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.

SYSTEM ARCHITECTURE



Service quality. 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 there 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 _ >eii; 8i 2 N, i.e., this bound is larger than the round-trip delay between a user and the data center in the same region. We will show that our dynamic optimal service migration algorithm can bound both the average round-trip delay and queueing delay experienced by users.

Operational cost. 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.

4. CONCLUSION

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. 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. We design a joint content placement and load distribution algorithm for dynamic content distribution service deployment in the hybrid cloud.

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. We 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 CloudFront, 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. Hajjat, 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.

[13] Energy optimal control for time varying wireless networks,‖ IEEE Tran. On Information Theory, no. 7, pp. 2915–2934, July 2006.

[14] M. M. Amble, P. Parag, S. Shakkottai, and L. Ying, Content- Aware Caching and Traffic Management in Content Distribution Networks,‖ in Proc. of IEEE INFOCOM, April 2011.

[15] M. J. Neely and L. Golubchik, ―Utility Optimization for Dynamic Peer-to-Peer Networks with Tit-For-Tat Constraints,‖ in Proc. Of IEEE INFOCOM, April 2011.

[16] M. Pathan, J. Broberg, and R. Buyya, ―Maximizing Utility for Content Delivery Clouds,‖ in Proc. of the 10th International Conference on Web Information Systems Engineering, 2009.

[17] F. Chen, K. Guo, J. Lin, and T. L. Porta, ―Intra-cloud Lightning: Building CDNs in the Cloud,‖ in Proc. of IEEE INFOCOM, 2012.

[18] H. Li, W. Huang, C. W. abd Z. Li, and F. C. Lau, ―Utility-Maximizing Data Dissemination in Socially Selfish Cognitive Radio Networks,‖ in Proc. of IEEE International Conference on Mobile Ad-hoc and Sensor Systems (IEEE MASS 2011), Oct 2011.

[19] S. T. Maguluri, R. Srikant, and L. Ying, ―Stochastic Models of Load Balancing and Scheduling in Cloud Computing Clusters,‖ in Proc. of IEEE INFOCOM, 2012.

[20] S. Ren, Y. He, and F. Xu, ―Provably-Efficient Job Scheduling for Energy and Fairness in Geographically Distributed Data Centers,‖ in Proc. of IEEE ICDCS, 2012.

[21] N. Laoutaris, G. Smaragdakis, K. Oikonomou, I. Stavrakakis, and A. Bestavros, ―Distributed Placement of Service Facilities in Large-Scale Networks,‖ in Proc. of IEEE INFOCOM, 2007.

[22] J. Leblet, Z. Li, G. Simon, and D. Yuan, ―Optimal Network Location in Distributed Virtualized Data-Centers,‖ Computer Communications, no. 16, pp. 1968–1979, 2011.