An architecture for selective caching and content distribution as an on-demand service for communities

Vittorio Manetti
COMICS Lab, Dipartimento di Informatica e Sistemistica
Università di Napoli Federico II, Via Claudio 21, 80125 Napoli, Italy
Email: vittorio.manetti@unina.it

***Abstract*— The multimedia contents provided by Internet are very heterogeneous, however it is possible to identify communities of users, so defined regarding the Internet access infrastructure used, and the common interests. The communities of users are often organized like subgroups, and they exploit services provided by third parties in order to receive better performance in regard to the fruition of contents given by one or more Content Providers. The main goal of our work is to realize a network infrastructure in which a cache system is available on demand and in a dynamic way, in order to increase the quality of service perceived by users.**

I. Introduction

A community can be considered like a collection of users with the same interests. By definition, community users can be located in the same zone and they can share the same network infrastructure in order to access Internet contents; users are often collected in clusters. As an instance, we can consider students from the same University like a community of users.

The generic Content Provider gives contents accessible by users; these last require contents with an high rate, and often with very stringent Quality of Service features. Usually, the Content Provider has the task to solve QoS problems; in the scenario we are depicting, the community by itself assures high performance in terms of QoS regarding the content distribution. The reasons: firstly, services provided by Content Providers can be very expensive; secondly, it is possible that the Content Providers are not be able to manage the QoS requested by users.

So, we can also consider a Service Provider: its task is to guarantee the required QoS. In this scenario, this kind of service is linked with the implementation of caching mechanisms that can decrease the latency perceived by a user when he is accessing to the requested content. We propose this solution in order to allow the community to select contents for which this service may be activated. Caches are provided by ISPs or other third parties as a service, or alternatively can be part of the community network infrastructure. In the first case, we have to take into account the cache management fee imposed by the Content Providers.

The main goal of our work is, on one hand, to define an objective function that has to determine the optimal placement of content replicas in a set of candidate web caches, considering a set of cost parameters. On the other hand, we have to realize a network infrastructure for content distribution to user communities, in order to minimize the aforementioned objective function. Considering both tasks, we can imagine by now two distinct scenarios: the first one for web contents delivery; the second one for multimedia flow streaming.

The rest of this paper is organized as follow: in section II we define a model for optimal placement of content replicas; in section III we introduce the features of an architecture for content distribution; in section IV we conclude the work.

II. Definition of a model for optimal placement of content replicas

In order to realize the model concerning the problem to solve, we proceed by incremental steps. We start from the Simple Plant Location classical model; this model provides the possibility to localize services with the main goal to minimize the total cost. We would like to exploit this kind of approach: exact solutions for medium-little networks, and heuristics development for big networks. On a first step, we consider a full mesh network. The objective function takes into account the following parameters:

- Objects update rate
- Request rate for a specific object
- Cache activation and cache management fee
- Cache space availability
- Cache localization fee
- Hit rate and miss rate on a specific cache
- Potential cache overload
- Objects transfer fee (server-cache / cache-client))

The model is formulated in a way in which a request from a very asset client is too powerful in comparison with the request from a less asset client. The defined constraints assure that:

- the generic request is attended only by open servers
- the caches have to satisfy the server requests
- the client requests have to be served by the cache

Moreover, it is necessary to consider the limited capacity of the generic cache; a cache can not contain an amount of contents that goes over the maximum capacity. Each content into a cache, has to be a specific dimension. The definition of the objective function has to be considered like a work in

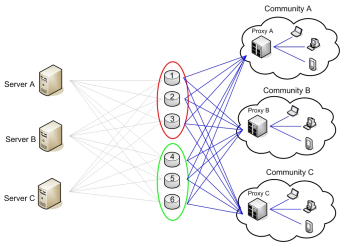

Fig. 1. Run Phase

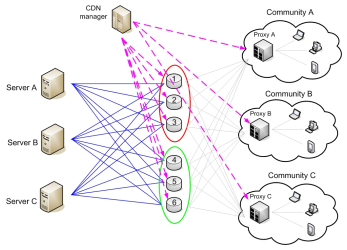

Fig. 2. Control Plane

progress; the main goal is to define a function that can count all the parameters in the introduced model, and to increase the complexity of the function by subsequent steps.

III. Design and Implementation of an Architecture for Content Distribution

The realization of the architecture requires the definition of new entities for management of content replicas into the cache system, in order to optimize the Quality of Experience measured by the generic community user. We introduce a Proxy node and a CDN Manager node.

A. The Proxy node

Policies for request routing content-aware have to be provided to the community users. The adopted solution consists in introduce a Proxy node for each cluster of users; all the users in a cluster, as aforementioned, share the same network infrastructure. The Proxy has to forward the user request; obiviously, this task is based on the content replicas location. The Proxies have to be instructed by CDN Manager on the routing policies to adopt, in order to forward the request on the right cache. The contents are organized in a way in which different caches can provide the same content. There is a direct connection between client and Proxy, and no direct connection between client and cache; a community Proxy refers to a single cache for a specific object, and it can refer to several caches in the same time in order to satisfy multiple requests of differet objects. Each proxy is equipped by a table with the binding between URL and cache identifier.

B. The CDN Manager node

The CDN Manager has to determine the optimal placement of content into the cache system; this task is based on the objective function previously designed. As we explained above, this function solves the location problem considering a set of cost parameters and metrics opportunely established. The CDN Manager, during the run phase, has to obtain information needed to compute the objective function, and, based on the achieved results, it has to determine the optimal placement of content replicas into the cache system. The CDN Manager collects information from servers, caches and Proxies, and it deliveries to caches and Proxies information needed to distribute the content replicas. In other words, it computes the optimal way to deploy content into the caches, and the best configuration of the proxy tables.

IV. Conclusion and Future Work

The proposed work is about the design and implementation of a system for optimal placement of multimedia contents for community of users; the model is based on the use of a distributed cache system and on the computation of a predetermined objective function. The work consists, on one hand, in the definition of the objective function, and, on the other hand, in the realization of a network infrastructure. Regarding this second task, we introduce in the architecture two specific nodes and their functionalities: Proxy and CDN Manager.

The proposed model has to be considered like a work in progress. In order to complete this model we have to take into account several aspects: further metrics have to be defined to enrich the objective function; some policies for content routing implemented on the Proxy have to be defined; we have to consider the possibility to realize a cache infrastructure organized in a hierarchical manner; we have to consider the possibility to exploit peer-to-peer technologies. In order to validate the effectiveness and the performance of our model, we also have to realize simulations and emulations, exploiting, for instance, a distributed system like PlanetLab.

V. Acknowledgements

This work has been supported by the European Union under the IST Content (FP6-2006-IST-507295) project. The CONTENT Network of Excellence targets Content Delivery Networks for Home Users, as an integral part of Networked Audio-Visual Systems and Home Platforms.

References

[1] N. Laoutaris, G. Smaragdakis, K. Oikonomou, I. Stavrakakis, A. Bestavros, *"Distributed Placement of Service Facilities in Large-Scale Networks,"* to appear in IEEE INFOCOM 2007.

[2] N. Laoutaris, V. Zissimopoulos, I. Stavrakakis, *"Joint Object Placement and Node Dimensioning for Internet Content Distribution,"* Information Processing Letters, Vol. 89, No. 6, pp. 273-279, March 2004.

[3] N. Laoutaris, V. Zissimopoulos, I. Stavrakakis, *"On the Optimization of Storage Capacity Allocation for Content Distribution"*, Computer Networks, Vol. 47, No. 3, pp. 409-428, February 2005.

[4] W. Shi, Y. Mao, *"Performance evaluation of peer-to-peer web caching systems"*, Journal of Systems and Software, Volume 79, Pages: 714 - 726, Year of Publication: 2006, ISSN:0164-1212

[5] G. Tsuchida, T. Okino, T. Mizuno, S. Ishihara, *"Evaluation of a replication method for data associated with location in mobile ad hoc networks,"*, ICMU'05, pp. 116–121, 2005.

[6] A. Wierzbicki, *"Models for internet cache location"*, in the 7th Int. Workshop on Web Content Caching and Distribution (WCW), 2002.

[7] A. Vakali, G. Pallis, *"Content delivery networks: Status and trends"*, IEEE Internet Computing, 7(6):68.74, December 2003.