Available online at www.sciencedirect.com
ScienceDirect
Computer Networks 50 (2006) 3763–3783
Computer Networks
www.elsevier.com/locate/comnet

Session based access control in geographically replicated Internet services

Novella Bartolini *

Computer Science Department, University of Rome "La Sapienza", Via Salaria 113, 00198 Rome, Italy

Received 18 June 2005; received in revised form 12 March 2006; accepted 27 March 2006
Available online 17 May 2006

Responsible Editor: J. Misic

Abstract

Performance critical services over Internet often rely on geographically distributed architectures of replicated servers. Content Delivery Networks (CDN) are a typical example where service is based on a distributed architecture of replica servers to guarantee resource availability and proximity to final users. In such distributed systems, network links are not dedicated, and may be subject to external traffic. This brings up the need to develop access control policies that adapt to network load changing conditions. Further, Internet services are mainly session based, thus an access control support must take into account a proper differentiation of requests and perform session based decisions while considering the dynamic availability of resources due to external traffic.

In this paper we introduce a distributed architecture with access control capabilities at session aware access points. We consider two types of services characterized by different patterns of resource consumption and priorities. We formulate a Markov Modulated Poisson Decision Process for access control that captures the heterogeneity of multimedia services and the variable availability of resources due to external traffic. The proposed model is optimized by means of stochastic analysis, showing the impact of external traffic on service quality. The structural properties of the optimal solutions are studied and considered as the basis for the formulation of heuristics. The performance of the proposed heuristics is studied by means of simulations, showing that in some typical scenario they perform close to the optimum.

Keywords: Content Delivery Networks; QoS; Session based access control

1. Introduction

Geographical replication of resources is at the basis of several performance critical services over the Internet. Content Delivery Networks (CDN) [1] are based on a placement of server replicas and on mechanisms for request redirection that guarantee resource availability, service quality and proximity of content to the user, together with an efficient and content aware routing. A proper placement of replica servers shortens the path from servers to clients thus hedging the risk of encountering bottlenecks in the non-dedicated environment of the Internet.

* Tel.: +39 06 49918357; fax: +39 06 854 1842.
E-mail address: novella@di.uniroma1.it

1389-1286/$ - see front matter
doi:10.1016/j.comnet.2006.03.007

Flash crowds and unpredictable link congestions could cause a critical performance degradation of some servers leaving few resources available to grant service availability and continuity. A simple scheme of replica placement and request redirection may not be sufficient to solve this problem. Options include increasing the capacity of servers and of network links and replicating more servers. However in many circumstances it is impossible to estimate the amount of resources required to fulfill all requests. In such a finite server capacity scenario, service-level agreements (SLA) specifying quality of service (QoS) probabilistic guarantees must be in place and penalties should be imposed when requests are not served or are served in violation of the agreements on quality.

Our study is valid for many types of service, but to make the analysis more concrete, we study the workload of two typical web services: informational web access and e-commerce requests and transactions. See [2] for a survey on types of services typically supported by CDN and more generally by geographically distributed replicated architectures. The list include informational services and e-commerce services among the others. We refer to [3] for the description of replicated architecture supporting dynamic caching for e-commerce sites.

Typical Internet services are based on the concept of *session* [4–6]. A session is a sequence of temporally and logically related requests issued by the same client. We believe that an access control mechanism that gives priority to requests belonging to an already active service session, and to critical phases of each session, could improve the client perceived performance and the profits of the service providers. The session concept must be at the basis of any access control mechanism in Web and more generally Internet services, as pointed out in [4,6]. If a service session has been started, all its composing requests should also be admitted, especially during critical phases in which more revenue could be gained or lost.

Aim of this paper is to investigate the performance of session based access control policies in a non-dedicated network environment, in presence of external traffic and possible congestions.

The proposed policies can be made available at any appliance that performs request interception and redirection over a distributed architecture. From now on, we will refer to this appliance with the name of *access point*.

Service sessions can be modelled as a sequence of service phases alternated to think phases. The sequence of phases traversed by the session during its lifetime strictly depends on the particular application. We formulate models of service sessions and underline the existence of critical phases that should not be interrupted in order not to lose profits and incur penalties.

We propose a stochastic decision model to optimize some performance parameters such as the probability of ongoing *session disruption* due to lack of resources, the probability of successful *session completion*, i.e. the probability that a session is terminated due to the client's will, and the *service refusal* probability that is the probability that a request is blocked by the admission control mechanism at the first access attempt.

The proposed decision model is based on a Markov Modulated Poisson Process (MMPP) [7] of service sessions that captures the dynamic nature of external traffic [8,9], and the phase model typical of multimedia network applications accessible through the Internet.

External traffic on the non-dedicated links is modelled by a Markov modulated vacation (or ON/OFF) process.

Through the proposed model, the optimal policy can be computed by means of common methods of operations research such as the value iteration algorithm [10,11].

The analysis of the optimal policy shows the impact of external traffic on service quality and how the access controller can adapt its decision to the external traffic scenario.

A structural analysis of the optimal solution is conducted to suggest a choice of possible heuristics to be applied by the access routers. Performance comparisons between the heuristics and the optimal policies are also given by means of simulations. Application traffic is simulated by means of a synthetic traffic generator that follows the proposed model of session lifetime. The use of a synthetic traffic generator is justified by the immense variability of application session model, that makes impossible the use of real traces to model sufficiently general situations.

The paper is organized as follows. In Section 2, we introduce the state of the art of admission control in web systems and in particular in CDN. In Section 3, we give some details on the reference architecture to enable access control to replicated servers. In Section 4, we introduce two typical CDN types of service and their related Markov modulated phase model: a purely informational