A QOS AFFIRMATION SYSTEM MODEL FOR CONTENT DELIVERY NETWORK

Manivannan .K
Department of Computer Science and Engineering,
PSNA College of Engineering and Technology, Dindigul, INDIA
Email : manivannan@psnacet.edu.in

Abstract

Examining of Content Delivery Networks (CDNs) can be a way to allow mutual aid between CDNs in a scalable manner and to achieve better overall service, as perceived by end-users. A CDN is expected to provide high performance Internet content delivery through global coverage, which might be a barrier for new CDN providers, as well as affecting commercial viability of the existing ones. A Quality of Service (QoS) driven performance modeling approach is carried out for peering CDNs, in order to predict its user performance. The peering between CDNs upholds user perceived performance by satisfying the target QoS is shown. The methodology presented in this paper provides CDNs a way to dynamically redirect user requests to other peering CDNs according to different request-redirection policies. The model-based approach helps an overloaded CDN to return to normal by offloading excess requests to the peers. It also assists in making concrete QoS guarantee for a CDN. Overall this model undertakings to achieve scalability for a CDN in a user transparent manner.

Keywords: Content Delivery Networks (CDNs), Quality of Service (QoS), Request-redirection, Service Level Agreements (SLAs)

I. INTRODUCTION

Content Delivery Networks (CDNs) are networks of surrogate servers spanning the Internet, aiming to offer fast and reliable Web services by distributing content to edge servers located close to end-users. The main objective of a CDN is to deliver competitive services according to user Quality of Service (QoS) requirements. The current deployment approach for a CDN requires building a global network of surrogate servers to host replicated content. Running a global CDN is challenging in financial, technical and administrative terms, both for deployment and operation of service. Furthermore, commercial CDNs make specific commitments to their customers by signing Service Level Agreements (SLAs). The SLA is a contract between the provider and the customer to describe the provider's commitment and to specify penalties if those commitments are not met. So, if a CDN is unable to provide QoS for user requests, it may result in SLA violation and end up costing the provider.

This large Internet-wide cooperation can be termed as a 'peering arrangement' or internetworking between CDNs, where some CDN providers may team up at some point in time to share resources and form an alliance in order to respond to or exploit a particular niche. It virtualize multiple providers, and provides flexible resource sharing and dynamic collaboration between autonomous individual CDNs. In such a system, a CDN serves user requests as long as the load can be handled by itself. If the load exceeds its capacity, the excess user requests are offloaded to the Web servers of the peers. The general objective of a peering CDNs model is to provide improved QoS performance through minimizing end-user response time. However, the proprietary nature of existing commercial CDNs makes it difficult to predict the performance a given user is expected to experience from a particular CDN.

An approach to perform QoS-driven modeling of the peering CDNs based on the fundamentals of queuing theory. The performance comparison of four request-redirection policies within the peering CDNs model is also done. The aim is to show that the cooperation between CDNs through a peering arrangement upholds user perceived performance by providing target QoS according to SLAs. QoS performance models can be used to reveal the effects of peering and to predict end-user perceived performance. This approach undertakes to assist in making concrete QoS guarantees by a CDN provider. The main contributions are:

- performance models to demonstrate the effects of peering and to predict user perceived performance;
- systematic performance analysis and measurement-based methodology to study the impact of key performance parameters such as server load and measurement errors that can be expected in a realistic system

- an approach to measure the QoS level of a CDN provider to ensure it provides efficient services; and

II. EXISTING WORKS

Research on Web server selection to redirect end-user requests has thus far focused primarily on techniques for choosing a server from a group administered by a single CDN. Peering of CDNs is gaining popularity among researchers of the scientific community, since such cooperation between CDNs can achieve better overall service, as perceived by end-users. Some projects are being conducted for finding ways to allow peering between CDNs. But many of them lack in virtualizing multiple CDNs for the management and delivery of content in a cooperative environment, and directing end-user requests to different CDNs based on performance to satisfy user QoS requirements.

Protocol architecture [14] for CDI attempts to support the interoperation and cooperation between separately administered CDNs. In this architecture, performance data is interchanged between CDNs before forwarding a request by an authoritative CDN (for a particular group), which adds an overhead on the response time perceived by the users. Moreover, being a point-to-point protocol, if one end-point is down the connection remains interrupted until that end-point is restored. Since no evaluation has been provided for performance data interchange, the effectiveness of the protocol is unclear.

CDN brokering [16] allows one CDN to intelligently redirect end-users dynamically to other CDNs in that domain. Though it provides benefits of increased CDN capacity, reduced cost and better fault tolerance, it does not explicitly consider the end-user perceived performance to satisfy QoS while serving requests. Moreover, it demonstrates the usefulness of brokering rather than to comprehensively evaluate a specific CDN's performance.

An approach to model traffic redirection [17] in geographically diverse server sets uses a novel metric Server Set Distance (SSD) to simplify the modeling and classification of redirection scheme. Though this modeling provides a foundation for intelligent server selection over multiple, separately administrated server pools, it does not try to show the effectiveness of any particular policy or evaluate the QoS performance of any given CDN. WARD (Wide Area Redirection of Dynamic Content) [18] presents a novel architecture for redirecting dynamic content requests from an overloaded Internet Data Center (IDC) to a remote replica. It demonstrates a simple analytical model to characterize the effects of wide area request-redirection on end-to-end delay.

III. AN APPROACH

A CDN serves end-user requests as long as the load can be handled by itself. If the load exceeds its capacity, the excess end-user requests are offloaded to the Web servers of other cooperating CDNs. A peering arrangement of CDNs is formed by a set of autonomous CDNs {CDN1, CDN2… CDNn}, which cooperate through a mechanism M that provides facilities and infrastructure for cooperation between multiple CDNs for sharing resources in order to ensure efficient service delivery. Assume S = {S1, S2… Sm}as the set of services provided by a CDN.

Figure 1, shows an abstraction of the peering CDNs. The initiator of a peering negotiation is called a primary CDN; while other CDNs who agree to provide their resources are called peering CDNs. The endpoint of a peering negotiation between two CDNs is a contract (SLA) that specifies the peer resources (Web servers, bandwidth etc.) that will be allocated to serve content on behalf of the primary CDN. The primary CDN manages the resources it has acquired insofar that it determines what proportion of the Web traffic (i.e. end-user requests) is redirected to the Web servers of the peering CDNs.

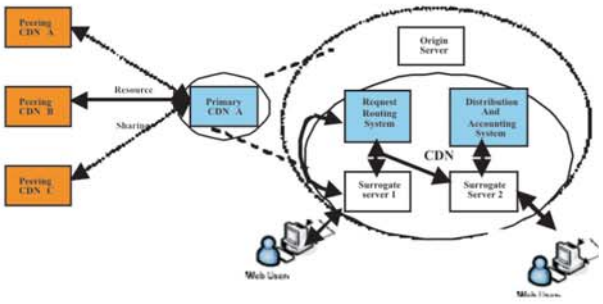

Fig. 1. Peering Abstraction CDNs

QoS in peering CDNs

QoS performance can be measured based on the user's experience of a service to compare the 'promise' against the delivery'. The definition of quality varies from different perspectives and views. This model adopts the conformance view and defines QoS as the experience perceived by a user when being served by a CDN.

Let A be a CDN provider and S = {S1, S2… Sm}be the set of services provided by it. Assume that for each service Si, Sip is the quality that A promised to offer to the

users and Sid is the actual quality delivered by A. Then, the QoS for CDN A is given by,

$QoSA = f(Si^p, Si^d)$

Where, f is the function that measures the conformance between Si^p and Si^d.

SLAs to ensure QoS

Ensuring QoS guarantees requires a means of establishing a set of common quality parameters and establishing which attributes are needed by a particular customer to describe its QoS requirements. These factors are combined to an SLA that both a customer and provider agree to and that the provider refers to when monitoring its QoS performance. Examples of QoS parameters that an SLA may specify are:

- The maximum response time for a service request will not exceed 0.4 seconds.
- 96% of user requests will be completed in less than 2 seconds.
- A service will be available for at least 99.9% of the time.

IV. PERFORMANCE MODELS

Develop the performance models based on the fundamentals of queuing theory to demonstrate the effects of peering between CDNs and to characterize the QoS performance of a CDN.

Single CDN model

Let us model a CDN as an M/G/1 queue as shown in Figure 2. The request streams coming to the Web servers of a CDN are abstracted as a single request stream. User requests arrive following a Poisson process with mean arrival rate λ. All requests in its queue are served on a first-come-first-serve (FCFS) basis with mean service rate μ. It is assumed that the total processing of the Web servers of a CDN is accumulated through the server and the service time follows a general distribution. The term 'task' is used as a generalization of a request arrival for service. The processing requirement of an arrival is 'task size'.

Hyper-exponential approximation

In order to quantify the performance perceived by the users while being served by a CDN, find the P.D.F. of waiting time distribution. The Bounded Pareto distribution has all moments finite; however advanced analysis is complex due to the difficulties in manipulating the Laplace transforms of the queuing metrics (e.g. waiting time, busy period). Hence, the 'heavy-tailed' Bounded Pareto distribution can be approximated with a series of exponential distributions (known as Hyper-exponential), while still maintaining the main characteristics of the original service distribution, such as heavy tail, first and second moments [9].

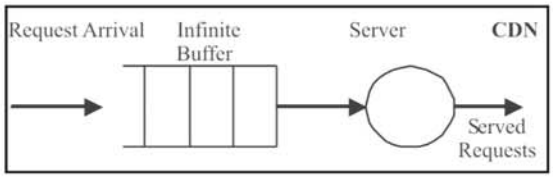

Fig. 2. Model of an M/G/1 queue

Service distribution and waiting time

The Laplace transform of the service distribution hn(t) is

$$L_{h_n}(s) = \int_0^\infty e^{-st} h_n(t)dt = \sum_{i=1}^{n} \frac{P_i \lambda_i}{\lambda_i + s}$$

The moments of the service distribution can be obtained as

$$E[X^n] = (-1)^n \left. \frac{d^n L_{h_n}(s)}{ds} \right|_{s=0}$$

Where, X is a continuous random variable with P.D.F hn(t). The first moment (mean) E[X] and the second moment E[X2] are

$$E[X] = \sum_{i=1}^{n} \frac{P_i}{\lambda_i} \text{ and } E[X^2] = \sum_{i=1}^{n} \frac{2P_i}{\lambda_i^2}$$

The Laplace transform of the waiting time, W(s) for an M/G/1 queue with the hyper-exponential approximation of a Bounded Pareto distribution is defined as follows

$$L_W(s) = \frac{s(1-\rho)}{s - \lambda + \lambda L_{h_n}(s)} = \frac{s(1-\rho)}{s - \lambda + \lambda \sum_{i=1}^{n} \frac{P_i \lambda_i}{\lambda_i + s}}$$

Peering CDNs Model

A CDN's inability to meet user QoS requirements according to the SLAs may lead to a collaboration of CDNs, so that it may redirect excess requests to the Web servers of the peers. In Figure 3, a conceptual view of the peering CDNs is provided where each CDN is modeled as an M/G/1 queue

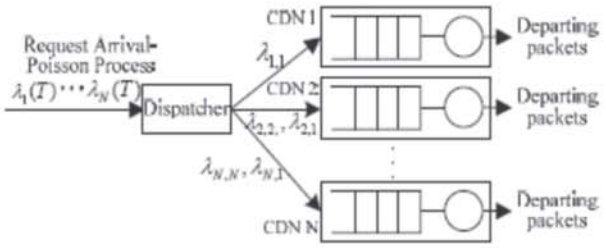

Fig. 3. Conceptual view of the peering CDNs

Arrive at a conceptual entity, called dispatcher, following a Poisson process with the mean arrival rate The dispatcher acts as a centralized scheduler in a particular peering relationship with independent mechanism to distribute content requests among partnering CDNs in a user transparent manner. If, on arrival, a user request can not be serviced by CDN i, it may redirect excess requests to the peers. Since this dispatching act on individual requests of Web content, it endeavors to achieve a fine grain control level. Anticipation is that it can also pave the ways in performing the request assignment and redirection at multiple levels – at the DNS, at the gateways to local clusters and also (redirection) between servers in a cluster. Thus, end-users can be assigned via DNS (by the CDNs updating their DNS records regularly) and also via redirection at the gateway (through dispatching) when appropriate [1]. The dispatcher follows a certain policy that assists to assign a fraction of requests of CDN I to CDN j. The request stream to a CDN in the peering CDNs model is defined as λj,i = request to CDN j for CDN i's content. For j ≠ i, λj,i denotes redirected user requests, where CDN i is the primary where CDN j is a peer. On the other hand, for j = i, λj,i denotes the user requests to a primary CDN i. For example, request to CDN B for CDN A's content can be denoted as λB,A. A peer always prioritizes the requests from the primary CDN over its own user requests. However, if a redirected request (higher priority) arrives to a peer when its own user request (lower priority) is being served, it never interrupts the current service. Thus, this priority discipline is non-preemptive during service quantum of end-user requests achieving effective service from it. he P.D.F of the waiting time distribution (through numerical inversion) for each given (primary) CDN, with independent priority class can be used to observe the expected waiting time perceived by majority of users in the peering CDNs system. Since a primary CDN's request has priority over any peer's own user requests, consider using (2) for a primary CDN, while (3) for any peering CDN. Though these equations are useful for computation, the iterative expression for G* i(s) in (4) is impossible to invert numerically. Therefore, the waiting time experienced by a primary CDN's user requests is found using (2), while the classical result presented in (1) is used to find the average expected waiting time for the peer's user requests.

V. EXPERIMENTAL RESULTS

The performance results are obtained using the models presented in Section 4. Each CDN is modeled as an M/G/1 queue with highly variable hyper exponential distribution which approximates a heavy-tailed Bounded Pareto service distribution (α, k, p) with variable task sizes. Thus, the workload model incorporates the high variability and self-similar nature of Web access. Table 1 shows the distributions, probability density functions and parameter ranges for the workload model. For experiments, consider the expected waiting time as an important parameter to evaluate the performance of a CDN. This peering scenario, an SLA of serving all user requests by the primary CDN in less than 20000 time units.

QoS performance of the primary CDN

First, provide the evidence that a peering arrangement between CDNs is able to assist a primary CDN to provide better QoS to its users. The C.D.F of the waiting time distribution of the primary CDN can be used as the QoS performance metric. In a highly variable system such as peering CDNs it is more significant than average values. The waiting time corresponds to the time elapsed by a user request before being served by the CDN. Figure 4 shows the C.D.F of waiting time of the primary CDN without peering at different loads. The Figure shows that for a fair load ρ = 0.6 there is about 55% probability that users will have a waiting time less than the threshold of 20000 time units. For a moderate load ρ = 0.7, there is about 50% probability for users to have waiting time below the threshold, while for a heavy load ρ = 0.9 the probability reduces to > 24%.

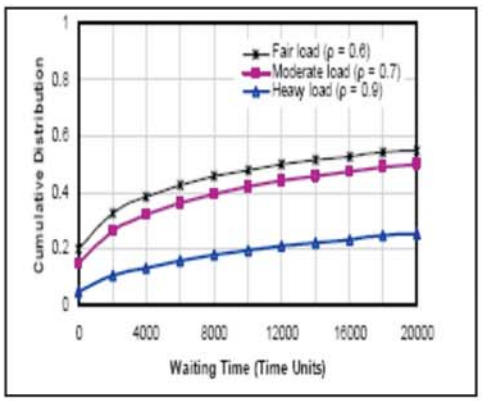

Fig. 4. Cumulative distribution of waiting time of the primary CDN without peering

Request-redirection policies

In the peering CDNs model, no redirection is assumed until primary CDN's load reaches a threshold load (ρ = 0.5). This load value is also used as the baseline load for comparing waiting times at different primary CDN loads. Any load above that will be 'shed' to peers.

A request-redirection policy determines which requests have to be redirected to the peers. Each peer is ready to accept only a certain fraction (acceptance threshold) of the redirected requests. Any redirected request to a given peer exceeding this acceptance threshold is simply dropped to maintain the system equilibrium. In face of sudden surge in demand, the load on a given primary CDNi, i={1,2,…,N}becomes, ρ* i = ρi –ρi redirect and the redirected load is distributed among the peering CDNs. The value of ρi redirect varies depending on the dispatcher chosen redirection policy. The initial and new load on the given primary CDN i is measured by, ρi =λi,iE[Xi] and ρ* i =λ*i,iE[Xi] respectively. λi,i is the initial arrival rate, whereas, *i,i = λi,i – λi rdirect is the new arrival rate.

Four request-redirection policies for evaluation within the peering CDNs model can defined as follows:

- Uniform Load Balanced (ULB) request-redirection policy distributes the redirected content requests uniformly among all the peering CDNs.
- Minimum Load Balanced (MLB) request-redirection policy assigns the redirected content requests to the peer with minimum expected waiting time.
- of traffic is uniformly distributed over all other participating peers.

VI. CONCLUSION AND FUTURE WORK

An innovative approach to model the peering CDNs is proposed here. Through the presented performance models, the demonstration of the effects of peering and predicted end-user perceived performance from a primary CDN is shown. A measurement-based methodology which endeavors to assist in making concrete QoS guarantees by a CDN provider is outlined. This approach assists an overloaded CDN to immediately stabilize by offloaded a fraction of the incoming content requests to the peers. This model provides a foundation for performing effective peering between CDNs though achieving target QoS in service delivery to end-users. Since the peering CDNs retain load-balancing control within their own Web server sets, using the approach a primary CDN can realize the QoS performance it can provide to the end-users, without requiring individual partners to provide expected service performance from it. A model-based approach is important since having each CDN provider communicate how it would service millions of potential end-users would introduce significant scalability issues, and requesting this information from each partnering provider at the user requests time would introduce substantial delays. The future work includes performing an advanced system analysis to study the impact of other performance parameters such as network latency and cost of peering. It also includes developing a proof-of-the-concept implementation for demonstrating the real-time application of the approach for peering between CDNs.

This methodology for modeling peering CDNs and predicting performance of a CDN provider in a peering arrangement will be a timely contribution to the content networking trend in the infrastructure-based CDNs domain.

REFERENCES

[1] Pathan A. M. K, Broberg J, Bubendorfer K, Kim K. H, and Buyya R, 2007, "Architecture for Virtual Organization (VO)-Based Effective Peering of Content Delivery Networks", UPGRADE-CN'07, In Proc. of the 16th IEEE International Symposium on High Performance Distributed Computing (HPDC).

[2] Pathan A.M.K and Buyya R, 2007, "Economy-Based Content Replication for Peering CDNs", TCSC Doctoral Symposium, In Proc. of the 7th IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2007).

[3] Buyya R, Pathan A.M.K, Broberg J and Tari Z, 2006, " A Case for Peering of Content Delivery Networks", IEEE Distributed Systems Online, 7(10).

[4] Pathan A.M.K and Buyya R, 2007," A Taxonomy and Survey of CDNs", Technical Report, GRIDS-TR-2007-4.

[5] Crovella M.E and Bestavros A, 1997, "Self-similarity in World Wide Web traffic: Evidence and possible causes", IEEE/ACM Transactions on Networking, 5(6), pp.835- 846.

[6] Crovella M. E, Taqqu M.S and Bestavros A, 1998, "Heavy- Tailed Probability Distributions in the World Wide Web- A Practical Guide To Heavy Tails, Birkhauser Boston Inc.

[7] Cobham A, 1954, "Priority Assignment in Waiting Line Problems", Journal of the Operations Research Society of America, Vol. 2, No.1, pp. 70-76.

[8] Conway R. W, Maxwell W. L and Miller L. W, 1967, "Theory of Scheduling", Addison-Wesley (Reading, Mass).

[9] Broberg J, Zeephongsekul P, and Tari Z, 2007, "Approximating Bounded General Service Distributions", In Proc of IEEE Symposium on Computers and Communications.

[10] Bouman J, Trienekens J and Zwan M, 1999, "Specification of Service Level Agreements, Clarifying Concepts on the Basis of Practical Research", In Proc. of the Software Technology and Engineering Practice Conference, pp. 169.

[11] Kleinrock L, 1975, "Queuing Systems", Vol. 2, Computer Applications, John Wiley & Sons, pp. 15-19.

[12] Padmanabhan V. N and Sripanidkulchai K, 2002, " The Case for Cooperative Networking", In Proc. of International Peer-To-Peer Workshop (IPTPS02).

Manivannan .K, Lecturer in the Department of Computer Science Engineering at PSNA College of Engineering & Technology. His research interests are Distributed Systems, Middleware technologies, Software Engineering methodologies and Network architecure.