

On the Effectiveness of Content Distribution Networks

Seraphin B. Calo, Dinesh C. Verma, James Giles and Dakshi Agrawal

IBM T J Watson Research Center

PO Box 704, Yorktown Heights, NY 10598

Abstract--Content Distribution has emerged as a preferred way to improve the client perceived response time of applications and services provided over the Internet. In this paper, we analyze the reduction in response time that can be obtained by deploying a content distribution network solution. We study the impact of caching hit ratios on content distribution, and its interaction with network latencies. We then examine the implication of the size of the content distribution network to study the relative merits of two approaches towards building such networks: one using a small numbers of powerful sites; and, the other using a large number of less powerful sites. We also analyze the optimal distribution of processing capacity between an origin site and surrogate sites in a content distribution environment.

Index terms-- Content Distribution Networks, Analysis, Performance Evaluation

A. INTRODUCTION.

Content distribution is the technique of improving user perceived response time by deploying a number of surrogate servers that are geographically distributed throughout the Internet. Each surrogate acts like a caching surrogate server that can provide locally cached content to its clients. A typical content distribution network (CDN) consists of one origin site, and many surrogate servers that are distributed throughout the network. A client is directed to the surrogate site that is close to it for some of the content that can be cached or otherwise provided from the surrogate site. Since the connection from the client to the surrogate is likely to have a better performance than the connection between the client and the origin server, the perceived performance by the client improves in a content distribution network. A variety of schemes [1] are used to determine and direct a client to the closest surrogate.

The amount of content that can be serviced from the surrogate site varies depending on the specific vendor of the CDN software provider, or the CDN service provider. Most providers of CDN services presently confine themselves to providing static images from the surrogates [2] [3]. However, there is effort underway to standardize limited amounts of processing at the surrogate sites [4]. Software for supporting CDN functions offered by some companies includes capability to offload general application processing to the surrogate sites [5]. CDN service providers also have differing philosophies on the best approach to providing

a content distribution service. The leading player in the arena [2] has a model of having a large number of servers deployed throughout the Internet, while others [6] have a small number of high capacity surrogates within the Internet. It is unclear which of them would be a more effective solution.

Given the rise in prominence of CDN sites, several researchers have tried to study their performance. Most of the studies have been done by performing empirical measurements of the effectiveness of the different CDN providers over the Internet [7] [8]. While empirical measurements are valuable, they lack the generality and rigor of an analytical approach. Empirical measurements and conclusions based upon them are also subject to becoming obsolete due to the changing nature of the Internet. In this paper, we propose a tractable model to analyze the theoretical effectiveness of the CDN approach. While the tractability is obtained by making some simplifying assumptions, the results will be true in all environments where our model provides a reasonable approximation. The analytical approach can also help compare alternative approaches - building a CDN with many small sites, versus a CDN with a few more powerful sites. Our work in this context is similar in nature to [9] but we use a more detailed model for server processing time in this paper. This approach allows us to determine the optimal network characteristics in a more precise fashion.

In the next section (Section B), we present the model for analyzing CDN performance that we will be using in the rest of the paper. In Section C, we study the impact of the offloaded content on client-perceived response time. In Section D, we also examine the number of surrogate sites that are optimal for content distribution, and in Section E, we study how to best divide processing capacity among the surrogates and the origin server. Finally, we summarize our findings in Section F.

B. THE CDN MODEL.

In order to analyze its performance, we model a CDN as consisting of an origin server at a central location with a set of n clients distributed symmetrically in a circle around it (See Figure 1). Each client generates requests at rate λ , while the server has a processing rate

of μ_s . The network latency between the client and the server is assumed to be a fixed value of τ_0 . The network latency is the effective round trip time that we would expect for a TCP connection between a client and the server. We assume that each client uses a surrogate server that is located at a network latency of τ_p from the client. The network latency between the client and origin server should be greater than that between the client and its surrogate. We model this effect by defining τ_0 to be equal to $\tau_p + \tau_d$, where τ_d is a positive quantity. We assume that each surrogate has the processing rate of μ_p . All arrival and service processes are assumed to be Poisson.

The system with n surrogates therefore consists of $n+1$ M/M/1 queues with one queue representing the origin server at the center of the circle, n queues distributed symmetrically around a circle of radius τ_d centered at the origin server, and n clients generating load distributed symmetrically around a circle of radius $\tau_p + \tau_d$. Each client is assumed to be at the linear distance of τ_p from a surrogate. The total processing capacity μ_0 in the system is $\mu_s + n\mu_p$, and the total load generated by all the clients, λ_0 , equals $n\lambda$.

While the symmetrical system is a simplified model of the Internet, we believe that it can be used as a fairly accurate approximation as far as studying the average response times are concerned. If we consider the client clouds in Figure 1 to be IP subnets generating client load requests at an average rate of λ per subnet, τ_p as the average latency between a client and the surrogate, and $\tau_p + \tau_d$ as the average latency between a client and the origin server, the simple model can be seen to be a good match for a real CDN.

In order to obtain a quantitative measure of the user response time, we use the linear model proposed in [9]. The user response time is a combination of two terms, one reflecting the contribution due to network latency and the other reflecting the server processing time. For a network client generating requests at rate λ that is accessing a server with rate μ located at a network latency of τ , the contribution of network latency to user response time is $N\tau$ where N is a scaling factor that incorporates the effect of network loss rates and network re-transmissions. The average time spent by a request at a M/M/1 server is $1/\mu - \lambda$. Thus, the average user response time is given by the expression:

$$R = N\tau + 1/\mu - \lambda,$$

where $\lambda < \mu$ is the stability condition for the server.

C. ANALYZING THE IMPACT OF CONTENT OFFLOAD.

We assume that the content distribution network operates by offloading a fraction p of its requests to the surrogate sites. Since we are focusing primarily on content that is offloadable, we assume that p can be between 0 and 1. We further assume that the CDN routing scheme will direct the offloadable requests to the closest surrogate site, and the non-offloadable requests directly to the server. The caches at the surrogate sites are assumed to be pre-loaded so that the surrogate is always able to service the requests that it receives. While some surrogates fetch the content on the first access, we assume that the impact of the initial miss on the load or service rate at the surrogates and the origin server is negligible. Since most surrogates operate in close collaboration with the origin server, these assumptions are fairly realistic.

The communication between the client and the surrogate can be modeled as an M/M/1 queue with an arrival rate of $p\lambda$, a service rate of μ_p , and a network latency between the client and server of τ_p . Each such request will have an average response time of $N\tau_p + 1/(\mu_p - p\lambda)$. The communication between the client and the origin server (for the non-offloadable components) can be modeled as an M/M/1 server with an arrival rate of $(1-p)\lambda$, a service rate of μ_s , and a network latency of $(\tau_p + \tau_d)$. Each such request will have an average response time of $N(\tau_p + \tau_d) + 1/[\mu_s - n(1-p)\lambda]$.

Given that a fraction p of requests goes to the surrogate, and the other fraction goes to the origin server, the overall response time R can be obtained as

$$R = p\left\{N\tau_p + \frac{1}{\mu_p - p\lambda}\right\} + (1-p)\left\{N(\tau_p + \tau_d) + \frac{1}{\mu_s - n(1-p)\lambda}\right\},$$

which after some algebraic manipulation can be shown to be equivalent to

$$R = N\tau_p + (1-p)N\tau_d + \frac{p}{\mu_p - p\lambda} + \frac{1-p}{\mu_s - n(1-p)\lambda}.$$

The constraints for the existence of this solution are that $p\lambda < \mu_p$ and that $n(1-p)\lambda < \mu_s$.

When we consider a solution that is totally centralized without using any content distribution, p is 0, and we get that the response time for no distribution, $R(0)$, is

$$R(0) = N(\tau_p + \tau_d) + \frac{1}{\mu_s - n\lambda}.$$

For an approach which offloads all of the content and does not have any processing at the origin server (full distribution), p is 1 and we have $R(1) = N\tau_p + \frac{1}{\mu_p - \lambda}$.

Let us now consider the conditions under which content distribution will result in a performance gain. By differentiating the response time with respect to p , we see that R' , the first derivative of R with respect to p , is $\frac{dR}{dp} = R' = -N\tau_d + \frac{\mu_p}{(\mu_p - p\lambda)^2} - \frac{\mu_s}{(\mu_s - n(1-p)\lambda)^2}$.

The second derivative with respect to p , R'' , is given by

$$R'' = \frac{2\lambda\mu_p}{(\mu_p - p\lambda)^3} + \frac{2n\lambda\mu_s}{(\mu_s - n(1-p)\lambda)^3}.$$

When the existence conditions are satisfied, the second derivative is always positive. Thus the first derivative R' is a monotonically increasing function of p . If we now look at $R'(p)$, exactly one of the following three mutually exclusive conditions must be true: (i) $R'(0) < R'(1) < 0$, (ii) $R'(0) \leq 0 \leq R'(1)$ or (iii) $0 < R'(0) < R'(1)$. Let us consider each of the above conditions in turn.

If the first derivative is negative for the case of no distribution ($p = 0$) and for the case of full distribution ($p = 1$), then the lowest value of response time is obtained at full distribution.

If the first derivative is negative for the case of no distribution ($p = 0$), but positive for the case of full distribution ($p = 1$), then the first derivative must be zero at exactly one value of p between 0 and 1. Since the second derivative is always positive, the value of p where $R'(p)$ becomes 0 is the point of optimum response time.

If the first derivative is positive for the case of no distribution, then response time increases with distribution and the best solution is to have processing concentrated at a single server.

We introduce a simplifying parameter $\Gamma = N\mu_p\tau_d$, and evaluate the expression of the first derivative to obtain

$$R'(0) = \left\{ 1 - \Gamma - \frac{\mu_p/\mu_s}{(1-n\lambda/\mu_s)^2} \right\} / \mu_p,$$

and,

$$R'(1) = \left\{ \frac{1}{(1-\lambda/\mu_p)^2} - \mu_p/\mu_s - \Gamma \right\} / \mu_p.$$

Thus, we can summarize our conclusions as the

following:

If $\Gamma \leq 1 - \frac{\mu_p/\mu_s}{(1-n\lambda/\mu_s)^2}$, then no distribution is the best solution.

If $1 - \frac{\mu_p/\mu_s}{(1-n\lambda/\mu_s)^2} < \Gamma < \frac{1}{(1-\lambda/\mu_p)^2} - \mu_p/\mu_s$, there is an optimum distribution level between 0 and 1.

If $\Gamma \geq \frac{1}{(1-\lambda/\mu_p)^2} - \mu_p/\mu_s$, then full distribution is the best solution.

Figure 2 shows three scenarios that can arise with different values of the Γ parameter. It shows the response time for a CDN under three different cases: one in which the response time increases with the offload percentage; another in which the response time decreases with offload percentage; and, a third in which there is an optimum percentage of data that ought to be offloaded. If the goal is to have content distribution improve the response time of the clients, then having a higher value of the Γ parameter will be helpful.

D. OPTIMUM SIZE OF A CDN.

There are two competing approaches towards building a CDN. Some content distribution providers [2] have opted to build a content distribution service incorporating thousands of sites distributed throughout the network. Other content distribution companies [6] have opted to build a few tens of more powerful sites. Since we have an expression describing the response time of the users of a CDN, it would be interesting to determine which of the two approaches is supported by our analysis.

In order to compare the two approaches, we assume that the total processing rate among all the surrogate servers is fixed. Thus, the processing rate of each surrogate server when the CDN system contains n surrogates is given by $\mu_p = \frac{\mu_o}{n}$, where μ_o is a constant. This models the fact that one can establish many weak sites or a small number of powerful sites. We also assume that the total load generated by all the clients together is a constant given by λ_o . Thus, the request arrival rate at each surrogate is $p\lambda_o/n$.

As a first step, let us assume that the value of τ_p is not affected by the number of surrogate servers deployed within the CDN. The response time expression is

$$R = N\tau_p + (1-p)N\tau_d + \frac{p}{\mu_p - p\lambda} + \frac{1-p}{\mu_s - n(1-p)\lambda},$$

in which substituting the values of μ_p and λ_p yields the simplified equation,

$$R = N\tau_p + (1-p)N\tau_d + \frac{np}{\mu_0 - p\lambda_0} + \frac{1-p}{\mu_s - (1-p)\lambda_0}.$$

Since N , τ_p , τ_d , μ_p and p are constant, the response time is linear in the number of surrogates!

The result implies that it is better to have a small number of powerful surrogates than to have a large number of weaker surrogates. This is because weaker surrogates result in larger processing delays than more powerful surrogates, and dividing the processing capability hinders, rather than helps, to reduce the client response time.

A corollary of the above result is that smart content distribution design should concentrate on locating surrogates so that the average latency between the clients and surrogates is reduced. The costs involved in placing additional surrogates in the network must be compared with the additional reduction in latency that results. Placing a larger number of surrogates in the network should result in reducing the average network latency between the clients and servers on the network, and it will have implications in determining the right size for a content distribution network.

Let us first note that the average distance between the client and the origin server remains unchanged by the placement of extra surrogates. In other words, the relation that $\tau_p + \tau_d = \tau_0$ where τ_0 is a constant remains valid. Let us further assume that n is large, and thus it would be reasonable to assume that n is a continuous variable rather than a discrete integer. In that case, to determine the optimum value of n , we can take derivatives with respect to n , and determine the maxima points. We can rewrite the response time equation and take derivatives with respect to n to obtain:

$$R = pN\tau_p + (1-p)N\tau_0 + \frac{np}{\mu_0 - p\lambda_0} + \frac{1-p}{\mu_s - (1-p)\lambda_0}$$

$$\frac{dR}{dn} = R'(n) = pN\tau_p' + \frac{p}{\mu_0 - p\lambda_0}$$

$$\frac{d^2R}{dn^2} = R''(n) = pN\tau_p''.$$

Let us assume that $\tau_p = Cn^{-\alpha}$, where n is the number of surrogates, and C as well as α are positive constants. Such a distribution would be consistent with the power law relationships exhibited within the Internet [10]. For some distributions of surrogates, it can be shown that α would be 1 or 2. (See Appendix 1).

With this expression for τ_p , we have that $\tau_p' = -(a/n)\tau_p$ and $\tau_p'' = \tau_p a(a+1)/n^2$. Note that the second derivative of τ_p is always positive. This implies that the response time is minimized when the first derivative is 0. Solving for that condition provides the

optimum number of surrogates as:

$$n_{opt} = ((\mu_0 - p\lambda_0)aCN)^{1/(1+a)}.$$

At the optimum point, the first derivative of R is zero.

This implies that:

$$N(a/n)\tau_p = \frac{1}{\mu_0 - p\lambda_0}.$$

Therefore, the optimum response time would be obtained as:

$$R_{opt} = \frac{n_{opt}(1+a)p}{a(\mu_0 - p\lambda_0)} + (1-p)N\tau_0 + \frac{1-p}{\mu_s - (1-p)\lambda_0}.$$

Let us first make the observation that the value $1/(\mu_0 - p\lambda_0)$ reflects the average time taken to service a request at a surrogate site if the entire load were sent to a single server with the combined processing capacity. For a real Internet based application, this would translate to a response time somewhere between 1 and 20 ms. The second observation is that the parameter C reflects the network latency between the client and surrogate when there is only one surrogate, and approximately equals the network latency between the clients and the server. This average latency ranges from 100 ms for unloaded enterprise intranets to 500 ms for the congested Internet. The parameter N approximates the average number of round trip exchanges needed for a single transactions, and would be between 5-10 on typical web transactions.

Figure 3 shows the optimum number of surrogates for some typical values of α that can be expected for enterprise intranet ($C=100$) environments. The surrogate in one case is assumed to be distributed randomly around the circle ($\alpha = 1$, see Appendix), located randomly along the bisector of a circular segment ($\alpha = 2$), and at an intermediary value of $\alpha = 1.6$. Figure 4 shows the same numbers for a congested Internet, with C equaling 500 ms. The constant N is assumed to be 7 in both of the graphs. As can be seen by the graphs, usually a reasonably small number of surrogates will be optimal in a variety of cases. The optimum number of surrogates in most cases is less than 100.

E. OPTIMUM CAPACITY DISTRIBUTION.

The analysis of the previous section assumed a fixed processing rate at the origin server, which is a valid assumption when we consider the case of a corporation with existing servers utilizing the services of a third party content distribution provider. When content distribution is deployed within an enterprise intranet, there is an option to trade-off the processing capacity at the origin server versus the processing capacity at the surrogate servers. In this section, we consider the situation where a fixed processing capacity is available and can be distributed among the surrogate servers or at

the origin server. Thus, we have the constraint that $n\mu_p + \mu_s = \mu_0$ where μ_0 is a constant and n is the number of surrogate servers. We also assume that the total load generated by all the clients together is a constant, λ_0 . Thus, the request arrival rate at each surrogate is $p\lambda_0/n$.

For the purpose of analyzing the distribution of capacities between the origin servers and the surrogate servers, the percentage of an application that can be offloaded to the surrogate servers is determined by the type of application, and thus p is considered to be a fixed quantity. We also assume that the number of surrogate sites is a fixed constant, and thus the only question is that of determining the right allocation between the processing capacity at the origin server and at the surrogate sites.

In the response time equation, we express the surrogate service rate as a function of the origin server service rate. Since $\mu_p = (\mu_0 - \mu_s)/n$, we get that

$$R = N\tau_p + (1-p)N\tau_d + \frac{n}{\mu_0 - \mu_s - p\lambda_0} + \frac{1-p}{\mu_s - (1-p)\lambda_0}.$$

The constraints for the existence of this solution are that $p\lambda_0 < \mu_0 - \mu_s$ and $(1-p)\lambda_0 < \mu_s$. These constraints can be rewritten as $\mu_s < \mu_0 - p\lambda_0$ and $(1-p)\lambda_0 < \mu_s$. It follows that for a feasible solution to exist, $\lambda_0 - p\lambda_0$ must be less than $\mu_0 - p\lambda_0$, which will be satisfied provided that $\lambda_0 < \mu_0$.

Since the number of surrogates is fixed, the values of τ_p and τ_d are also constant. Taking the derivative of the response time with respect to μ_s , we get that

$$\frac{dR}{d\mu_s} = \frac{np}{(\mu_0 - \mu_s - p\lambda_0)^2} - \frac{1-p}{(\mu_s - (1-p)\lambda_0)^2}.$$

The second derivative of the response time with respect to μ_s is given by the expression:

$$\frac{d^2R}{d\mu_s^2} = \frac{2np}{(\mu_0 - \mu_s - p\lambda_0)^3} + \frac{2(1-p)}{(\mu_s - (1-p)\lambda_0)^3}.$$

For the feasible values of μ_s between the values of $\lambda_0 - p\lambda_0$ and $\mu_0 - p\lambda_0$, the second derivative is always positive. Let us examine the behavior of the first derivative at the two extreme values of this range. From the expression of the first derivative, we can see that the limiting value of $\frac{dR}{d\mu_s}$ as $\mu_s \rightarrow (\lambda_0 - p\lambda_0)$ is $-\infty$, and the limiting value of $\frac{dR}{d\mu_s}$ as $\mu_s \rightarrow (\mu_0 - p\lambda_0)$ is $+\infty$. This coupled with the fact that the second derivative is always positive implies that the first derivative is 0 at exactly one point in this space, and that is the point where the response time is minimized. The optimum response time will be obtained when the processing rate

at the origin site is given by:

$$\mu_s = \{\mu_0 - \lambda_0(\beta p + p - \beta)\}/(1 + \beta),$$

where $\beta = \sqrt{\frac{1-p}{np}}$.

It follows that the primary determination of the optimal processing capacity between the servers and surrogates is determined by the fraction of traffic that is offloaded, and the number of surrogates involved in the interchange. When the number of surrogates is relatively large, the limit of β as $n \rightarrow \infty$ is 0. Thus, the limiting value of the optimal processing capacity of the origin server is $\mu_0 - p\lambda_0$. This is the smallest possible capacity at the origin server that is feasible under the stability constraints. In other words, the optimum solution in the limiting case would be to distribute as much of the processing capacity as possible.

F. CONCLUSION AND FURTHER WORK.

In this paper, we have analyzed the performance aspects of content distribution networks, and have shown the following results:

For networks where the Γ parameter is less than unity, a centralized solution may be the better option in many circumstances.

For networks where the Γ parameter is very large, full distribution is the better solution. For networks with an intermediate values of the Γ parameter, an optimum distribution point exists.

Having extra surrogate servers in a content distribution network does not help to reduce user response time if the distance between the clients and surrogates does not decrease as a result.

When designing a CDN with a fixed amount of total surrogate processing rate, a smaller number of powerful surrogates provides a better solution than many weaker surrogates. For parameters approximating Internet characteristics, the optimum number of surrogates is typically less than 100 depending on network environment.

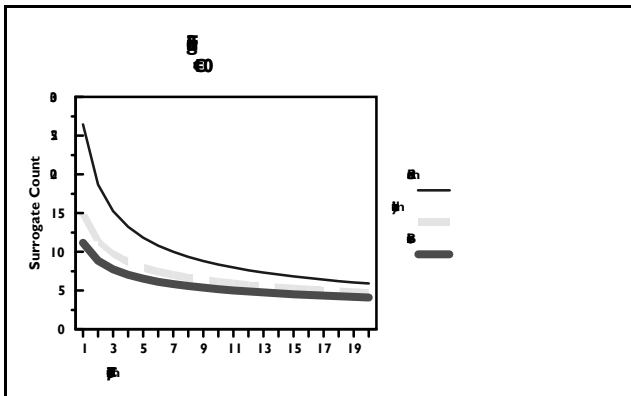
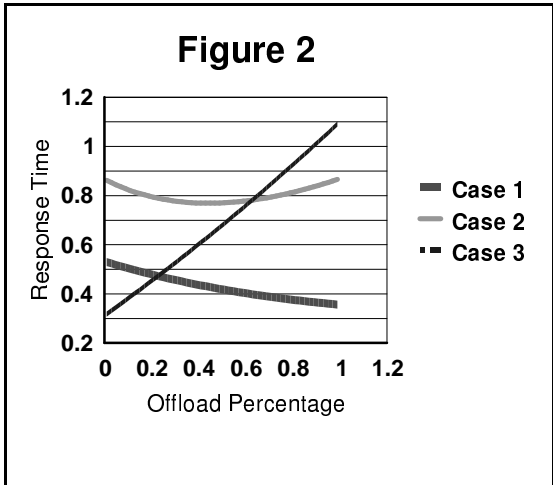
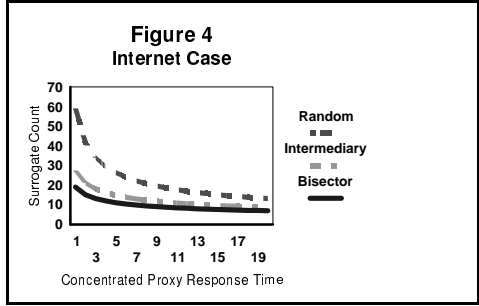
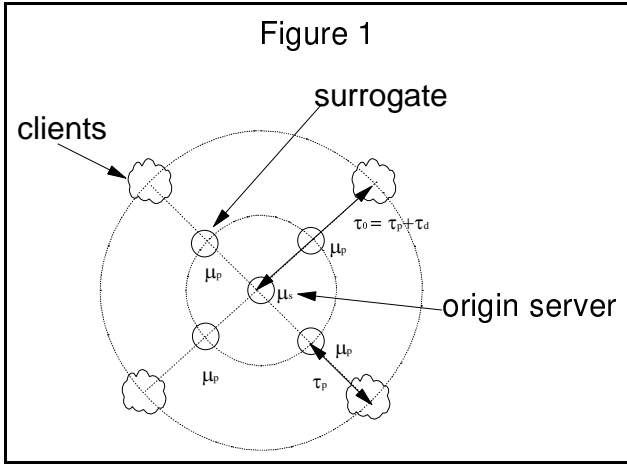
When designing a CDN with a fixed amount of total surrogate and origin server processing rate, the best allocation of processing power is determined by p , the fraction of requests offloadable to the surrogates. In the limiting case of large numbers of surrogates, the best approach is to distribute as much processing power as possible to the surrogate sites.

As future enhancements to this analysis, we intend to

determine the appropriate value of the α parameter in Internet like topologies, and to develop the model further to account for the overhead involved in content distribution, as well as asymmetrical load generation from different client clusters.

G.REFERENCES

- [1] P. S. M. Sayal and P. Vingralek. *Selection algorithms for replicated web servers*. In The 1998 SIGMETRICS/Performance Workshop on Internet Server Performance, June 1998.
- [2] Akamai Technologies Inc., *FreeFlow content distribution service*, www.akamai.com.
- [3] Digital Island Inc., *Footprint content distribution service*, Description available at URL <http://www.digitalisland.net/services/cd/footprint.shtml>.
- [4] Edge Side Includes Overview, Available at URL <http://www.edge-delivery.org/overview.html>.
- [5] IBM Corp., *IBM WebSphere Edge Services Architecture, - Application Offload capability*, <http://www-4.ibm.com/software/webservers/edgeserver/doc/esarchitecture.pdf>
- [6] Mirror Image, *Content Access Point (CAP) Network*, <http://www.mirror-image.com/cap/overview.html>.
- [7] Kirk L. Johnson, John F. Carr, Mark S. Day, and M. Frans Kaashoek. *The measured performance of content distribution networks*. In *Proceedings of the 5th International Web Caching and Content Delivery Workshop*, May 2000.
- [8] A. Shaikh, R. Tewari, and M. Agrawal, *On the Effectiveness of DNS-based Server Selection*, *Proc. IEEE INFOCOM 2001*, April 2001.
- [9]. D. Agrawal, J. Giles and D. Verma, *On the Performance of Content Delivery Networks*, *Proceedings of SPECTS 2001*, Orlando, FL, July 2001.
- [10] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "*On Power-Law Relationships of the Internet Topology*," *Proc. SIGCOMM '99*, Aug. 1999.



Appendix 1: Network Latency & the Number of Surrogates

Distributing a larger number of surrogates in the network should result in reducing the average network latency for the clients. We propose a simple model to get a feel for the type of relationship that would exist between these two variables, i.e., network latency as a function of the number of surrogate servers. We assume that the origin server is at the center of a circle of radius τ_0 , and the clients are distributed along the perimeter of this circle. If we divide the circle into n segments, then each segment would subtend an angle of $2\pi/n$ radians. We then place a surrogate server in each segment along the bisector of this angle at a distance of r_p from the center of the circle. From simple geometry and the Law of Cosines, we get that the distance, D , from any point on the perimeter of the circle within a given segment to the surrogate server in that segment must satisfy the inequality:

$$D \leq [r_p^2 + \tau_0^2 - 2\tau_0 r_p \cos(\pi/n)]^{1/2}.$$

If there are a large number of surrogates, $n \gg 1$, the cosine function can be approximated by the first two terms of its Taylor Series expansion and we get that:

$$D \sim [(\tau_0 - r_p)^2 + \tau_0 r_p (\pi/n)^2]^{1/2}.$$

If the surrogates are close to the clients, so that r_p has a value close to τ_0 , then the distance from any client to the surrogate server in its segment would be approximated by:

$$D \sim \tau_0 \pi/n.$$

We note that such a relationship (inversely proportional to n) would also arise if the client were located at a distance τ_0 from the server, surrogates were uniformly randomly placed in between them, and we calculated the expected distance of the client from the closest server. While this linear model may appear to be overly simplified, it captures circumstances in which the clients are clustered and the surrogates are placed advantageously close to them, so that the placement of each surrogate has a significant affect on each client. Indeed, the same results would pertain for a particular client if all the surrogates were placed randomly within a circle of radius τ_0 around the client itself.

More generally, if we take the original inequality, and

assume that the surrogate is now randomly placed on the bisector of its segment with a uniform probability distribution that lies between the values τ_1 and τ_2 , we can get an expression for the expected value of the distance bound, but it is somewhat complex. Assuming again that $n \gg 1$, the expression simplifies to the one below.

$$D \sim \tau_0 - (\tau_1 + \tau_2)/2 + \tau_0^2 (\pi/n)^2 \log[(\tau_0 - \tau_1)/(\tau_0 - \tau_2)]/2(\tau_2 - \tau_1)$$

If we pick values for τ_1 and τ_2 (e.g., $\tau_1 = \tau_0/k$ and $\tau_2 = \tau_0(k-1)/k$ for some integer k), then we get that

$$D \sim \tau_0/2 [1 + \beta(\pi/n)^2],$$

where the constant β is determined solely by the parameters of the distribution. This inverse square relationship seems to capture situations in which clients are spaced sufficiently far apart that they are affected only by the locations of a relatively small number of surrogates.

The above analysis leads us to conclude that $\tau_p = Cn^{-\alpha}$, where n is the number of surrogates, and C as well as α are positive constants. It also shows that reasonable values for α are 1 or 2, depending upon the geographic clustering of the clients.