# Bandwidth versus storage trade-off in a content distribution network and a single server system

Stijn Van Rompaey, Kathleen Spaey, Chris Blondia
University of Antwerp, Department of Mathematics and Computer Science
Performance Analysis of Telecommunication Systems Research Group
Middelheimlaan 1, B-2020 Antwerpen, Belgium
Email: {stijn.vanrompaey, kathleen.spaey, chris.blondia}@ua.ac.be

***Abstract*— CDN architectures are recently being deployed to lower the client-perceived response time of clients retrieving information from the Internet. This paper addresses the question under which conditions a CDN system or an increase in capacity of the already installed single server system could be used for this task. The trade-off between storage (CDN system) and bandwidth (upgraded single server system) is investigated by comparing the average client-perceived response times of both solutions. The influence of the different parameters that specify each system is also examined in more detail.**

***Index Terms*— Content distribution network (CDN), single server system, bandwidth versus storage, response time, performance, modeling.**

## I. Introduction

Despite the faster access of end users and the increasing Internet backbone capacity, end users often encounter poor quality when retrieving content from the Internet. Therefore, the concept of Content Distribution Networks (CDNs) [1], [2], [3] has recently gained a lot of attention. In a CDN, popular content that initially resides at the origin server is replicated to so-called surrogate servers at the edges of the Internet, and clients requesting that content are served by such a surrogate server. Since the connection from the client to the surrogate server is likely to have a better performance than that between the client and the origin server, the client-perceived response time will likely be better in a CDN than in the classical client-server architecture.

The effectiveness of a CDN depends on many factors such as the type of the hosted application, the cache-hit ratio, the average round trip delay between clients and servers, and the architecture of the CDN. Using a linear model for the client-perceived response time, Agrawal et al. [4] analytically compare the performance of a CDN architecture with that of the classical client-server architecture and study the effect of some of these factors.

A simple CDN model considered in [4] consists of a client, an origin server and a surrogate server. If possible, the surrogate server serves client requests. In case the surrogate is unable to satisfy a request, it redirects the client to the origin server. The client-server model considered is similar to the CDN model, but now without a surrogate server in place, such that the origin server always serves a client request. The round trip delay between the client and the origin server is identical in both models. The focus in [4] is on the comparison between the performance of the single server system and the better performance of the CDN system.

In this paper we will consider a similar CDN and client-server model as in [4], and also compare the client-perceived response time obtained in both systems. However, our starting point is different: assume that for a given client-server model the client-perceived response time is considered to be poor. Then two solutions that will improve the response time are (i) to upgrade the capacity of the bottleneck links and network elements between client and server, or (ii) to transform the system into a CDN, i.e., to introduce a surrogate server closer to the client that keeps duplicates of part of the content, such that part of the requests can be served by the surrogate server. So we focus on the bandwidth versus storage trade-off by comparing the performance of an 'upgraded' single server system (bandwidth) to that of a CDN system (storage).

The paper is organized as follows. First the CDN and single server model of [4] and its parameters and assumptions are briefly summarized in Section II. In Section III, the model is slightly changed and a new parameter, the round trip delay ratio is introduced, to study the bandwidth versus storage trade-off. Next, this trade-off and the influence of surrogate placement, round trip delay ratio, cache-hit ratio and the type of the hosted application is analyzed and illustrated by numerical examples in Section IV. Finally, conclusions are given in Section V.

## II. CDN and single server model

### A. A linear model for web transactions

To quantify the average client-perceived response time for the download time of a web page, a linear model is used:

$$T = N\tau + P, \tag{1}$$

where the variable $\tau$ denotes the average round trip delay (RTD) between client and server, and $P$ and $N$ are constants. The value of $N$ is strongly correlated with the amount of data

This work was carried out within the framework of the project CoDiNet sponsored by the Flemish Institute for the promotion of Scientific and Technological Research in the Industry (IWT).

to be transferred. The processing time needed by both server and client is denoted by $P$. So the average client-perceived response time is modeled by the sum of the data transfer time ($N\tau$) and the processing time ($P$).

Experimental validation of this linear model was obtained in [5]. In this paper the average client-perceived response time is measured by downloading different popular web pages using a WAN emulator. The linear model is also used in [6] where other CDN issues, like the optimal number of surrogate servers and the optimal processing capacity distribution, are investigated.

### B. Model and assumptions

The CDN model considered throughout this article consists of a client, an origin server and a surrogate server. To retrieve the data, the client contacts the surrogate server. Subsequently there are two possibilities: either the data is cached at the surrogate server and the surrogate server responds by transmitting this data to the client, or the surrogate server is unable to satisfy the request and redirects the client to the origin server that stores the data by definition.

Let $f$ denote the cache-hit ratio, i.e., the fraction of the total number of requests that is satisfied by the surrogate server. Assuming that the overhead of the redirection transaction is insignificantly small compared to the other delays and can consequently be omitted, the average client-perceived response time of the CDN system is given by

$$T_{CDN} = f(N_{cs}^{hit}\tau_{cs} + P_{cs}^{hit}) + (1-f)(N_{co}\tau_{co} + P_{co}). \quad (2)$$

Notice that the indices denote which transaction is modeled: *cs* refers to a transaction between client and surrogate server and *co* refers to a transaction between client and origin server.

The average client-perceived response time of a transaction in a classical single server model is given by

$$T_S = N_{co}\tau_{co} + P_{co}. \quad (3)$$

To simplify Equation (2), some assumptions are made. First it is assumed that the surrogate server sends the same data and needs the same number of round trips to serve the client as the origin server. So $N_{cs}^{hit} = N_{co}$. Also the processing time incurred by the transaction between the client and the surrogate server on a cache-hit is assumed to be equal to the processing time of a transaction between the client and the origin server: $P_{cs}^{hit} = P_{co}$. Because in a CDN the surrogate server is positioned closer to the client than the origin server, the RTD between the client and the surrogate server is typically smaller than that between client and origin server. This is modeled by assuming that $\tau_{cs} = \alpha_\tau \tau_{co}$, where $0 < \alpha_\tau \leq 1$.

In [3] a standard web page is downloaded from numerous surrogate and origin servers and the corresponding DNS and download times are recorded. These measurements give some indication about values of $\alpha_\tau$ in the real world. Most $\alpha_\tau$ values are situated in the interval [0.08 (Speedera), 0.98 (Fasttide)] when the median measurements (Jan. 2001) are used. Only Adero manages to exceed this interval with $\alpha_\tau$ equal to 1.07, which is couterintuitive.

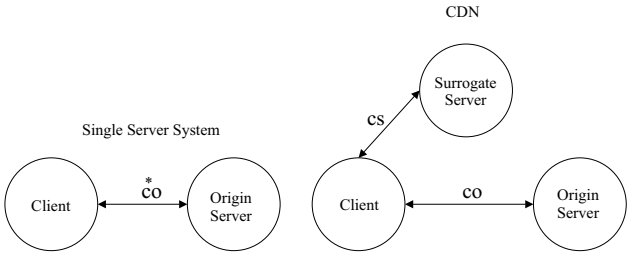


Fig. 1. The different RTDs between servers and clients.

### C. Client-perceived response time ratio

With the assumptions made in section II-B, the ratio of the client-perceived response times of the CDN and the single server system is given by

$$R = \frac{T_{CDN}}{T_S} = \frac{((1-f) + f\alpha_\tau)\Gamma + 1}{\Gamma + 1}, \quad (4)$$

where $\Gamma = N_{co}\tau_{co}/P_{co}$. Hence, $\Gamma$ relates the transfer time of an application to its processing time and will be called the critical ratio. Notice that under the assumption of a negligible redirection time, $R$ is never larger than one.

## III. Trade-off between storage and bandwidth

### A. The model

From the previous section one could conclude that the usage of a CDN will always result in a reduction of the average client-perceived response time. Now, an alternative to the deployment of a CDN will be explored in order to decrease the average client-perceived response time. This solution consists of an upgrade of the capacity of the bottleneck links and network elements between client and server. The result will be a single server system with a reduced RTD $\tau_{co}^*$ between client and server, indicated by $\tau_{co}^* = \beta\tau_{co}$, where $0 < \beta < 1$. Consequently, the average client-perceived response time in an upgraded single server architecture is given by

$$T_S^* = N_{co}\beta\tau_{co} + P_{co}. \quad (5)$$

The variable $\beta$ will be referred to as the *RTD ratio*. A large RTD ratio represents only a modest upgrade of the single server system, whereas a small RTD ratio models a large upgrade. The relevant RTDs for an upgraded single server system and a CDN are illustrated in Fig. 1.

### B. Client-perceived response time ratio

The trade-off between storage and bandwidth can now be incorporated in the client-perceived response time ratio, which is given by

$$R^* = \frac{T_{CDN}}{T_S^*} = \frac{((1-f) + f\alpha_\tau)\Gamma + 1}{\beta\Gamma + 1}. \quad (6)$$

The CDN option is the preferred one if $R^*$ is smaller than 1, or $f(\alpha_\tau - 1) + 1 < \beta$. Otherwise upgrading the single server

system is preferable. Both systems perform equally well when the following condition is satisfied,

$$\beta = f(\alpha_\tau - 1) + 1. \quad (7)$$

Notice the absence of the variable $\Gamma$ in these conditions.

## IV. Results and discussion

The values for $N$ and $P$ that are being used throughout this section are the ones found in [5] for the top-level pages of the CNN website when performing experiments for 0% packet loss using a fast client ($N = 36.0, P = 659$ ms).

### A. Influence of the surrogate placement

The surrogate placement in the CDN system is determined by the parameter $\alpha_\tau$. The smaller $\alpha_\tau$, the closer to the client the surrogate server is placed. The influence of parameter $\alpha_\tau$ is examined more closely in Fig. 2 and Fig. 3.

In Fig. 2 a scenario where the cache hit ratio and the RTD ratio are fixed at 0.8 is considered. The different values of $\Gamma$ are determined by choosing different RTDs (20 ms, 50 ms, 100 ms, 300 ms) from client to origin server in the CDN model with constant values of $N$ and $P$. When the surrogate placement corresponds to $\alpha_\tau = 0.75$, $R^*$ equals 1, or the average client-perceived response time for both the CDN and the upgraded single server system perform equally well. Values of $\alpha_\tau$ smaller than 0.75 make the CDN architecture the preferred choice, while larger values favour the upgraded single-server model. Notice that larger values of $\Gamma$ enlarge the performance difference.

Fig. 3 assumes a fixed RTD of 100 ms between client and origin server in the CDN model, which results in a value for $\Gamma$ equal to 5.4628. Again the cache hit ratio is assumed to be a constant equal to 0.8. Smaller values of $\beta$, representing a larger capacity upgrade of the single server model, result in a smaller average client-perceived response time for this system. As long as the surrogate server is positioned close enough to the client, the CDN model will be able to outperform the upgraded single server model. E.g., when the RTD ratio $\beta$ is fixed at 0.9, a surrogate server location characterized by $\alpha_\tau < 0.875$ makes the CDN model the better option. For a RTD ratio of 0.25, the surrogate server needs to be positioned very close to the client, $\alpha_\tau < 0.0625$, in order for the CDN architecture to perform better than the upgraded single server system.

In accordance with Equation (6) a linear relation between the surrogate placement and the average client-perceived response time ratio is observed.

### B. Influence of the RTD ratio

The RTD ratio $\beta$ represents the capacity improvement of the upgraded single server system. Fig. 4 and Fig. 5 show in more detail the influence of $\beta$ on the client-perceived response time ratio.

In Fig. 4, the surrogate server is positioned halfway between the client and the origin server in terms of the RTD and the cache hit ratio $f$ is fixed at 0.8. The two systems perform equally well when the RTD ratio equals 0.6, which is verified by Equation (7). When the value of $\beta$ is smaller than 0.6, the upgraded single server system outperforms the CDN system. A value of $\beta$ larger than 0.6 favours the CDN system. Notice again that a larger critical ratio enlarges the performance difference between the two systems.

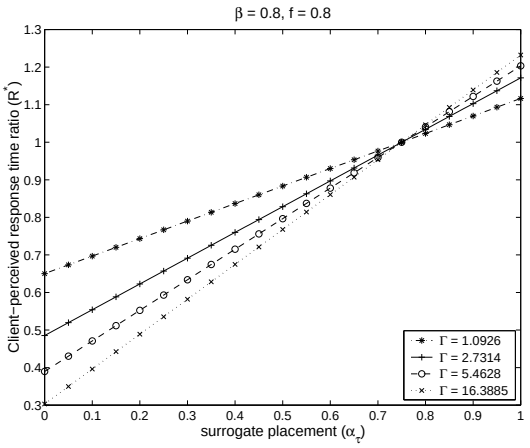


Fig. 2. Influence of the surrogate placement on the client-perceived response time ratio for a fixed $\beta$ and $f$.

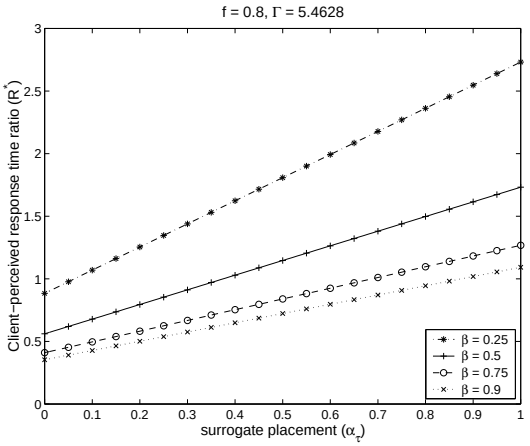


Fig. 3. Influence of the surrogate placement on the client-perceived response time ratio for a fixed $f$ and $\Gamma$.

Fig. 5 considers a similar scenario, but now instead of the cache hit ratio $f$ the critical ratio $\Gamma$ is fixed. Larger values of the cache hit ratio improve the performance of the CDN system. But small values of the RTD ratio $\beta$, indicating a substantial capacity upgrade of the single server system, cannot be compensated for by a high cache hit ratio to make the CDN system the better option. This is because the surrogate server is only located halfway.

### C. Influence of the cache hit ratio

The effect of the cache hit ratio $f$ at the surrogate server is examined in Fig. 6 and Fig. 7.

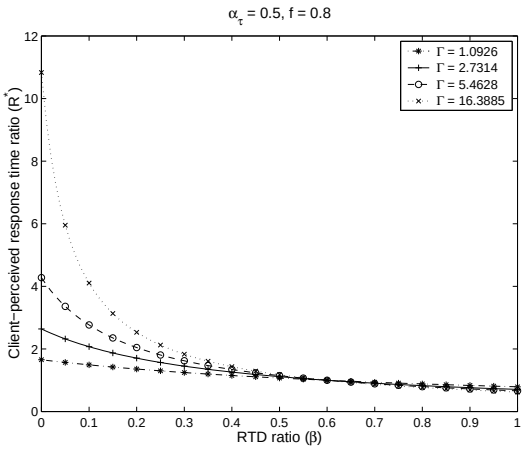


Fig. 4. Influence of the RTD ratio on the client-perceived response time ratio for a fixed $\alpha_\tau$ and $f$.

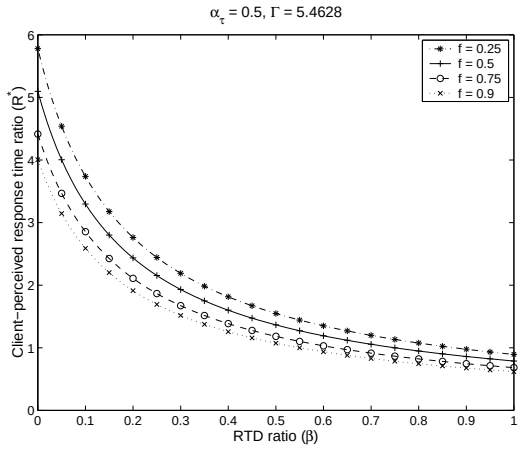


Fig. 5. Influence of the RTD ratio on the client-perceived response time ratio for a fixed $\alpha_\tau$ and $\Gamma$.

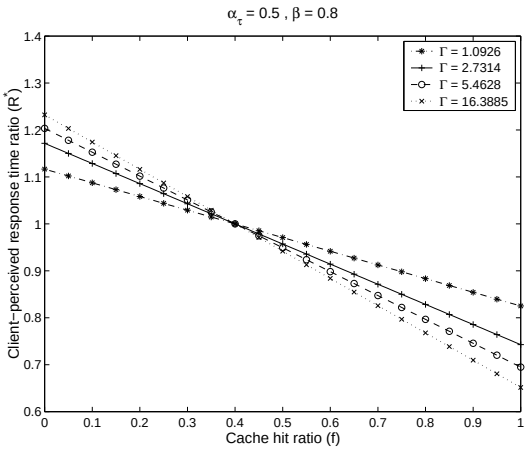


Fig. 6. Influence of the cache hit ratio on the client-perceived response time ratio for a fixed $\alpha_\tau$ and $\beta$.

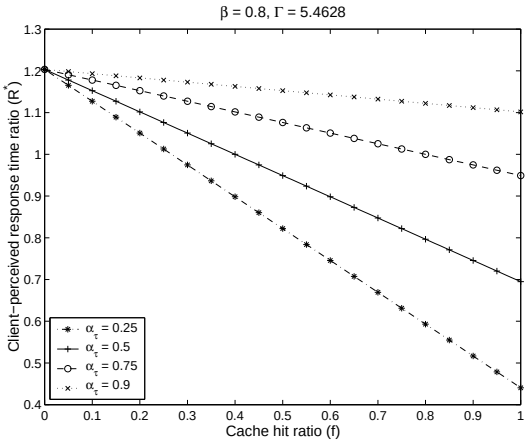


Fig. 7. Influence of the cache hit ratio on the client-perceived response time ratio for a fixed $\beta$ and $\Gamma$.

Fig. 6 describes the scenario where the surrogate server is positioned halfway, $\alpha_\tau = 0.5$, and the single server system is modestly upgraded, $\beta = 0.8$. Both systems perform equally well when the cache hit ratio equals 0.4. For values of $f$ larger than 0.4 the CDN model is the preferred option, while values smaller than 0.4 favour the upgraded single server system.

In Fig. 7, instead of the surrogate placement the critical ratio is assumed to be fixed. The figure shows that for smaller values of $\alpha_\tau$, the positive influence for the CDN system of an increased cache hit ratio is larger than for larger values of $\alpha_\tau$.

Both figures show a linear relation between the cache hit ratio $f$ and the average client-perceived response time ratio $R^*$, which is consistent with Equation (6).

### *D. Influence of the critical ratio*

The critical ratio $\Gamma$ equals the data transfer time divided by the processing time for a transaction between client and origin server in the CDN model. Fig. 8 and Fig. 9 show its influence on the client-perceived response time ratio.

In Fig. 8 the surrogate server is assumed to be located halfway and the RTD ratio equals 0.8. The figure shows that a higher critical ratio value enlarges the performance difference between the two systems. The client-perceived response time ratio stagnates when the critical ratio approaches to infinity, as can also be understood from Equation (6). E.g., the limiting value of $R^*$ equals 1.09375 if the cache hit ratio equals 0.25. When the cache hit ratio is increased from 0.25 to 0.4 or larger, the upgraded single server system is no longer the better solution.

In Fig. 9 the cache hit ratio is constant instead of the surrogate placement. If the surrogate placement is defined by $\alpha_\tau = 0.75$, then the critical ratio has no influence on the client-perceived response time ratio. A surrogate server located closer than this favors the CDN model, while the upgraded single server system performs better than the CDN model when $\alpha_\tau > 0.75$.

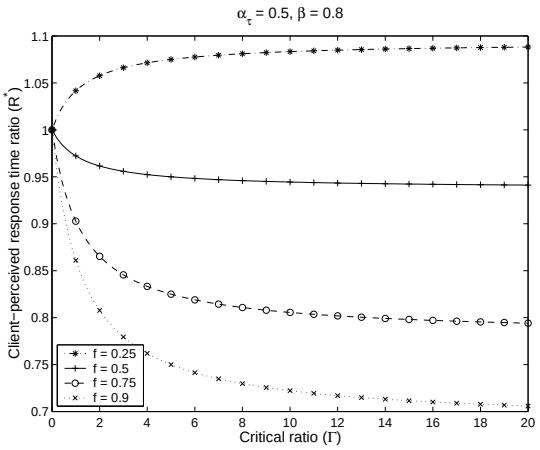


Fig. 8. Influence of the critical ratio on the client-perceived response time ratio for a fixed $\alpha_\tau$ and $\beta$.

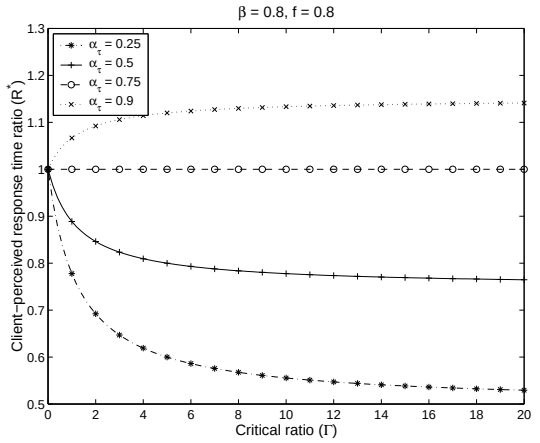


Fig. 9. Influence of the critical ratio on the client-perceived response time ratio for a fixed $\beta$ and $f$.

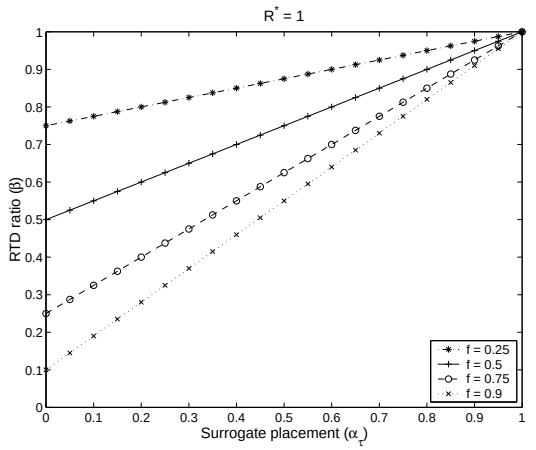


Fig. 10. Relation between surrogate placement and RTD ratio when R equals one.

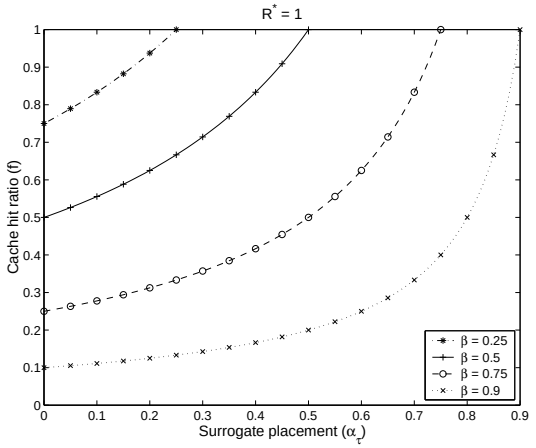


Fig. 11. Relation between surrogate placement and the cache hit ratio when R equals one.

### *E. The average client-perceived ratio equals one*

In this section, the client-perceived response time improvement obtained by both systems is assumed to be equal, which is indicated by $R^* = 1$. The combinations of the parameters $\alpha_\tau$, $\beta$ and $f$ that result in $R^* = 1$ are shown by the curves in Fig. 10 and Fig. 11. From these figures it is also easy to determine when one solution outperforms the other.

Fig. 10 shows a linear relation between the surrogate placement and the RTD ratio. The area above a chosen straight line, which represents a certain cache hit ratio, corresponds to all combinations $(\alpha_\tau, \beta)$ for which the CDN model is the preferred option. The area under the curve specifies all combinations for which the upgraded single server system is the better solution. The CDN system can never be the optimal choice if the RTD ratio is smaller than the surrogate placement.

The relation between the surrogate placement and the cache hit ratio when the two systems perform equally well is plotted in Fig. 11. Again the area under a curve specifies all combinations for which the upgraded single server system is the better option, while the area above the curve determines all possible combinations that make the CDN architecture the best choice. When the RTD ratio equals 0.25, a high cache hit ratio and a surrogate server close to the client are needed for the CDN solution to be the better option. A higher RTD ratio softens these restrictions such that also a lower cache hit ratio and a surrogate server located further away from the client could make the CDN system to perform equally well as the upgraded single server system.

## V. Conclusions

In this article a simple model to investigate the trade-off between bandwidth and storage in an upgraded single server system (bandwidth) and a CDN system (storage) was considered. Using this model one can decide under which conditions on the capacity improvement of the single server system and on the surrogate placement and the cache hit ratio of the CDN system the client-perceived response time in the CDN system is smaller or larger than that in the upgraded

single server system. Specifically, Equation (7) specified as an inequality can be employed to choose one of the two architectures. This equation is independent of the parameters $N$ and $P$, which characterize a certain application.

Using numerical results the influence of the different parameters that define the two models was illustrated. One could easily see that to have the CDN model outperform the upgraded single server system, the cache hit ratio, the RTD ratio and the critical ratio should be high, while the surrogate placement should be small. The different examples however showed in more detail what happens if a certain parameter is changed.

## REFERENCES


[1] D. C. Verma, *Content Distribution Networks: An Engineering Approach*, 1st ed. John Wiley & Sons, Dec. 2001.

[2] K. L. Johnson, J. F. Carr, M. S. Day, and M. F. Kaashoek, "The measured performance of content distribution networks," in *Proc. 5th International Web Caching and Content Delivery Workshop*, Lisbon, Portugal, May 2000.

[3] B. Krishnamurthy, C. Wills, and Y. Zhang, "On the use and performance of content distribution networks," in *Proc. First ACM SIGCOMM Workshop on Internet Measurement*, San Francisco, CA, USA, June 2001.

[4] D. Agrawal, J. Giles, and D. Verma, "On the performance of content distribution networks," in *Proc. SPECTS 2001*, Orlando, FL, USA, July 2001.

[5] D. Agrawal and J. Giles, "Modeling the response time of web pages," 2003, submitted for publication.

[6] S. B. Calo, D. Verma, D. Agrawal, and J. Giles, "On the effectiveness of content distribution networks," in *Proc. SPECTS 2002*, San Diego, CA, USA, July 2002.