Mathematical Assessment of CDN Servers in a Cloud Computing Environment: A Case of Big Data for e-Governance

Riktesh Srivastava, PhD
Associate Professor, Information Systems
Skyline University College
Sharjah, UAE.

ABSTRACT
Cloud computing environment delivers a countless tractability and accessibility of computing resources at a lesser price. This evolving technology unlocks an innovative eon of e-services provided by Government. However, as the number of users retrieving these e-services are cumulative, it is problematic for the current e-Government Infrastructure to accomplish these requirements. The exceptional way to succeed the problem is the employment and usage of CDN Servers. A CDN is a network of geographically distributed content delivery nodes that are settled for effectual delivery of digital content on behalf of content providers. The paper organizes the mathematical calculation to inspect the average response time for exploring the content from the e-Government Infrastructure implementing CDN Servers. There are three possible situations which are calculated in the research, as cited below:

- When the request is available at the first CDN Server
- When the request is not available at the first CDN Server but at the other CDN Servers
- When the request is not available at any CDN Servers, but to be transferred to e-Government Cloud Computing Infrastructure

This mechanism may best serve as a guideline to identify the best content server to respond to a request directed to the CDN Servers.

Keywords
CDN Servers, Hit Ratio, Miss Ratio, Average Response Time, Cloud Computing, e-Government Infrastructure.

1. INTRODUCTION
As e-Government Infrastructure started using Cloud Computing, the problem of maintaining the infrastructure, platform and even software was solved. However, as Cisco predicted on May 29, 2013[1], that every individual will be using 5 devices connected to Internet, speed of data transfer from remotely located Data Center was the major concern. Moreover, the spread of smart phones has enhanced people's interest in individual users' mobile media servers [2,3,4] and activated various media services.

The Solution to the above mentioned problem is called CDN (Content Delivery Network). Over the years, CDN [5, 6] has advanced to become a well-established technology for delivering a wide range of contents including Web objects, downloadable objects, applications, live streaming media and social networks. Although CDNs deliver content from e-Government Infrastructure to citizens and businesses with high availability and performance, they fail to meet the more recent, quickly increasing demand of multimedia functions on the delivery/server side [7,8], it is far more superior than delivering the contents with CDN Servers in place.

CDN is a network constructed from the group of "Caching Servers", which are strategically and geographically located. CDN is one of the most efficient mechanism, by which data from various departments in an e-Government Infrastructure would be serving a large number of devices. The architecture of CDN Servers along with Cloud Computing (Hybrid Clouds) for an e-Government infrastructure is given in Figure 1:

As the numbers of electronic devices are increasing (in billions), it becomes very difficult for the web and application server to handle all the request fast and hence CDN architecture was introduced.

A. CDN Structure
CDN includes Content Providers (CP) and Caching Servers (as mentioned in the Figure above). CP includes all the contents (usually, these are department websites/applications in e-Government Infrastructure). Amongst these, the most popular contents are loaded in the distributed set of caching servers.

User requests a content to its nearest Caching Server, CDN

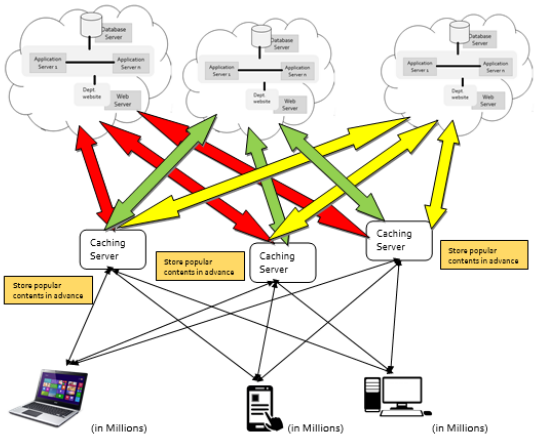

Fig 1.CDN Structure

Servers with Cloud Computing for e-Government Infrastructure which gets delivered to the user. Such a condition is called HIT RATIO.

If the content does not reside in the Caching Server, there are two possibilities:

1. The content may be in another Caching Server, thus, may be redirected from other Caching Server to the user, or
2. The content does not exist in any of the Caching Server, and needs to be redirected from remotely located (in Cloud) CP.

Both of these conditions are called MISS RATIO. Thus the biggest challenge for Caching Servers is to maintain the balance of contents, thereby, increasing HIT RATIO and reducing MISS RATIO.

The rest of the paper is divided as follows: Section 2 defines the mathematical study to calculate the response time at CDN Server. Section 3 calculates the average response time considering the hit ratio and miss ratio, thereby, calculating the average total response time. Section 4 gives the conclusion and forthcoming study to be conducted.

2. MATHEMATICAL STUDY TO EVALUATE THE RESPONSE TIME

For estimation of probability of $"n"$ requests at CDN Servers, certain assumptions are to be made. This can be given as follows:

1. Δt is a small time, in which only one requests arrive at any CDN Server.
2. The state of arrival is λ and state of departure is μ.

Probability of one arrival = $\lambda\Delta t$

and, Probability of one departure = $\mu\Delta t$

Then, Probability of no arrival = $1-\lambda\Delta t$

and, Probability of no departure = $1-\mu\Delta t$

If there are $"n"$ requests present at any time $"t"$. This, will be represented by $P_n(t)$. If the time is increased from $"t"$ to $(t+\Delta t)$, then there are three possibilities as mentioned in the equation 1 given below:

$$P_n(t+\Delta t)=\begin{cases} P_n(t).(1-\lambda\Delta t).(1-\mu\Delta t) \\ P_{n+1}(t).\mu\Delta t \\ P_{n-1}(t).\lambda\Delta t \end{cases} \quad (1)$$

$$P_n(t+\Delta t)=P_n(t).(1-\lambda\Delta t)(1-\mu\Delta t)+P_{n-1}(t).\lambda\Delta t+P_{n+1}(t).\mu\Delta t \quad (2)$$

Or

$$\frac{P_0(t+\Delta t)-P_0(t)}{\Delta t}=-\lambda P_n(t)-\mu P_n(t)+\lambda P_{n-1}(t)+\mu P_{n+1}(t) \quad (3)$$

But,

$$\underset{\Delta t\to 0}{Lim}\left\{\frac{P_n(t+\Delta t)-P_n(t)}{\Delta t)}\right\}=\frac{d}{dt}\{P_n(t)=0\}\ \textit{for stable condition}$$

Thus, the R.H.S. of equation 3 becomes

$$P_{n-1}(t).\lambda-(\lambda+\mu).P_n(t)+P_{n+1}(t).\mu=0 \quad (4)$$

To solve, equation 4, we need to consider the initial condition, i.e., there is 0 requests arriving at any of the CDN Servers at time$(t+\Delta t)$. This can be obtained from the states as given under:

$$P_0(t+\Delta t)=P_0(t).(1-\lambda.\Delta t)$$

$$=P_1(t).\mu\Delta t$$

$$=P_0(t)(1-\lambda\Delta t)+P_1(t)\mu\Delta t$$

$$\left(\frac{(P_n(t+\Delta t)-P_n(t)}{\Delta t}\right)=-P_0(t)\lambda+P_1(t).\mu \quad (5)$$

Thus, L.H.S. of equation 5 becomes

$$\underset{\Delta t\to 0}{Lim}\left\{\frac{(P_n(t+\Delta t)-P_n(t)}{\Delta t}\right\}$$

$$\frac{d}{dt}\{P_0(t)\}=0,\ \textit{at stable state} \quad (6)$$

Hence, equation 6 becomes

$$P_1(t)=\left(\lambda/\mu\right).P_0(t) \quad (7)$$

From equation 4 and equation 7, we derive the following:

$$\left.\begin{aligned} P_0(t)&=\left(\lambda/\mu\right)^0.P_0(t) \\ P_1(t)&=\left(\lambda/\mu\right)^1.P_0(t) \\ P_2(t)&=\left(\lambda/\mu\right)^2.P_0(t) \\ P_3(t)&=\left(\lambda/\mu\right)^3.P_0(t) \\ &\vdots \\ &\vdots \\ P_n(t)&=\left(\lambda/\mu\right)^n.P_0(t) \end{aligned}\right\} \quad (8)$$

Summation of all the equations in equation 8 is given as under:

$$\sum_{i=0}^{n}P_i(t)=\left\{(\lambda/\mu)^0+(\lambda/\mu)^1+(\lambda/\mu)^2.....+(\lambda/\mu)^n\right\}P_0(t) \quad (9)$$

Based on limiting condition, when $n\to\infty$, and $\lambda/\mu<1$, L.H.S. becomes 1 and R.H.S. becomes

$$\left[\frac{1}{\left(1-\lambda/\mu\right)}\right]P_0(t)$$

Thus equation 8 becomes

$$1=\left[\frac{1}{\left(1-\lambda/\mu\right)}\right]P_0(t) \quad (10)$$

If equation 9 is substituted in equation 8, then

$$P_n(t)=\left(\lambda/\mu\right)^n\left(1-\lambda/\mu\right) \quad (11)$$

Hence, the probability for the presence of $"n"$ data can be computed at any time $"t"$ provided rate of arrival and rate of departure at any CDN Server is calculated.

For variable $"n"$ the average value can be written as:

$$\text{Avg(Requests)} = \sum_{n\to\infty}^{N} nP_n(t)$$

$$=\sum_{n\to\infty}^{N}\left(\lambda/\mu\right)^n\left(1-\lambda/\mu\right)$$

$$=\left(1-\lambda/\mu\right)\sum_{n\to\infty}^{N}\left(\lambda/\mu\right)^n$$

$$=\left(1-\lambda/\mu\right)\left\{\lambda/\mu+2\left(\lambda/\mu\right)^2+3\left(\lambda/\mu\right)^3+.....\right\}$$

$$=\left(1-\lambda/\mu\right)\left(\lambda/\mu\right)\left\{1+2\left(\lambda/\mu\right)+3\left(\lambda/\mu\right)^2+.....\right\}$$

$$=\left(1-\lambda/\mu\right)\left(\lambda/\mu\right)\frac{d}{d\left[\lambda/\mu\right]}\left\{\left(\lambda/\mu\right)+\left(\lambda/\mu\right)^2+\left(\lambda/\mu\right)^3+.....\right\}$$

$$=\left(1-\lambda/\mu\right)\left(\lambda/\mu\right)\left(\left(\left(1-\lambda/\mu\right)+\lambda/\mu\right)\Big/\left(1-\lambda/\mu\right)^2\right)$$

$$=\left(\lambda/\mu\right)\Big/\left(1-\lambda/\mu\right) \quad (12)$$

Thus equation 12 mathematically measures the number of requests arrive at any CDN Server.

3. MATHEMATICAL EVALUATION-HIT RATIO AND MISS RATIO

In order to evaluate the requests at CDN Servers, there are three possible situations:

1) When the request is available at the first CDN Server
2) When the request is not available at the first CDN Server but at the other CDN Servers
3) When the request is not available at any CDN Servers, but to be transferred to e-Government Cloud Computing Infrastructure

The generic mathematical formula to evaluate the Response time is given in equation 13 below:

$$T_{avg} = \text{Hit Time} * \text{Hit Rate} + \text{Miss Time} * \text{Miss Rate} \quad (13)$$

Note that Hit Time is the time to find the data at CDN Server.

Also note that

$$\text{Miss Rate} = 1 - \text{Hit Rate} \quad (14)$$

Substituting the value of equation 14 in 13, we get

$$T_{avg} = \text{HT} * (1 - \text{MR}) + \text{MT} * \text{MR}$$

Upon evaluating equation 14,

$$T_{avg} = \text{HT} + \text{MR} * (\text{MT} - \text{HT}) = \text{HT} + (\text{MR} * \text{MP}) \quad (15)$$

where,

$$\text{MT} - \text{HT is Miss Penalty}$$

Based on equation 15, we can straightforwardly estimate the average response time for the three cases.

A. *Request available at the first CDN Server*

In this case, user requests are served at the first CDN Server, as indicated in Figure 2.

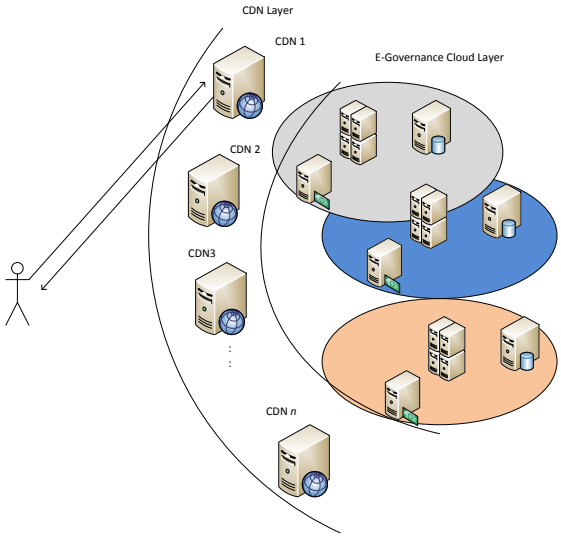

Fig 2.Case 1: User requests served at first CDN Server

For this case, MR=0

$$T_{avg\ (Case\ 1)} = \text{HT} + 0 * \text{MP} = \text{HT} \quad (16)$$

Substituting the value of equation 13 in 16, we get

$$T_{avg\ (Case\ 1)} = \left(\frac{\left(\frac{\lambda}{\mu}\right)}{\left(1-\frac{\lambda}{\mu}\right)}\right) \quad (17)$$

B. Request available at k^{th} CDN Server

This is the instance when the content is not accessible at the first CDN Server, but available at kth CDN Server. Figure 2 clarifies the condition in much detail:

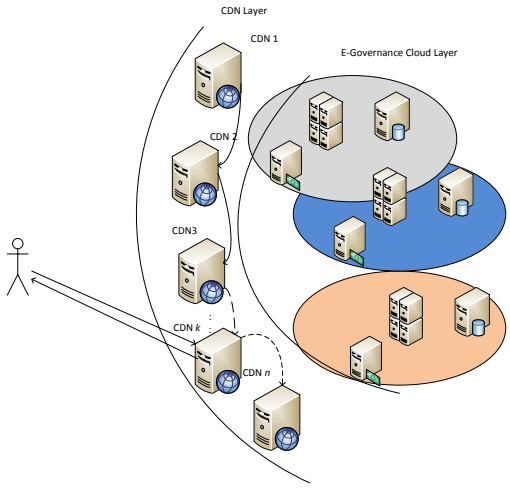

Fig3.Case 2: User request get served at k^{th} CDN Server

Mathematical formulation for this case is as follows:

$$T_{avg(Case2)} = CDN_1^{HT} + (CDN_1^{MR} + CDN_1^{MP}) \quad (18)$$

Evaluating the conditions, we get:

$$T_{avg\ (Case\ 2)} = \sum_{i=1}^{n-k} MR_i * \sum_{i=1}^{k} HT_i + \sum_{i=1}^{k} (MR_i * MP_i) \quad (19)$$

C. Request not available at any CDN Servers

This situation can be explained using the following figure (Figure 4)

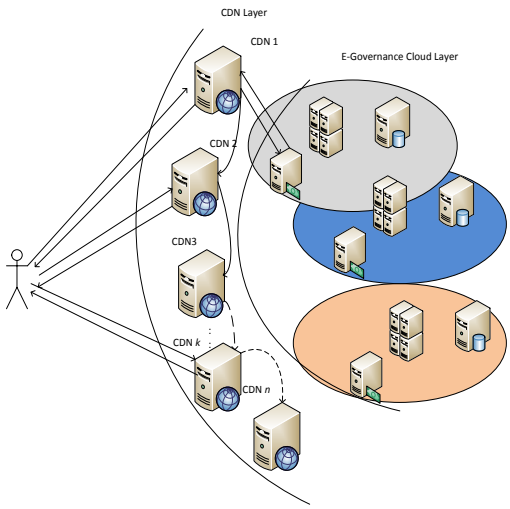

Fig 4.Case 3: User request not available at any CDN Server

If the user request is not available at any CDN Server, then it gets transferred to the specific Department and then creates a copy in one of the CDN Servers.

The mathematical equation can be mentioned as

$$T_{avg\ (Case\ 3)} = \sum_{i=1}^{k} (MR_i * MP_i) + \left(\frac{\left(\frac{\lambda}{\mu}\right)}{\left(1-\frac{\lambda}{\mu}\right)}\right) \quad (20)$$

D. Classifying the Cache Hit Ratio

The most suitable mechanism to classify the cache performance is to divide it into 3 categories as mentioned below:

Table 1.CDN Types

CDN Type	Mathematical Evaluation		
	Terminology	***Success Rate***	***Formula***
Case 1	Hot Cache	99%	$T_{avg(Case1)} = \left(\frac{\left(\frac{\lambda}{\mu}\right)}{\left(1-\frac{\lambda}{\mu}\right)}\right)$
Case 2	Warm Cache	80%	$T_{avg(Case2)} = \sum_{i=1}^{n-k} MR_i * \sum_{i=1}^{k} HT_i + \sum_{i=1}^{k} (MR_i * MP_i)$
Case 3	Cold Cache	0-15%	$T_{avg(Case3)} = \sum_{i=1}^{k} (MR_i * MP_i) + \left(\frac{\left(\frac{\lambda}{\mu}\right)}{\left(1-\frac{\lambda}{\mu}\right)}\right)$

4. CONCLUSION

Cloud Computing (SaaS) is used by many e-Government Infrastructures. These government departments use CDN Servers to improve performance, reliability, and scalability. CDN Servers replicate content over several copied CDN servers to improve response time. A CDN also increases network performance by exploiting bandwidth, refining availability, and upholding accuracy through content replication. Unfortunately, although many viable CDN providers exist, they don't collaborate in delivering content to end users in an accessible manner. In accumulation, the e-Government Infrastructures typically subscribe to one CDN provider and thus can't use multiple CDNs at the same time. Such a closed, non-cooperative model results in "isles" of CDNs. The paper does the mathematical assessment of the average response time for three conditions. These conditions are- when the request is available at first instance, when the request is available at any of the CDN Servers or the request is not available at any CDN Servers and needs to be taken from Web Server. The study may provide the guidelines to create open content and service delivery networks (CSDNs) that scale well and can share resources with other CSDNs through cooperation and coordination, thus overcoming the island CDN problem.

5. REFERENCES

[1] http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[2] H. Mao, N, Xiao, W, Shi and Y Lu, "Wukong: A cloud-oriented file service for mobile Internet devices,' Journal of Parallel and Distributed Computing, vol. 72, pp. 171-184, 2012.

[3] Wikipedia,http://en.wikipedia.org/wiki/Mobile_media (2011).

[4] Y, W, Kao, C,F, Lin, K. A. Yang and S.M. Yuan, "A Web-based, Offline-able, and Personalized Runtime Environment for executing applications on mobile devices," Computer Standards & Interfaces, vol. 34 (2012), pp. 212-224, 2012.

[5] Peng G. "CDN: Content distribution network". Stony Brook University, Technical Report, TR-125; 2008.

[6] Pallis G, Vakali A. "Insight and perspectives for content delivery networks," Commun ACM 2006;49(1):pp. 101–106, 2006.

[7] Zhu W, Luo C, Wang J, Li S. "Multimedia cloud computing," IEEE Signal Proc Mag, pp. 59–69, 2011.

[8] Ranjan R, Mitra K, Georgakopoulos D. "MediaWise cloud content orchestrator," J Internet Serv Appl;4(2), 2013.