Investigating Structure of Modern Web Traffic

Noriaki Kamiyama[*†], Yuusuke Nakano[*†], Kohei Shiomoto[†],
Go Hasegawa[‡], Masayuki Murata[*], and Hideo Miyahara[*]
[*]Department of Information Science, Osaka University, Osaka 565-0871, Japan
Email: {kamiyama.noriaki, nakano.yuusuke, murata, miyahara}@ist.osaka-u.ac.jp
[†]NTT Network Technology Labs, Tokyo 180-8585, Japan, Email: shiomoto.kohei@lab.ntt.co.jp
[‡]Cybermedia Center, Osaka University, Osaka 560-0043, Email: hasegawa@cmc.osaka-u.ac.jp

***Abstract*—Modern websites consist of many rich objects dynamically produced by servers and client terminals at diverse locations. Consequently, we face complications in understanding the communication structure generated when accessing websites. To reduce the response time at browsed websites, many website objects are delivered using content delivery networks (CDNs), in which data objects are delivered from cache servers located close to user terminals. Although the use of CDNs have been assumed to reduce web response time, the actual effect of CDNs on this reduction has not been clarified. To answer this fundamental question, we measured the communication structure of traffic generated when accessing the 1,000 most popular websites from 12 locations worldwide. We found, for example, that it will be desirable to give high priority to entertainment websites at night and to business-related websites during the day.**

I. Introduction

In recent years, a large percentage of Internet traffic has been dominated by HTTP traffic. For example, an analysis of traffic measured on a backbone link between Japan and the USA from 2006 to 2008 indicated that about 60% of the traffic consisted of HTTP packets [6]. However, it was reported that two-thirds of the users encountered slow websites every week, and about half of the users abandoned websites after experiencing performance issues [7]. More concretely, it was reported that 67% of users experienced a long waiting time every week when browsing websites, and 17% of users would not wait if the time exceeded 5 seconds [11]. Another report claimed that users expect a page to load in 2 seconds or less, and that 40% of them will wait for no more than 3 seconds before leaving a site [17].

It was also reported that a 400-millisecond delay resulted in a 0.74% decrease in searches on the Google search engine [32], and that Amazon increased revenue 1% for every 0.1 second reduction in web-page load time [33]. It was also reported that users on high-performance sites were 15% more likely to complete purchases and 9% less likely to abandon the sites after viewing only one page [18]. Therefore, reducing web response time is an urgent issue for many Internet service providers (ISPs) and content providers. It is important to adequately control web traffic to improve user-perceived quality and reduce the amount of network resources consumed.

Each website consists of a large number of data objects, and objects are transmitted from object servers using HTTP. With traditional websites, static objects are stored at servers, and web browsers simply download them. Content delivery networks (CDNs), which use a number of cache servers deployed in multiple networks, have been widely used to efficiently transmit web traffic and reduce response time [18][20][31]. Currently, 74% of the 1,000 most frequently accessed websites use CDNs [20]. Although CDNs are typically operated by CDN providers, e.g., Akamai, the number of CDNs operated by large-scale content providers, such as Google, and by tier-1 ISPs, such as AT&T, has been increasing recently [15].

Although the use of CDNs has been assumed to reduce web response time, the actual effect of CDNs on this reduction has not been clarified. To improve the response time and suppress the amount of traffic transferred in networks, we need to adequately control caches, i.e., the placement and replacement of objects, based on the communication structure of web traffic, so it is important to clarify the structure of modern web traffic. To answer this fundamental question, we measured the traffic generated when browsing the 1,000 most popular websites from 12 locations worldwide and investigated the communication structure of traffic generated when accessing various websites from various global locations. The contributions of this paper are summarized as follows:

- We developed and implemented an active-based framework for measuring the communication structure of web traffic and implemented a method of classifying web objects based on website categories as well as CDN use.
- To investigate the object-deployment tendencies of websites in the current Internet, we measured the distance and round-trip-time (RTT) to the origin or cache servers from the client terminals as well as the delay caused in obtaining objects.
- Based on the experimental results, we clarified various tendencies of the geographical distribution of original objects and CDN caches.

After briefly reviewing current methods for measuring the patterns of web traffic in Section II, we propose a framework for measuring the structure of web traffic in Section III. In Section IV, we explain the experimental results of object-deployment tendencies in each website category by accessing websites from various access points using PlanetLab. Finally, we conclude the manuscript in Section V.

II. Related Work

Several studies have pointed out the problem of taking a long time when accessing websites consisting of many objects [32][33]. A number of techniques to improve web performance have been investigated [28]. Baeza-Yates et al. and Butkiewicz et al. investigated web traffic using the active measurement approach [4][7]. Baeza-Yates et al. compared the tendencies of properties including the size of web traffic and degree of connected graphs in each country based on the results of crawling websites from 12 countries in 2004 [4]. Butkiewicz et al. periodically accessed randomly selected websites and investigated properties of HTTP traffic including the object count for each website and number of servers accessed [7].

Ager et al., Bent et al., Gill et al., Ihm et al., and Schneider et al. analyzed web traffic based on passive measurement [1][5][10][12][25]. Ihm et al. analyzed the changes in various properties of web traffic using the proxy access logs of websites over a five-year period from 2006 to 2010 [12]. Bent et al. investigated the frequency of using cookies based on packet capture data from one day in 2004 [5]. Gill et al. analyzed the web access traffic from an enterprise and a university and clarified the patterns of web use [10]. Ager et al. proposed a method for identifying the content location over a CDN or data center and a method of selecting servers based on the measurement of control packets for domain name system (DNS) and a snapshot of a Border Gateway Protocol (BGP) routing table [1]. Schneider et al. extracted the HTTP and AJAX sessions from packet capture data and investigated the difference in generated-traffic patterns [25].

However, these studies based on active or passive measurements did not focus on the geographical communication structures, such as the distance between servers and client terminals, when accessing websites.

III. Measurement Procedure

First, we briefly describe the procedure used to measure traffic properties when accessing various websites.

A. Generation of URL List

The traffic generated when users access popular websites should be analyzed to investigate trends in the communication patterns of websites. Quantcast provides a ranking list of websites accessed by users in each country [23], and we used this ranking list for selecting the most popular websites. The URLs shown in the access ranking list are the home pages of each website, so we analyzed the properties of traffic generated only when accessing the home pages. We did not evaluate which pages can be accessed from the home pages; therefore, we did not evaluate user behavior when they were browsing websites. However, we were able to roughly investigate the trends in web traffic by analyzing the communication properties generated when users access many websites.

We classified the extracted data into URL categories to investigate the differences in communication patterns among various types of websites. The website of Alexa provides URL lists classified by category, e.g., Arts and Business [3], and we classified the extracted data into the 16 URL categories listed in Table I based on this list.

TABLE I
Number of websites for each website category used in clustering analysis

ID	Category	Midnight	Noon
C1	Business	59	40
C2	Computers	112	91
C3	News	39	27
C4	Reference	112	109
C5	Regional	80	73
C6	Science	95	86
C7	Society	79	83
C8	Health	86	52
C9	Home	85	47
C10	Shopping	69	68
C11	Adult	112	102
C12	Arts	55	60
C13	Games	87	58
C14	Kids & teens	106	64
C15	Recreation	86	52
C16	Sports	38	53

B. Acquisition of Web Traffic Properties

We acquired the HTTP archive (HAR) files [19] to obtain the communication properties generated when sending a GET message of an HTTP request from the probing terminal, i.e., client terminal. In the HAR files, various communication properties, e.g., the host URL from which each object is downloaded, the size of each object, and the delay caused in obtaining each object, are output as JavaScript Object Notation (JSON).

We continuously and automatically accessed a large number of websites by using the netsniff.js executable on phantomjs in which JavaScript can be executed on the command line [9]. Many cacheable objects, e.g., video or pictures, are cached at client terminals, and these objects in the local cache are reused when accessing the same websites from the same client terminal. When obtaining objects from the local cache, no communication is generated on the networks, so we need to focus on cases in which objects are downloaded from remote hosts to investigate the communication structure of web traffic. We obtained all the objects from remote servers by invalidating the local cache of the probing terminal.

After obtaining the HAR file for each website accessed, we could extract various data from the information of each object included in each obtained HAR file. We specifically focused on the total delay, which is the time required to obtain each object, i.e., the time from the completion of sending request packets at the client terminal to the beginning of the arrival of response packets at the client terminal. The total delay of object includes the network delay between the client terminal and object server as well as the object server processing time, such as for generating objects.

We also evaluated the average distance of objects, which is defined as the Euclidean distance between the probing client terminal and object server. To calculate the average distance, we obtained the country name, city name, and coordinates of the host from which each object was transmitted by using the GeoIP API provided by MaxMind [16]. Because the object distance is the Euclidean distance between the access host and object server, and this metric is different from the distance in the Internet, we also measured the RTT from the probing client terminal to each object server. To measure the RTT of each object, we automatically sent a ping command to each object server immediately after obtaining the HAR file of each website at the probing client terminal.

C. Measurement from Various Access Locations

We accessed websites from multiple access locations using PlanetLab [21], which is an overlay network constructed on the Internet and consists of over 500 hosts worldwide. With PlanetLab, we were able to execute various types of programs on a number of selected hosts. By executing the procedure described in the previous two subsections, we were able to access various websites from various locations worldwide. We selected a total of 12 measurement locations on PlanetLab: three points in North America (NA), two in Europe (EU), one in Russia (RU), two in Oceania (OA), one in Asia (AS), two in South America (SA), and one in Africa (AF). These 12 locations are shown in Table II. We summarize the procedure of accessing websites from various locations as follows.

(1) *Construction of measurement environment on PlanetLab*:
Before starting the measurement, we constructed a measurement environment on PlanetLab. First, we booted hosts for measurement using the GUI provided by PlanetLab. Next, we uploaded the URL list of the measurement target, as well as some programs for collecting HAR files for extracting statistical data from the HAR files.

(2) *Measurement*:
To compare the trend in web traffic patterns among various access points, we needed to start the measurement procedure described in the two previous subsections at the same local time at all the access locations. To satisfy this requirement, we used cron command of UNIX, which is a demon process to automatically start jobs. Because the time was set in all the PlanetLab hosts based on coordinated universal time (UTC), we derived the local time from UTC.

(3) *Derivation and collection of measurement results*:
The size of the HAR files obtained at each host of PlanetLab was huge, so we extracted only the required information from the HAR files and generated statistical data files at each PlanetLab host. Finally, the produced data files were sent to the collector terminal in our laboratories.

We selected 300 websites with the highest access count from each of the 16 website categories. Because some websites are categorized into multiple categories, we selected 4,290 websites in total after removing the duplications in multiple categories. From each location, we continuously accessed these 4,290 websites starting at midnight (0:00) and noon (12:00) (local time of each access location). The total number of websites that were evaluation targets was 1,124 at midnight and 927 at noon. Table I also summarizes the number of websites that were evaluation targets in each website category.

TABLE II
MEASUREMENT LOCATIONS

ID	Area	Location	ID	Area	Location
L1	NA	Massachusetts	L7	OA	Australia
L2	NA	Wisconsin	L8	OA	New Zealand
L3	NA	California	L9	AS	Japan
L4	EU	Ireland	L10	SA	Ecuador
L5	EU	Germany	L11	SA	Argentina
L6	RU	Russia	L12	AF	Reunion

D. Classification of Objects Based on CDN Use

By investigating the tendencies in web traffic of objects delivered using CDNs (denoted as *CDN objects*) and objects delivered without using CDNs (denoted as *non-CDN objects*), we can investigate the geographical tendency of cache deployment of CDNs and the locations where original objects of websites are provided. We classified the objects extracted from the HAR files into the two sets, i.e., CDN objects and non-CDN objects, by creating a list of second-level domains of hosts delivering CDN objects.

First, we listed the second-level domains of various CDN providers by manually checking websites of various CDN providers. We obtained a total of 44 second-level domain names of CDN caches, e.g., edgesuite.net, cloudfront.net, and akamaiedge.net. The domain names of objects extracted from the HAR files are those of content providers, e.g., www.yahoo.com, and the domain names of hosts delivering objects. e.g., host1.akamaiedge.net, are different from those extracted from the HAR files. We obtained the domain names of the hosts actually delivering objects by using the dig command. Finally, we identified CDN objects by comparing the second-level domain obtained by the dig command with each of the second-level domains included in the generated list.

E. Clustering Analysis of Web Traffic Properties

To understand the manner in which properties tend to differ when accessing websites from various locations, we used the approach of clustering analysis. As shown in Fig. 1, we accessed various websites, Y_1 and Y_2, from various measurement locations, X_1, X_2, and X_3, at each measurement time t. When we accessed each website from N access locations, we obtained N results of each property, e.g., the average RTT, for the same website. Therefore, we could construct $\boldsymbol{v}(y,t)$, a vector of N dimensions in which each element $v_{y,t,k}$ $(1 \leq k \leq N)$ is the value of v measured at location k when accessing website y at time t. When we accessed M websites at time t from N locations, we obtained M vectors $\boldsymbol{v}(y,t)$ for $1 \leq y \leq M$. Using the obtained M vector $\boldsymbol{v}(y,t)$, we applied the clustering analysis to investigate the trends in the way each property v differs among the accessed locations.

The k-means method is the most widely used method for centroid-based clustering, and we applied it to cluster websites on the basis of the property vectors. Let us consider the case in which there are n members, and each member i is associated with the property vector $\boldsymbol{x}_i$. Given a set of members, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\cdots$, $\boldsymbol{x}_n$, where each member is a d-dimensional real vector, k-means clustering aims to partition the n members into k sets $\boldsymbol{S} = \{S_1, S_2, \cdots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\mathrm{WCSS} \equiv \arg\ \min_{\boldsymbol{S}} \sum_{i=1}^{k} \sum_{\boldsymbol{x}_j \in S_i} |||\boldsymbol{x}_j - \boldsymbol{m}_i|||^2$$

where $\boldsymbol{m}_i$ is the mean of members in S_i, i.e., the centroid of cluster S_i.

It is widely known that k-means method results are strongly affected by the initial clusters, i.e., the initial k centroids. One of the major problems with the k-means method is that the approximations found can be arbitrarily bad with respect to the objective function compared to the optimal clustering. To address this problem, Arthur et al. proposed the k-means++ method, which involves initializing the cluster centers before proceeding with the standard k-means optimization iterations [2]. The basic idea of this method is spreading out the k initial cluster centers. The first cluster center is chosen uniformly at random from the data points that are being clustered, after which each subsequent cluster center is chosen from the remaining data points with probability proportional to its squared distance from the points closest to the existing cluster center. We used the k-means++ method to avoid the initial cluster problem.

Moreover, the k-means method results strongly depend on the parameter k, i.e., the number of clusters. In this study, we used the JD method proposed by Jain et al. to optimally determine k [13]. We can anticipate that the optimum k minimizing the distance between each member and the centroid of the cluster it belongs, while maximizing the distance between centroids between any pair of clusters. The JD method is based on this insight, and the JD method selects k minimizing the cost function $p(k)$ which is defined as

$$p(k) = \frac{1}{k} \sum_{i=1}^{k} \max_{1 \le j \le k} \left\{ \frac{\eta_i + \eta_j}{\xi_{ij}} \right\}$$

where

$$\begin{aligned} \eta_j &= \frac{1}{n_j} \sum_{i=1}^{n_j} D\left(\boldsymbol{x}_i^{(j)}, \boldsymbol{m}_j\right), \\ \xi_{ij} &= D(\boldsymbol{m}_i - \boldsymbol{m}_j). \end{aligned}$$

n_j is the number of members classified into cluster j, and $D(\boldsymbol{a} - \boldsymbol{b})$ is the distance between vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. k minimizing $p(k)$ in the range of $2 \le k \le 1 + \log_2 n$ is selected.

IV. Measurement Results

A. Average Properties

The average object size, average object count, and average total data size over all the 12 locations was 14.4 kbytes, 56.9, and 0.796 Mbytes, respectively. Table III summarizes the average object size (kbytes), number of objects, and total data size (Mbytes) of each website in each URL category. It should be noted that these values are the averages over all the 12 locations. The total amount of data and the number of objects accessed tended to be large in entertainment websites, i.e., Arts, Shopping, and Sports[1], whereas it tended to be small in websites for gathering information, i.e., Business, Computers, Health, and Reference. We obtained similar results for the number of locations and hosts accessed as well.

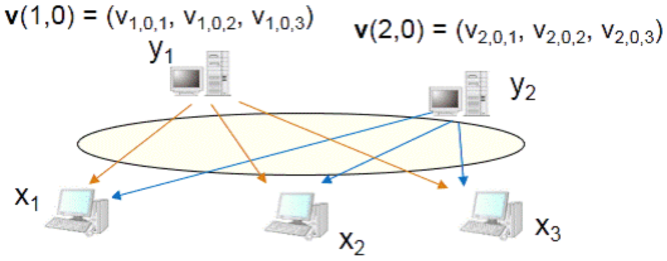

v(1,0) = (v1,0,1, v1,0,2, v1,0,3)
y1
v(2,0) = (v2,0,1, v2,0,2, v2,0,3)
y2
x1
x2
x3

Fig. 1. Clustering websites on basis of location properties

Figure 2(a) plots the average ratio of CDN objects in each URL category observed at each of the 12 locations at midnight, and Fig. 2(b) plots the same properties at noon. Almost all the curves of the 12 locations had similar values at all 16 URL categories, and we confirmed that the ratio of CDN objects was almost identical among the 12 access locations. On the other hand, the ratio of CDN objects largely differed among the 16 URL categories and varied between about 0.2 and 0.5 depending on the URL categories. The websites of Computers, News, Society, Shopping, Arts, and Kids & teens tended to consist of more CDN objects, whereas those of Business, Regional, Science, Adult, and Games tended to consist of fewer CDN objects.

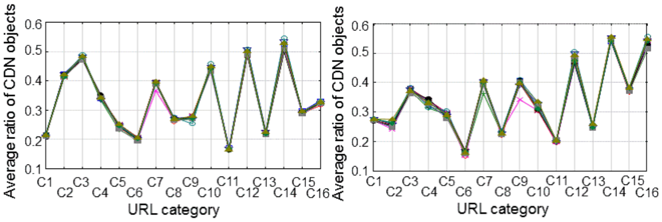
TABLE III
Average object size, object count, and total data size in each URL category

	Object size (kbytes)	Object count	Total size (Mbytes)
Business	14.70	55.14	0.810
Computer	16.26	43.63	0.709
News	13.55	72.45	0.982
Reference	13.09	43.42	0.568
Regional	17.77	50.59	0.899
Science	14.04	52.86	0.742
Society	15.01	66.86	1.003
Health	14.27	54.30	0.775
Home	15.66	55.39	0.867
Shopping	15.67	70.77	1.109
Adult	10.49	53.04	0.557
Arts	15.43	68.18	1.052
Games	15.28	54.12	0.827
Kids&teens	13.23	54.59	0.722
Recreation	13.55	57.30	0.776
Sports	16.62	86.67	1.440

B. Geographical Distribution of Original Objects

We then investigated the tendency in geographical distribution of original objects in each URL category through the

[1] This agrees with the results of Butkiewicz et al. [7].



(a) midnight

(b) noon

Fig. 2. Average ratio of objects delivered using CDN in each URL category at each access location

clustering analysis of average distance and RTT for non-CDN objects. Figure 3(a) plots the centroids of the average distance of non-CDN objects of websites classified into each cluster at each of the 12 access locations using the midnight dataset. It should be noted that the centroid of each cluster is an N-dimensional vector whose k-th element is the average object distance from the access location k among the websites classified into this cluster, and it represents the geographical tendencies of object distance on this cluster. Figure 3(b) shows the ratio of websites classified into each cluster in each of the 16 website categories (denoted as *C1, C2, $\cdots$, C16*) as well as in all the categories (denoted as *All*). We labeled the clusters in descending order of the website count classified into each cluster.

The geographical tendencies of original objects were classified into three clusters. About 80% of the websites were classified into the two largest clusters (Clusters 1 and 2), and we observed identical tendencies in the average object distance in these two clusters: close in North America, moderate in Europe, South America, Russia, and Africa, and far in Oceania and Asia. Therefore, we can assume that a large number of original objects of various websites are located in North America and many content providers exist in North America. In Cluster 3, on the other hand, original objects tended to be provided in just Europe. The difference in geographical tendencies of original objects among URL categories was small.

Similarly, Figs. 3(c) and (d) show the same properties for the noon dataset, and websites were also classified into three clusters. Cluster 1 in the noon dataset corresponds to Clusters 1 in the midnight dataset, and Cluster 2 in the noon dataset corresponds to Cluster 3 in the midnight dataset. Although a new cluster, i.e., Cluster 3, in which original objects tended to be provided in Asia and Oceania, appeared in the noon dataset, just a limited number of websites were classified into Cluster 3.

Although we can roughly estimate the physical distance between two nodes by using the Euclidean distance, the distance over the networks does not agree with this distance. Therefore, we also investigated website clustering based on the average RTT to object servers. Figures 4(a) and (b) respectively plot the centroids of the average RTT of non-CDN objects of websites classified into each cluster and the ratio of websites classified into each cluster when using the midnight dataset. The websites were classified into four clusters, and we observed the different tendencies among the URL categories in average RTT, unlike the case of average distance.

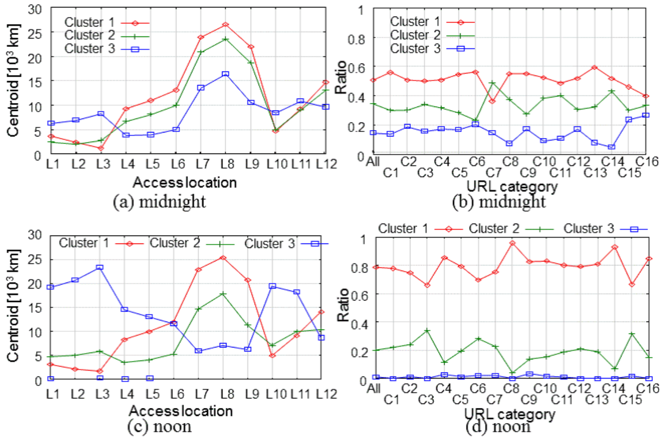
The average RTT to servers providing original objects of websites classified into Cluster 1 was small only in North America, and the ratio of Society, Adult, Recreation, and Sports websites classified into Cluster 1 was large. The websites of these categories tended to be provided by content providers in North America. The centroid of Cluster 2 was small in North America, Europe, and Asia, and more websites of Computers, News, Reference, Science, Arts, Games, and Kids & teens were classified into this cluster. Many content items of these categories were provided by content providers in North America, Europe, and Asia. For example, we can guess that many websites of Games and Kids & teens were provided by major Japanese content providers. We can say that the geographical locality of many websites classified into Clusters 1 and 2 was weak, and identical content tended to be viewed from various regions.

On the other hand, the centroid of Cluster 3 was small in all the areas excluding Africa, and more websites of Home and Shopping tended to be classified into this cluster. The geographical locality of many websites of Home and Shopping was high, and the websites of these categories tended to be provided from various countries. Therefore, original objects were obtained from servers provided at locations close to each access location. Finally, the average RTT to servers providing original objects of websites classified into Cluster 4 was small only in Europe, and only less than 10% of websites of all the categories were classified into this cluster.

Figures 4(c) and (d) respectively plot the same properties using the noon dataset in which the websites were classified into five clusters. Clusters 1, 2, and 3 in the noon dataset correspond to Clusters 2, 1, and 3 in the midnight dataset, respectively. Moreover, Clusters 4 and 5 in the noon dataset correspond to Cluster 4 in the midnight dataset. We confirmed that the tendency of the centroids of clusters was similar to that in the midnight dataset.

C. Geographical Distribution of CDN Caches

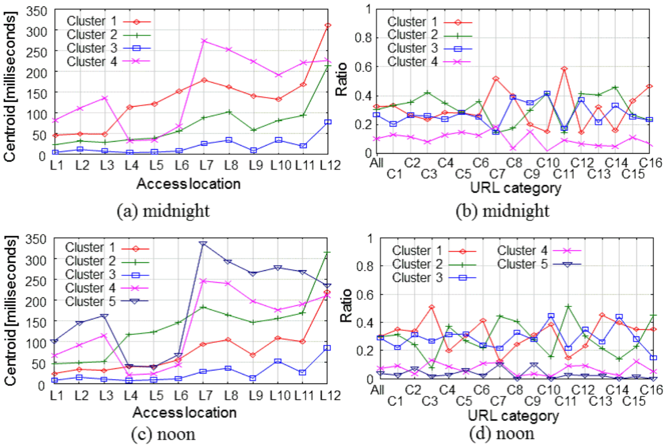
Next, we investigated the geographical tendencies of cache servers of CDNs through clustering analysis of the average distance and RTT of CDN objects. Figures 5(a) and (b) respectively plot the centroids of the average distance of CDN objects of websites classified into each cluster and the ratio of websites classified into each cluster when using the midnight dataset. Websites were classified into five clusters, and we confirmed that the geographical-deployment patterns of CDN caches were not identical, and various deployment patterns of CDN caches existed. The centroid of Cluster 1 was small in all the areas except Asia and Africa, and more websites of News, Reference, Regional, Health, and Kids & teens tended to be classified into this cluster. The centroid of Cluster 2 was small in North America, Europe, and South America, and more websites of Home, Shopping, Arts, and Recreation tended to be classified into this cluster. The average distance of websites classified into the other three clusters, Clusters 3, 4, and 5, was



(a) midnight (b) midnight

(c) noon (d) noon

Fig. 3. (a)(c) Centroids of average distance of non-CDN objects at each access location, (b)(d) ratio of websites classified into each cluster in each website category



(a) midnight (b) midnight

(c) noon (d) noon

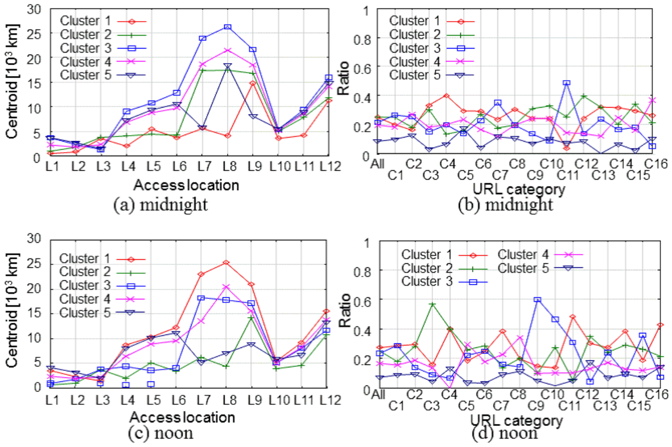
Fig. 4. (a)(c) Centroids of average RTT of non-CDN objects at each access location, (b)(d) ratio of websites classified into each cluster in each website category

small in North America, and Business, Computers, Society, and Adult websites were more likely to be classified into this cluster.

Figures 5(c) and (d) show similar results when using the noon dataset. Cluster 2 in the noon dataset corresponds to Cluster 1 in the midnight dataset, and more websites of News, Reference, and Arts tended to be classified into this cluster. Cluster 3 in the noon dataset corresponds to Cluster 2 in the midnight dataset, and Home, Shopping, and Recreation were more likely classified into this cluster. The other three clusters, Clusters 1, 4, and 5 correspond to Clusters 3, 4, and 5, respectively, in the midnight dataset, and websites of Society, Health, and Adult tended to use the CDNs in which cache servers were provided in North America.

Figures 6(a) and (b) respectively plot the centroids of the average RTT of CDN objects of websites classified into each cluster and the ratio of websites classified into each cluster when using the midnight dataset. The websites were classified into five clusters, and we also confirmed that the geographical-deployment patterns of CDN caches were not identical, and various deployment patterns of CDN caches existed. The centroids of Clusters 1, 2, and 3 were small in all the regions except South America and Africa, and more than 80% of websites of all the categories except Adult were classified into one of these three clusters. The centroids of the other clusters, i.e., Clusters 4 and 5, were small in North America and Europe, and Adult websites were more likely to be classified into one of these clusters.

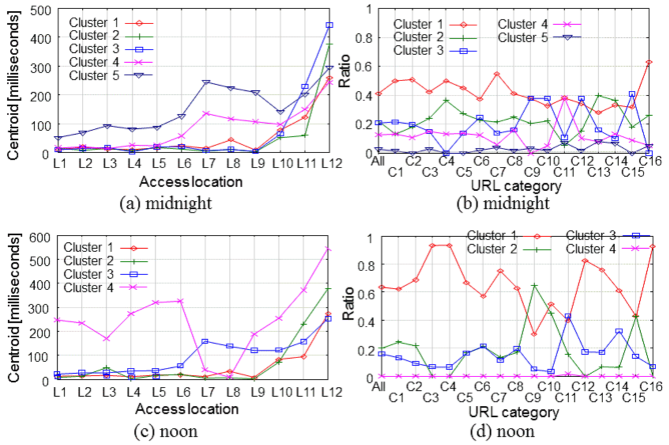
Figures 6(c) and (d) show similar results when using the noon dataset. We obtained similar tendencies with the midnight dataset. Although we observed a cluster, i.e., Cluster 4, that was not obtained in the midnight dataset, and the centroid of this cluster was small in Oceania, only a few websites were classified into this cluster.



(a) midnight (b) midnight

(c) noon (d) noon

Fig. 5. (a)(c) Centroids of average distance of CDN objects at each access location, (b)(d) ratio of websites classified into each cluster in each website category



(a) midnight (b) midnight

(c) noon (d) noon

Fig. 6. (a)(c) Centroids of average RTT of CDN objects at each access location, (b)(d) ratio of websites classified into each cluster in each website category

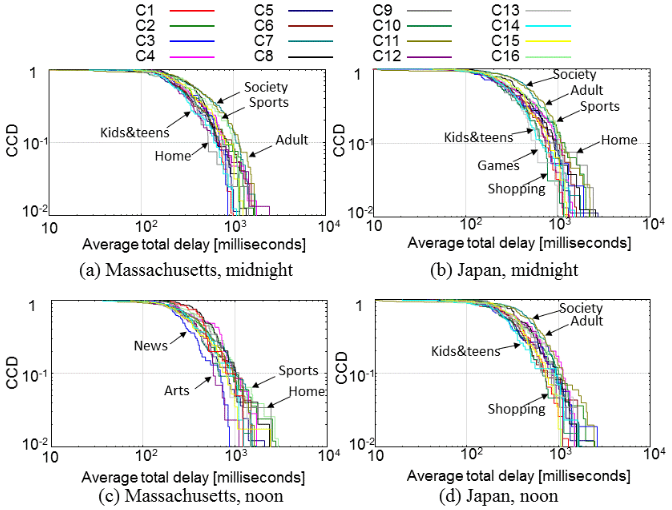
D. Average Total Delay of Objects

Finally, we investigated the tendencies of the average total delay of objects in each URL category and evaluated the effectiveness of using CDNs on the reduction of delay by comparing the average total delay of CDN objects and that of non-CDN objects. As mentioned in Section III-B, the total delay of objects is defined as the time required to obtain each object, i.e., the time from the completion of sending request packets at the client terminal to the beginning of the arrival of response packets at the client terminal. The total delay of objects includes the network delay between the client terminal and object server as well as the object server processing time, such as for object generation.

Figures 7(a) and (b) plot the complementary cumulative distribution (CCD) of the average total delay of non-CDN objects of each URL category using the midnight dataset measured at the PlanetLab host in Massachusetts and Japan, respectively. In Figs. 7(c) and (d), we also show the CCD of the average total delay of non-CDN objects using the noon dataset. The dependence of the average total delay of non-CDN objects on the access location as well as the measurement time was weak. We also observed the tendencies in which the average total delay of non-CDN objects was small in Shopping and Kids & teens websites, whereas it was large in Adult, Sports, and Society websites. Therefore, we can assume that the response time of Adult, Sports, and Society websites will be effectively reduced by increasing the use of CDNs.

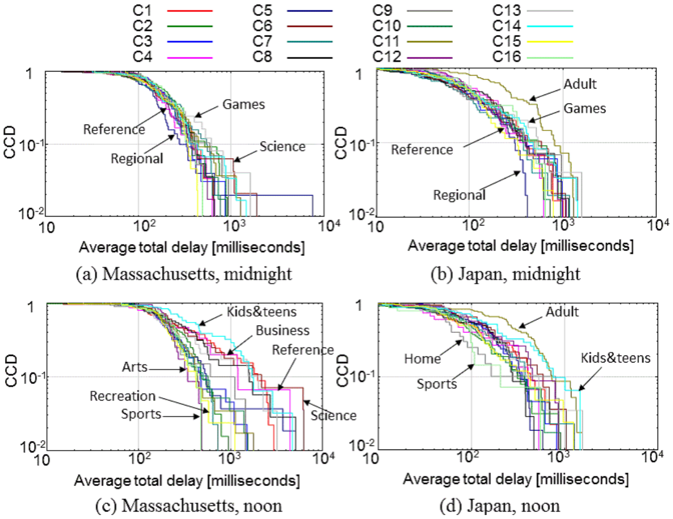
Finally, Fig. 8 plots the CCD of the average total delay of CDN objects in each URL category measured at the two access location at the two access times. Compared with the results of non-CDN objects shown in Fig. 7, the tendency of the average total delay of CDN objects strongly depended on the access location and time. The results depended on the access time more strongly than the access location, and the difference in the results among the URL categories was stronger during the day than at night. As shown in Fig. 7, the average total delay of non-CDN objects of about 20% to 80% of websites was 500 milliseconds or more. We confirmed that the ratio of websites whose average total delay exceeded 500 milliseconds decreased to about 5% to 20% at night and about 2% to 40% during the day by delivering objects using CDNs.

We also observed the tendency in which the average total delay of CDN objects of entertainment websites, e.g., Games, was large at night, and that of business-related websites, e.g., Business and Reference, was large during the day. Because kids are mainly active during the day, the average total delay of CDN objects of Kids & teens websites was also large during the day. The demand of websites of each URL category during the day is difference from that at night, and the total delay of objects of each URL category depends on the time of day, so differentiating the priority based on the URL categories in the cache control, i.e., replacement of cached objects, will be effective. For example, it will be preferable to give high priority to entertainment websites at night and to business-related websites during the day.



(a) Massachusetts, midnight
(b) Japan, midnight
(c) Massachusetts, noon
(d) Japan, noon

Fig. 7. Complementary cumulative distribution (CCD) of average total delay of non-CDN objects at two PlanetLab hosts



(a) Massachusetts, midnight
(b) Japan, midnight
(c) Massachusetts, noon
(d) Japan, noon

Fig. 8. CCD of average total delay of CDN objects at two PlanetLab hosts

E. Main Findings

We summarize the main findings obtained from the experiments in accessing websites from various probing locations as follows.

- The ratio of CDN objects against non-CDN objects was independent of access location. However, the ratio of CDN objects largely differed among the URL categories, and this ratio varied between about 0.2 and 0.5.
- Many original objects of Society, Health, Adult, Recreation, and Sports websites are provided in North America. The content of these categories has weak geographical locality, and identical content provided in North America tends to be viewed by users worldwide. On the other hand, many content items of Home and Shopping websites have strong geographical locality, and the original objects of these websites tend to be provided in all

regions, so different content items that are unique in each region are provided in each region. The geographical tendency of original objects of other categories, such as Computers, News, Reference, Science, Arts, Games, and Kids & teens, is moderate between these two extreme cases, and the original objects of these categories tend to be provided in North America, Europe, and Asia.

- The geographical-deployment patterns of CDN caches are not identical, and various deployment patterns of CDN caches exist. The first pattern is the placing of caches over many regions, i.e., North America, Europe, Asia, and Oceania; whereas, the second pattern is the providing of caches in North America and Europe, and the third pattern is the placing of caches in just Oceania. More than 80% of websites of all the categories except Adult use the CDNs of the first cache-deployment pattern, and objects are delivered from caches deployed in many regions. Adult websites were more likely to use the CDNs of the second cache-deployment pattern.
- The demand of websites of each URL category during the day is different from that at night, and the total delay of objects of each URL category depends on the time of day, so differentiating the priority based on the URL categories in the cache control, i.e., replacement of cached objects, will be effective. For example, it will be preferable to give high priority to entertainment websites at night and to business-related websites during the day.

V. Conclusion

The communication structure of traffic generated when accessing modern websites has become more complex. Various techniques have been used to effectively reduce web response time and suppress network traffic volume, and one of the most widely used techniques involves CDNs. Although the use of CDNs is assumed to reduce web response time, the actual effect of CDNs on the reduction of web response time has not been clarified. To improve the response time and suppress the amount of traffic in networks, it is preferable to design cache control methods, i.e., object deployment and cache replacement, based on the communication structure of web traffic. To answer this fundamental question, we analyzed the communication patterns of non-CDN objects and CDN objects generated when accessing various websites from 12 probing client terminals worldwide and investigated the geographical distribution of original objects and CDN caches. Although we analyzed only the home pages of popular websites, we plan to evaluate the geographical distribution of original objects and CDN caches when accessing other pages reachable from the home page and the change in the communication structure over one-day and one-week periods in the future.

References

[1] B. Ager, W. Muhlbauer, G. Smaragdakis, and S. Uhlig, Web Content Cartography, ACM IMC 2011.

[2] D. Arthur and S. Vassilvitskii, k-means++: the advantages of careful seeding, ACM SODA 2007.

[3] Alexa, http://www.alexa.com/topsites/category.

[4] R. Baeza-Yates, C. Castillo, E. N. Efthimiadis, Characterization of national Web domains, ACM Trans. Internet Technology (TOIT), 7(2), Article No. 9, 2007.

[5] L. Bent, M. Rabinovich, G. M. Voelker, Z. Xiao, Characterization of a Large Web Site Population with Implications for Content Delivery, ACM WWW 2004.

[6] P. Borgnat, G. Dewaele, K. Fukuda, P. Abry, and K. Cho, Seven Years and One Day: Sketching the Evolution of Internet Traffic, IEEE INFOCOM 2009.

[7] M. Butkiewicz, H. V. Madhyastha, and V. Sekar, Understanding Website Complexity: Measurements, Metrics, and Implications, ACM IMC 2011.

[8] J. Erman, V. Gopalakrishnan, R. Jana, and K. Ramakrishnan, Towards a SPDY′ier Mobile Web?, ACM CoNEXT 2013.

[9] GitHub, Network Monitoring, http://github.com/ariya/phantomjs/wiki/Network-Monitoring

[10] P. Gill, M. Arlitt, N. Carlsson, and A. Mahanti, Characterizing Organizational Use of Web-based Services: Methodology, Challenges, Observations, and Insights, ACM Trans. The Web, 5(4), Article No. 19, 2011.

[11] When seconds count. http://www.gomez.com/wp-content/downloads/GomezWebSpeedSurvey.pdf.

[12] S. Ihm and V. Pal, Towards Understanding Modern Web Traffic, ACM IMC 2011.

[13] A. L. Jain and R. C. Dubes, Algorithms for Clustering Data, NJ Prentice-Hall, 1988.

[14] M. Karlsson and M. Mahalingam, Do We Need Replica Placement Algorithms in Content Delivery Networks?, WCW 2002.

[15] C. Labovitz, S. Iekel-Johnson, J. Oberheide, and F. Jahanian, Internet Inter-Domain Traffic, ACM SIGCOMM 2010.

[16] MaxMind, GeoIP Downloadable Databases, http://dev.maxmind.com/geoip/downloadable.

[17] J. Mickens, Silo: Exploiting JavaScript and DOM Storage for Faster Page Loads, USENIX WebApps 2010.

[18] E. Nygren, R. Sitaraman, and J. Sun, The Akamai Network: A Platform for High-Performance Internet Applications, ACM SIGOPS 2010.

[19] J. Odvarko, HAR Viewer, Software is hard, http://www.softwareishard.com/blog/har-viewer.

[20] J. Ott, M. Sanchez, J. Rula, F. Bustamante, Content Delivery and the Natural Evolution of DNS, ACM IMC 2012.

[21] PlanetLab, https://www.planet-lab.org/

[22] G. Podjarny, Not as SPDY as You Thought, http://www.guypo.com/technical/not-as-spdy-as-you-thought/

[23] Quantcast, http://www.quantcast.com/top-sites-1.

[24] J. Ravi, Z. Yu, and W. Shi, A survey on dynamic Web content generation and delivery techniques, Elsevier J. Network and Computer Applications, 32(5), pp. 943-960, 2009.

[25] F. Schneider, S. Agarwal, T. Alpcan, and A. Feldmann, The new web: characterizing AJAX traffic, ACM PAM 2008.

[26] A. Sharma, A. Venkataramani, and R. Sitaraman, Distributing Content Simplifies ISP Traffic Engineering, ACM SIGMETRICS 2013.

[27] S. Sivasubramanian, G. Pierre, M. Steen, and G. Alonso, Analysis of Caching and Replication Strategies for Web Applications, IEEE Internet Computing, 11(1), pp. 60-66, 2007.

[28] S. Souders, High Performance Web Sites: Essential Knowledge for Front-End Engineers, O′Reilly Media, 2007.

[29] SPDY: An experimental protocol for a faster web, http://www.chromium.org/spdy/spdy-whitepaper.

[30] Squid cache replacement policy, http://www.squid-cache.org/Doc/config/cache_replacement_policy/.

[31] A. Su, D. Choffnes, A. Kuzmanovic, and F. Bustamante, Drafting Behind Akamai: Inferring Network Conditions Based on CDN Redirections, ACM Trans. Networking, 17(6), pp. 1752-1765, 2009.

[32] S. Sundaresan, N. Feamster, R. Teixeira, and N. Magharei, Characterizing and Mitigating Web Performance Bottlenecks in Broadband Access Networks, ACM IMC 2013.

[33] X. Wang, A. Balasubramanianm A. Krishnamurthy, and D. Wetherall, Demystifying Page Load Performance with WProf, NSDI 2013.

[34] X. Wang, A. Balasubramanianm A. Krishnamurthy, and D. Wetherall, How speedy is SPDY?, NSDI 2014.