社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

# Web コンテンツの CDN 利用状況とオリジナル配置傾向に関する分析

上山　憲昭[†,††]　中野　雄介[†,††]　塩本　公平[††]
長谷川　剛[†††]　村田　正幸[†]　宮原　秀夫[†]

† 大阪大学大学院情報科学研究科　〒 565-0871 大阪府吹田市山田丘 1–5
†† 日本電信電話株式会社 NTT ネットワーク基盤技術研究所　〒 180–8585 東京都武蔵野市緑 3–9–11
††† 大阪大学サイバーメディアセンター　〒 560-0043 大阪府豊中市待兼山町 1–32
E-mail: †kamiyama.noriaki@ist.osaka-u.ac.jp

**あらまし**　近年の Web サイトは，Ajax 等によって動的に生成されたオブジェクトを様々な配信ホストから取得する傾向を強めており，Web サイト閲覧時に発生する通信パタンが複雑化している．オブジェクトの配信には，ユーザの近くに存在するキャッシュサーバからコンテンツを配信する CDN(contents delivery network) が用いられることが多い．CDN を用いることで Web 閲覧時の応答時間の低減が可能となるが，CDN の効果を高めるには，オブジェクトの地理的配置に基づき適切にキャッシュを制御する必要がある．しかしオブジェクトの地理的配置傾向についての分析は見られない．そこで本稿では，PlanetLab を用いて世界の 12 の拠点から，アクセス頻度の多い約 1,000 の Web サイトにアクセスしたときに発生する通信パタンを測定し，サーバ距離，遅延時間，オブジェクト数といった各種特性値の地域的な傾向について分析し，オリジナルオブジェクトの地域的な配置傾向や，CDN を用いて配信されるオブジェクトの比率に関する傾向を明らかにする．例えば，夜間には娯楽性の高いサイトを日中はビジネスに関連するサイトを優先的にキャッシュするといった制御が望ましい等の知見を得た．
**キーワード**　Web，CDN，アクティブ測定

# Analizing Geographical Usage of CDN and Original Content of Website

Noriaki KAMIYAMA[†,††], Yusuke NAKANO[†,††], Kohei SHIOMOTO[††],
Go HASEGAWA[†††], Masayuki MURATA[†], and Hideo MIYAHARA[†]

† Department of Information Science, Osaka University　1-5, Yamadaoka, Suita, Osaka 565-0871
†† NTT Network Technology Laboratories, NTT Corporation　3-9-11, Midori, Musashino, Tokyo 180-8585
††† Cybermedia Center, Osaka University　1-32, Machikaneyama, Toyonaka, Osaka 560-0043
E-mail: †kamiyama.noriaki@ist.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 objects of websites are delivered using content delivery networks (CDNs), in which data objects are delivered from cache servers located close to user terminals. Although the web response time has been expected to be reduced by using CDNs, the actual effect of CDNs on the reduction of web response time depends on the geographical deployment of website objects. However, no works have investigated the tendency of geographical deployment of website objects. In this paper, to answer this fundamental question, we measure the communication structure of traffic generated when accessing about the 1,000 most popular websites from 12 locations in the world, and we obtain various findings. For example, it will be desirable to give high priority to entertainment websites at nighttime, whereas give high priority to business-related websites at daytime.
**Key words**　Web, CDN, active measurement

## 1. はじめに

近年，インターネットのトラヒックの多くの部分を，Web サービスで用いられる HTTP トラヒックが占めている．例えば 2006 年〜2008 年の間に日米間のバックボーンリンクで測定されたトラヒックの分析によると，HTTP パケットが約 60%を占めている [6]．しかし 2/3 のユーザは毎週のように Web サイト閲覧時の低速性を経験しており [7]，17%のユーザは待ち時間が 5 秒を超えた場合には閲覧を諦めるといった報告もなされている [11]．また 400 ミリ秒の遅延により Google サーチエンジンでの検索回数が 0.74%低減することや [33]，Web 応答時間が 0.1 秒だけ削減するごとに Amazon の利益が 1%増加することが報告されており [34]，また高速に表示される Web サイトは，ユーザが購買を完了する回数が 15%も多く，また 1 ページだけ閲覧した後にサイトから離脱する回数が 9%も少ないことが報告されている [19]．そのためインターネットの主流サービスである Web トラヒックを適切に制御することが，ユーザの体感品質を向上させ，ネットワーク資源の消費量を抑えるためには重要である．

従来の Web サイトは静的なテキストや画像といったオブジェクトがサーバに用意され，Web ブラウザは HTTP を用いてこれら静的オブジェクトを単にダウンロードして表示していた．しかし近年，クライアント PC からのリクエスト受信時に，サーブレットや JSP(Java server pages) のプログラムをサーバ側で

実行するか，JavaScript で書かれた Ajax や DOM(document object model) によるプログラムを HTML に埋め込みクライアント PC 側で実行することで生成される動的オブジェクトの割合が増加している [7]．また，広告を専用のサーバから取得するなど，各オブジェクトの配信元が多元化している．このように一つの Web サイトを構成するオブジェクトは複雑性を増している．Web 閲覧時のユーザのレスポンス時間を低減する技術としては CDN(contents delivery network) が一般的であり [21] [32]，アクセス数上位 1,000 のサイトの中では 74%が CDN を利用している [21]．CDN は主に Akamai 等の CDN 事業者が運営してきたが，近年，Google 等の大規模コンテンツプロバイダや，AT&T 等の Tier-1 ISP が自身で CDN を運用するケースも増えてきており [16]，CDN の提供形態が多様化している．

CDN を用いることで Web 応答時間の改善が期待されるが，CDN の効果を高めるには，オブジェクトの地理的配置に基づき適切にキャッシュを制御する必要がある．しかしオブジェクトの地理的配置傾向についての分析は見られない．応答時間を改善し，ネットワーク内を流れるトラヒック量を低減するためには，Web トラヒックの通信構造に基づきオブジェクトのキャッシュ位置の決定やキャッシュ置換といった制御を適切に行う必要があることから，Web トラヒックの通信構造を明らかにすることが重要である．そこで筆者らは，以前，PlanetLab を用いて世界中の 12 の地点から約 1,000 の高人気サイトを閲覧したときに発生するトラヒックを測定することで，Web トラヒックの通信構造の傾向について明らかにした [14]．しかし CDN を利用して配信されるオブジェクトを区別しないで分析していたため，オリジナルオブジェクトや CDN のキャッシュサーバの地理的な配置傾向については明らかにされていない．そこで本稿では，CDN の利用の有無によりオブジェクトを分離した分析を行うことで，これら傾向について明らかにする．2 節で既存の Web トラヒック測定手法に関して簡単にまとめた後，3 節にて Web トラヒックの通信構造の測定分析法について述べる．そして 4 節にて Web トラヒックの測定実験結果について述べ，最後に 5 節にて全体をまとめる．

## 2. 関 連 研 究

アクティブ測定による Web トラヒック分析に関する研究としては，Baeza-Yates らの，2004 年前後に 12 か国で実施した Web クローリングの測定結果から，国ごとの Web 閲覧トラヒックのサイズ，接続グラフの次数等の各種傾向を比較した研究や [4]，Butkiewicz らの，ランダムに選択した Web サイトを 9 週間にわたり周期的にアクセスし，構成オブジェクト数やアクセスサーバ数等の指標について分析した研究が見られる [7]．

一方，パッシブ測定による Web トラヒック分析に関する研究としては，Ihm らの，2006 年〜2010 年の 5 年間にわたる Web の proxy access log を用いて Web トラヒックの各種指標の変移を分析した研究や [12]，Bent らの，2004 年の 1 日のパケットキャプチャデータから，Web サイトの Cookiee 利用頻度等について分析した研究が見られる [5]．また Gill らは，企業や大学からの Web アクセストラヒックを分析し，Web サービスの利用傾向について明らかにしており [10]，Ager らは CDN やデータセンタ上のコンテンツ配置や，配信元サーバの選択がどのように行われているかを，DNS に関する制御パケットの測定と，BGP routing table の snapshot に基づき識別することを検討している [1]．また Schneider らはパケットキャプチャデータから HTTP や AJAX セッションを抽出し，生成されるトラヒックパタンの差異を分析している [26]．しかしこれらの研究では，Web サイトにアクセスしたときの，オブジェクトの配信元サーバとクライアント端末間の距離といった，地理的な通信構造については分析されていない．

## 3. Web トラヒックの測定分析の手順

本節では，多地点からの Web 通信構造分析のための測定実験手順について述べる．測定手順は，(i)PlanetLab 上での測定環境の構築と測定地点の選択，(ii) 評価 URL リストの生成，(iii) 各測定地点から各評価 URL にアクセスしたときの HAR(HTTP Archive) ファイルの取得，(iv)HAR ファイルからのデータ抽出，(v)RTT の測定，の 5 つの手順で構成される．以下に，各々の手順について述べる．

### 3.1 PlanetLab 上での測定環境の構築と測定地点の選択

PlanetLab はインターネット上に構築されたオーバレイネットワークで，世界の様々な地域に存在する約 500 のノードから構成される．PlanetLab を用いることで，選択したノード上で様々なプログラムを実行することができる．そのため (ii) 以降の手順を PlanetLab 上の複数のノードで独立に実行することで，世界中の様々な地域から様々な Web サイトにアクセスし，通信特性の情報を収集する．実験に先立ち，PlanetLab 上での測定実験環境を構築する必要があるが，PlanetLab が提供する GUI を用いて測定に用いるノードを起動する．北米 (NA) から三つ，欧州 (EU) から二つ，ロシア (RU) から一つ，オセアニア (OA) から二つ，南米 (SA) から二つ，アジア (AS) から一つ，そしてアフリカ (AF) から一つの，合計で 12 の PlanetLab ノードを測定ホストとして選択した．これら 12 の測定地点を表 1 にまとめる．

表 1　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 |

### 3.2 評価 URL リストの生成

Web 通信構造の傾向を分析するためには，アクセス数の多い，高人気のサイトにアクセスしたときに発生する通信を分析対象とすることが望ましい．そこで Alexa のサイト [3] 上で公開されているアクセスランキングを元に，表 2 に示す 16 の各 URL カテゴリから，最もアクセス数の多い上位 300 の Web サイトを測定対象として選択した．いくつかの Web サイトは複数のカテゴリに重複して分類されているため，重複したサイトを削除することで 4,290 の Web サイトを測定対象に選択した．

表 2　Website count used in clustering analysis, average object size, object count, and total data size in each URL category

| ID | Category | Website count | | Object size | Object count | Total size |
|---|---|---|---|---|---|---|
| | | 0:00 | 12:00 | (kbytes) | | (Mbytes) |
| C1 | Business | 59 | 40 | 14.70 | 55.14 | 0.810 |
| C2 | Computers | 112 | 91 | 16.26 | 43.63 | 0.709 |
| C3 | News | 39 | 27 | 13.55 | 72.45 | 0.982 |
| C4 | Reference | 112 | 109 | 13.09 | 43.42 | 0.568 |
| C5 | Regional | 80 | 73 | 17.77 | 50.59 | 0.899 |
| C6 | Science | 95 | 86 | 14.04 | 52.86 | 0.742 |
| C7 | Society | 79 | 83 | 15.01 | 66.86 | 1.003 |
| C8 | Health | 86 | 52 | 14.27 | 54.30 | 0.775 |
| C9 | Home | 85 | 47 | 15.66 | 55.39 | 0.867 |
| C10 | Shopping | 69 | 68 | 15.67 | 70.77 | 1.109 |
| C11 | Adult | 112 | 102 | 10.49 | 53.04 | 0.557 |
| C12 | Arts | 55 | 60 | 15.43 | 68.18 | 1.052 |
| C13 | Games | 87 | 58 | 15.28 | 54.12 | 0.827 |
| C14 | Kids & teens | 106 | 64 | 13.23 | 54.59 | 0.722 |
| C15 | Recreation | 86 | 52 | 13.55 | 57.30 | 0.776 |
| C16 | Sports | 38 | 53 | 16.62 | 86.67 | 1.440 |

### 3.3 評価 URL アクセス時の HAR ファイルの取得

生成した評価 URL リストの各 URL に対して，測定用 PlanetLab ホストから GET の HTTP リクエストを送信した際に発生する通信特性を，HAR(HTTP Archive) ファイルとして取得した [20]．HAR ファイルは，クライアント PC とサーバ間で転送される HTTP データのヘッダ情報から，クライアント PC において，各オブジェクトのサーバ URL，サイズ，取得に要した遅延時間等の各種通信特性を算出し，JSON(JavaScript Object Notation) 形式で出力したものである．HAR ファイルは NetExport の拡張を施した Firebug を適用した Firefox を用いたり，Google Chrome の Developer Tools を用いたりすることで取得可能である．しかし個別に手動で各測定 URL にアクセスして HAR ファイルを取得すると，評価できるサイト数が非常に限られたものとなる．そこで多数のサイトにアクセスするために，コマンドラインで JavaScript を実行できる phantomjs 上で動作するスクリプト netsniff.js [9] を用いることで，多数のサイトに連続してアクセスし，各々の HAR ファ

イルをバッチ処理で取得した．この際に，測定用 PlanetLab ホストのローカルキャッシュを無効化することで，全てのオブジェクトをリモートのサーバから取得した．

Web サイトにアクセスする時間帯によって，生じる通信特性が異なることが予想されるため，様々な測定地点間で Web 通信構造の傾向を比較するためには，全ての測定地点において同一の現地時刻に開始する必要がある．そこで UNIX の cron コマンドを用いて，UTC(coordinated universal time) より取得した各測定地点の現地時刻が midnight(0:00) と noon(12:00) となるときに各々，各測定用 PlanetLab ノードから 4,290 の Web サイトに連続してアクセスする実験を開始した．12 の全ての測定地点において HAR ファイルが正しく取得されたサイト (midnight で 1,124，noon で 927) を最終的に分析対象とした．表 2 に，各 URL カテゴリの分析対象 Web サイト数をまとめる．

### 3.4 HAR ファイルからのデータ抽出

取得された各 HAR ファイルに含まれる各構成オブジェクトの情報から，分析に必要となるデータを各測定用 PlanetLab ノードにて抽出する．具体的には，ホスト名，ホストの位置，オブジェクトサイズ，オブジェクト取得遅延時間，オブジェクト種別 (MIME Type) に関する情報を取得するため，HAR ファイル中の各 key に対応する value を Python で抽出した．オブジェクト取得遅延時間は，各オブジェクトに対する Request が測定用ノードから送信開始された時刻から，そのオブジェクトの測定用ノードでの到着が完了した時刻までの経過時間である．なお，MaxMind の提供する GeoIP API [17] を用いて，URL からサーバの存在する国名・都市名・位置座標を取得し，各オブジェクトの配信元サーバの位置座標と測定用 PlanetLab ノード間のユークリッド距離をオブジェクト距離と定義して算出した．

### 3.5 RTT の測定

オブジェクト距離は測定ノードとオブジェクトサーバ間のユークリッド距離であり，インターネット上での実際の距離とは異なる．そこで前ステップで述べた総計データに加えて，PlanetLab の各測定ノードにおいて各 Web サイトにアクセスして HAR ファイルを取得した直後に，各オブジェクトサーバに対して測定ノードから ping コマンドを送ることで RTT(round-trip time) を測定した．

### 3.6 CDN の利用の有無によるオブジェクトの分類

CDN を用いて配信されたオブジェクト (以後，CDN オブジェクトと表記) と，CDN を用いないで配信されたオブジェクト (以後，non-CDN オブジェクトと表記) を区別して，各々の通信パタンの傾向を分析することで，CDN のキャッシュサーバやオリジナルオブジェクトの地理的な配置傾向を分析可能となる．そこで本稿では，CDN オブジェクトを配信したホストのセカンドレベルのドメイン名のリストを作成することで，HAR ファイルから抽出されたオブジェクトを，これら CDN オブジェクトと non-CDN オブジェクトの二つのグループに分類する．

まず，様々な CDN 事業者の Web サイトを調べることで，edgesuite.net，cloudfront.net，akamaiedge.net などの，CDN 事業者の 44 のセカンドレベルドメイン名を含むリストを作成した．オブジェクトの HAR ファイルから抽出されたドメイン名は，www.yahoo.com 等のコンテンツプロバイダのドメイン名であり，host1.akamaiedge.net 等の実際にオブジェクトを配信したホストのドメイン名とは異なる．そこで Linux の dig コマンドを用いて，オブジェクトを実際に配信したホストのドメイン名を取得する．そして事前に作成した CDN 事業者のセカンドレベルドメイン名リストと照合し，本リストに含まれるものを CDN オブジェクトに，含まれないものを non-CDN オブジェクトに分類する．

### 3.7 Web サイトのクラスタ分析

様々な地域からアクセスされたときの各特性値の傾向の違いを明らかにするため，We サイトの各特性値の地理的傾向に基づいたクラスタ分析を行う．図 1 に示すように，各測定時刻 $t$ において $N$ 個の様々な特定地点 $X_1$，$X_2$，$\cdots X_N$ から，$M$ 個の様々な Web サイト $Y_1$，$Y_2$，$\cdots Y_M$ にアクセスしたとき，各 Web サイトに対して平均 RTT 等の各特性値の $N$ 個の結果が取得される．よって，時刻 $t$ に Web サイト $y$ に測定地点 $k$ よりアクセスしたときの特性値を $v_{y,t,k}$ $(1 \leqq k \leqq N)$ とするとき，$v_{y,t,k}$ を要素にもつ $N$ 次元のベクトル $\boldsymbol{v}(y,t)$ を構成することができ，$1 \leqq y \leqq M$ の各 $y$ に対して，$M$ 個のベクトル $\boldsymbol{v}(y,t)$ が得られる．各特性値 $v$ のアクセス地点ごとの傾向の違いを分析するために，得られた $M$ 個のベクトル $\boldsymbol{v}(y,t)$ を用いて，k-menas 法を用いてクラスタ分析を行う．

k-means 法の分類結果は初期クラスタに強く依存するため，最適なクラスタ構成が生成されるよう初期クラスタを構成することが重要となるが，本稿では Arthur らによって提案された k-means++法を用いて初期クラスタを構成する [2]．k-means++法は，$k$ 個の初期クラスタの重心をできるだけ偏りなく散らばらせるものであり，最初に一つのメンバをクラスタ重心としてランダムに選択し，以後，残るメンバの各々の最近接クラスタ重心までの距離の自乗に比例する確率でランダムに一つのメンバをクラスタ重心として選択する処理を，$k$ 個のクラスタ重心が選択されるまで反復する．ところで各特性値 $v$ のクラスタ重心は，そのクラスタを構成する全てのメンバの $v$ の平均値ベクトルであり．各要素は，構成メンバの各測定地点における $v$ の平均値を表す．

k-means 法の結果はクラスタ数 $k$ にも強く依存するため，$k$ を適切に設定することが重要となる．本稿では，Jain らによって提案された JD 法を用いてパラメタ $k$ を最適設計する [13]．各クラスタに属するメンバとそのクラスタ重心間の距離を最小化するのと同時に，任意の二つのクラスタの重心間の距離を最大化するために，JD 法は次式で定義されるコスト関数 $p(k)$ を最小化する $k$ を選択する．

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

ただし，$\eta_j = \sum_{i=1}^{n_j} D(\boldsymbol{x}_i^{(j)}, \boldsymbol{m}_j)/n_j$，$\xi_{ij} = D(\boldsymbol{m}_i - \boldsymbol{m}_j)$ であり，$n_j$ はクラスタ $j$ に分類されたメンバ数で，$D(\boldsymbol{a}-\boldsymbol{b})$ は二つのベクトル $\boldsymbol{a}$ と $\boldsymbol{b}$ との間の距離である．そして $1 \leqq k \leqq 1+\log_2 n$ の範囲で $p(k)$ を最小化する $k$ を選択する．

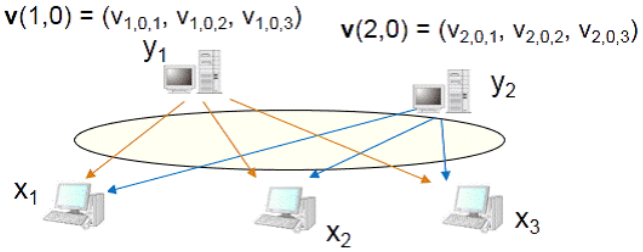


図 1 Clustering websites on basis of location properties

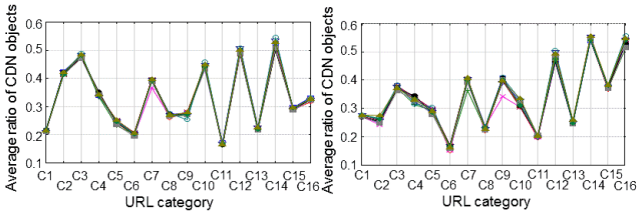


図 2 Average ratio of objects delivered using CDN in each URL category at each access location

## 4. 実験結果

### 4.1 平均特性

表 2 に，各 URL カテゴリの Web サイトの，平均オブジェクトサイズ (kbytes)，オブジェクト数，総データサイズ (Mbytes) をまとめる．ただしこれらの値は 12 の全測定地点における平均値である．総データ量とオブジェクト数は，Arts，Shopping，Sport といった娯楽系サイトで多い傾向が見られるのに対して(注1)，Business，Computers，Health，Reference といった情報収集系サイトで少ない傾向が見られる．アクセス都市数，アクセスホスト数に関しても同様の傾向が得られた．

(注1)：Butkiewicz らの結果と一致する [7].

また図 2(a) に 0:00 において 12 の各測定地点からアクセスしたときの，各 URL カテゴリを構成する Web サイトにおいて CDN オブジェクトの占める割合の平均値をプロットする．同様に図 2(a) には 12:00 にアクセスした際の結果を示す．12 のほぼすべての地点において，各 URL カテゴリの CDN オブジェクト数の平均比率はほとんど同一の値となった．一方，16 の URL カテゴリ間では平均 CDN オブジェクト数比率は大きく異なり，0.2 から 0.5 程度の値をとる．Computers，News，Society，Shopping，Arts，Kids & teens といったサイトは CDN オブジェクトの構成比率が高いのに対して，Business，Regional，Science，Adult，Games といったサイトは CDN オブジェクトの構成比率が低い傾向が見られる．

### 4.2 オリジナルオブジェクトの地理的な配置傾向

本節では，non-CDN オブジェクトの平均距離と平均 RTT に対して各カテゴリの Web サイトをクラスタ分析することで，オリジナルオブジェクトの地理的な配置傾向を分析する(注2)．図 3(a) に，midnight の測定データを用いた non-CDN オブジェクトの平均距離に関するクラスタ分析の結果，生成された各クラスタの重心を，12 の各測定地点に対して示す．3.7 節で述べたように，Web サイトを 12 の各地点における測定値を用いてクラスタ分析を行ったため，各クラスタに分類された Web サイトの各地点における平均オブジェクト距離の平均値が重心の各要素となる．ただしクラスタ番号は意味をもたないため，ここでは構成 Web サイト数の降順にクラスタ番号を付与した．各クラスタの重心は，そのクラスタに分類された Web サイトを構成するオリジナルオブジェクトの各測定地点からの平均距離に関する傾向を表す．また図 3(b) に，16 の各カテゴリ (C1, C2, ···, C16) の Web サイトの中で各クラスタに分類されたものの比率と，全 Web サイト (All) の中で各クラスタに分類されたものの比率をプロットする．

オリジナルオブジェクトの地理的配置傾向は三つのクラスタに分類された．約 80%の Web サイトは二つの最大クラスタ (Cluster 1 と Cluster 2) に分類されたが，これら二つのクラスタは共に，北米からは近く，欧州と南米とロシアとアフリカからは中程度の距離で，オセアニアとアジアからは遠いという特徴を有する．よって様々なカテゴリの Web サイトのオリジナルオブジェクトの多くは北米に存在し，北米に多くのコンテンツプロバイダが存在することが予想される．一方，残る Cluster 3 はオリジナルオブジェクトが欧州のみに存在する傾向が見られる．URL カテゴリ間のオリジナルオブジェクトの地理的配置傾向の差異は小さい．

図 3(c) と図 3(d) に同様に，noon の測定データを対象としたクラスタ分析結果をプロットする．やはり 3 つのクラスタが生成されており，noon の Cluster 1 と Cluster 2 は各々，night の Cluster 1 と Cluster 3 に相当する．オリジナルオブジェクトがアジアとオセアニアで提供される傾向の高いサイトが分類されている Cluster 3 が新たに出現しているが，ごく少数の Web サイトのみが Cluster 3 に分類されている．

ユークリッド距離を評価することで，二つのノード間の物理的な距離をおおまかに見積もることが可能であるが，パケットが転送される総延長距離といったネットワーク上の距離は，からずしもユークリッド距離とは一致しない．そこで測定ホストから non-CDN オブジェクトの配信サーバまでの実測 RTT の平均値に基づき Web サイトをクラスタ分析する．図 4(a) と図 4(b) に各々，midnight データを対象に，non-CDN オブジェクトの平均 RTT に基づき生成された各クラスタの重心と，各クラスタに分類された各カテゴリの Web サイトの比率をプロットする．Web サイトは 4 つのクラスタに分類されており，平均距離とは異なり，平均 RTT については URL カテゴリ間の傾向の違いが確認された．

Cluster 1 に分類された Web サイトのオリジナルオブジェクトは北米からアクセスしたときにのみ RTT が小さく，Society，Adult，Recreation，Sports といったサイトの分類比率が高い．これらカテゴリの Web サイトは北米に存在するコンテンツプロバイダから提供される傾向が強い．Cluster 2 の重心は，北米，欧州，アジアにおいて小さく，Computers，News，Reference，Science，Arts，Games，Kids & teens といったカテゴリの Web サイトの分類比率が高い．これらカテゴリのサイトの多くのコンテンツは北米や欧州やアジアで提供されていることが確認できる．例えば Games や Kids & teens の Web サイトの多くは，これら分野に強い日本のコンテンツプロバイダによって提供されていることが予想される．Cluster 1 と 2 に分類された多くの Web サイトの地域性は低く，同一のコンテンツが世界の様々な地域で閲覧される傾向が高い．

一方，Cluster 3 の重心はアフリカを除く全ての地域で小さく，Home や Shopping といった Web サイトの多くが本クラスタに分類されている．これらカテゴリのサイトの地域性は高く，各々の地域で固有のコンテンツが提供される傾向が見られ，各々の地域において近隣に存在する配信サーバからオリジナルオブジェクトが取得される傾向が確認される．最後に Cluster 4 に分類された Web サイトのオリジナルオブジェクトを提供する配信サーバまでの平均 RTT は欧州においてのみ小さく，全てのカテゴリにおいて 10%程度未満の Web サイトのみが本クラスタに分類された．

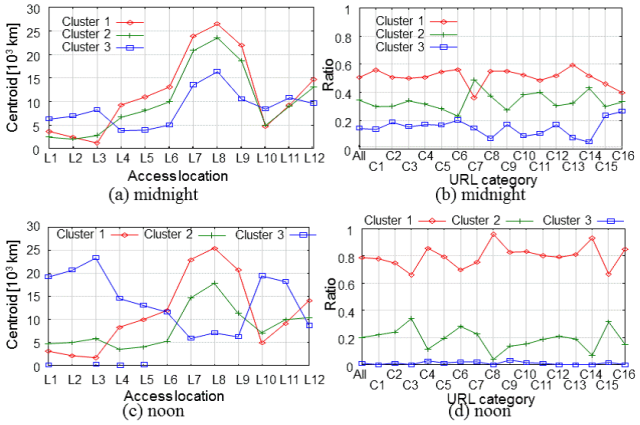


図 3 (a)(c) Centroids of average distance of non-CDN objects at each access location, (b)(d) ratio of websites classified into each cluster in each website category

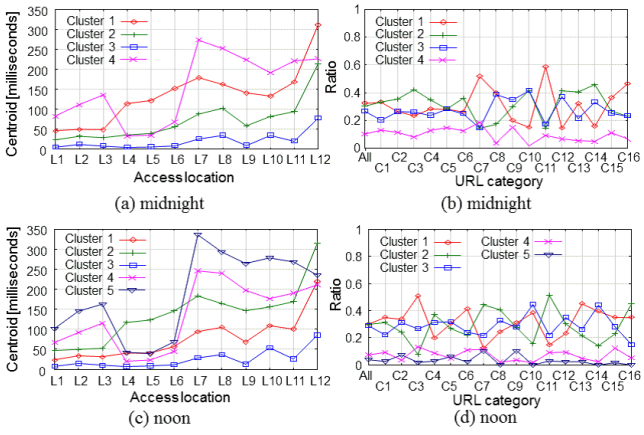


図 4 (a)(c) Centroids of average RTT of non-CDN objects at each access location, (b)(d) ratio of websites classified into each cluster in each website category

図 4(c) と図 4(d) に各々，noon データを対象としたクラスタリング結果を同様にプロットする．Web サイトは 5 つのクラスタに分類され，Cluster 1，2，3 が各々，midnight データにおける Cluster 2，1，3 に相当する．また noon データの Cluster 4 と 5 は，midnight データの Cluster 4 に相当し，noon データの各クラスタの重心に対しても midnight データと同様の傾向が見られる．

### 4.3 CDN のキャッシュサーバの地理的な配置傾向

本節では，CDN オブジェクトを対象とした平均距離と平均 RTT に基づくクラスタ分析により，CDN のキャッシュサーバの地理的な配置傾向を分析する．図 5(a) と図 5(b) に，midnight データを対象に生成された各クラスタに分類された Web サイ

(注2)：non-CDN に分類されたオブジェクトの中には，分類に用いた CDN 事業者のドメイン名リストに含まれていない CDN 事業者のキャッシュから配信されたオブジェクトが存在する可能性があるが，本稿では non-CDN のオブジェクトの多くがオリジンサーバから配信されていることを想定する．

トの CDN オブジェクトの平均距離に関する重心と，各クラスタに分類された Web サイトの比率を各々プロットする．Web サイトは 5 つのクラスタに分類されており，CDN のキャッシュサーバの地理的な配置パタンは同一ではなく，様々な配置パタンが存在する．Cluster 1 の重心はアジアとアフリカを除く全ての地域で小さく，News，Reference，Regional，Health，Kids & teens のサイトの分類比率が高い．Cluster 2 の重心は北米，欧州，南米において小さく，Home，Shopping，Arts，Recreation といったサイトの分類比率が高い．残る他の三つのクラスタに分類された Web サイトにおける平均距離は北米において小さく，Business，Computers，Society，Adult といったサイトの分類比率が高い．

図 5(c) と図 5(d) に，noon データに対する同様の結果を示す．noon の Cluster 2 は midnight の Cluster 1 に相当し，News，Reference，Arts といったサイトの分類比率が高い．Cluster 3 は midnight の Cluster 2 に相当し，Home，Shopping，Recreation の分類比率が高い．残る三つのクラスタ，Cluster 1，Cluster 4，Cluster 5 は各々，midnight の Cluster 3，Cluster 4，Cluster 5 に相当し，Society，Health，Adult といったサイトの分類比率が高く，北米にキャッシュサーバが展開された CDN を用いて配信される傾向が高い．

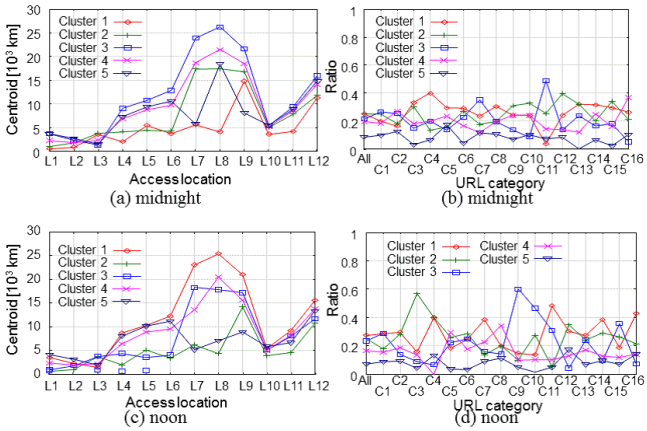


図 5　(a)(c) Centroids of average distance of CDN objects at each access location, (b)(d) ratio of websites classified into each cluster in each website category

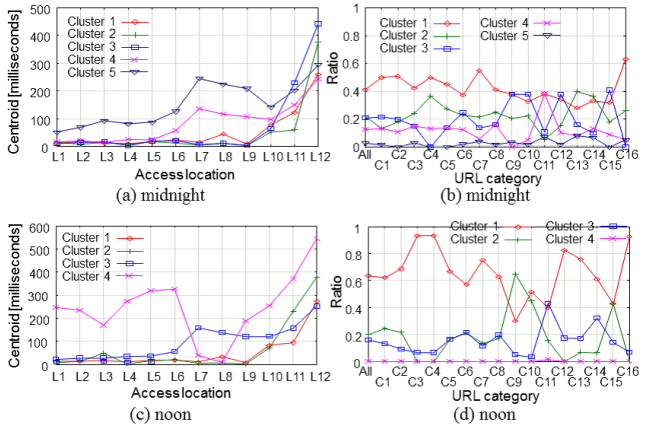


図 6　(a)(c) Centroids of average RTT of CDN objects at each access location, (b)(d) ratio of websites classified into each cluster in each website category

図 6(a) と図 6(b) に各々，CDN オブジェクトの平均 RTT に基づくクラスタ分析で生成された各クラスタの重心と，各カテゴリのサイトの各クラスタへの分類比率をプロットする．Web サイトは 5 つのクラスタに分類されており，やはり CDN のキャッシュサーバの地理的な配置パタンは同一ではなく，様々な配置パタンが存在することが確認できる．Cluster 1 と 2 と 3 の重心は南米とアフリカを除く全ての地域で小さく，Adult を除く全てのカテゴリの 80%以上もの多数のサイトがこれら三つのクラスタのいずれかに分類された．他の二つのクラスタ，Cluster 4 と 5 の重心は北米と欧州において小さく，Adult のサイトはこれら二つのクラスタのいずれかに分類された．

図 6(c) と図 6(d) に noon データに対する同様の結果をプロットするが，midnight データと同様の結果が確認される．midnight データでは得られなかった，オセアニアにおいて重心が小さい Cluster 4 が新たに出現しているが，ごく少数の Web サイトのみが本クラスタに分類された．

### 4.4　オブジェクトの平均取得遅延

最後に本節では，各 URL カテゴリの Web サイトにおける，各オブジェクトの取得に要した平均遅延時間についての傾向を分析し，CDN オブジェクトの結果と non-CDN オブジェクトとの結果を比較することで，CDN を用いることで得られる遅延低減効果を評価する．オブジェクト取得遅延時間をオブジェクトの取得に要した時間，すなわち測定ホストから HTTP request パケットの送出が完了してから，それに対する HTTP response パケットが最初に測定ホストに到着するまでに要した時間で定義する．オブジェクト取得遅延時間はネットワーク上のパケット転送遅延時間と，オブジェクト生成等の配信サーバにおける処理時間とを含む．

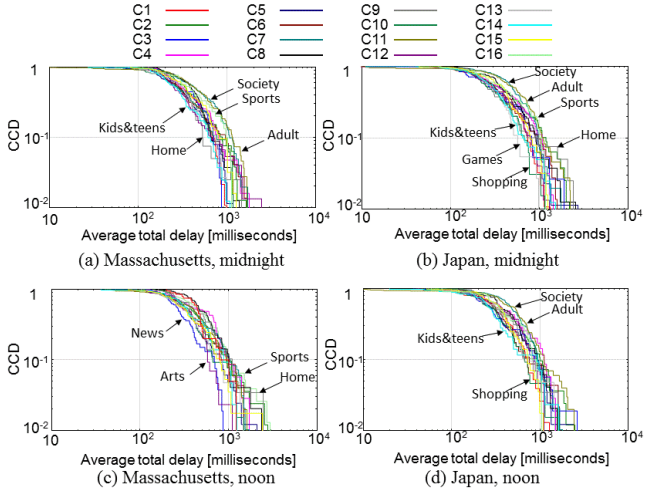


図 7　Complementary cumulative distribution (CCD) of average total delay of non-CDN objects at two PlanetLab hosts

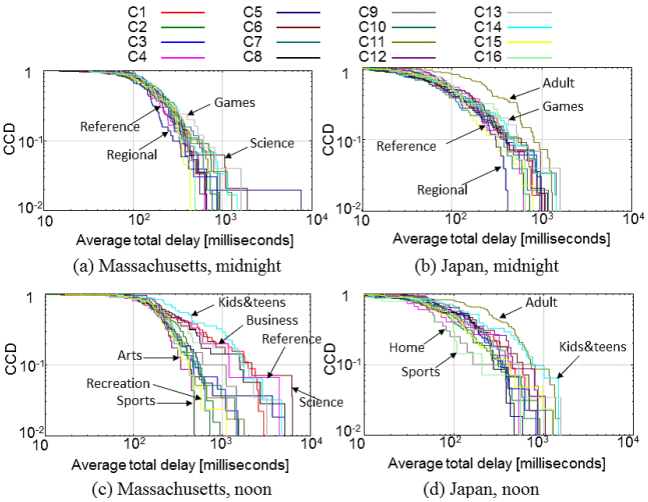


図 8　CCD of average total delay of CDN objects at two PlanetLab hosts

図 7(a) と図 7(b) に，Massachusetts と Japan の PlanetLab ホストで各々測定された midnight データを用いて算出された，各 URL カテゴリの各 Web サイトにおける non-CDN オブジェクトの平均取得遅延の累積補分布 (CCD) をプロットする．また図 7(c) と図 7(d) には noon データを用いて算出された結果を同様に示す．測定時刻が non-CDN オブジェクトの平均取得遅延時間に与える影響は小さい．Shopping や Kids & teens の Web サイトの non-CDN オブジェクトの平均取得遅延時間は小さい傾向があるが，Adult，Sports，Society のサイトは大き

い傾向が見られる．そのため CDN の利用比率を増加させることで，Adult，Sports，Society のサイトの応答時間が効果的に低減する可能性がある．

図 8 に CDN オブジェクトを対象とした平均取得遅延時間の CCD を同様に示す．non-CDN オブジェクトの結果と比較して，測定時刻や測定場所が結果に与える影響が大きく，特に測定時刻に強く依存する．また URL カテゴリ間のオブジェクト平均取得遅延時間の差異は日中の方が深夜と比べて大きい．図 7 に示したように，20%から 80%程度の Web サイトの non-CDN オブジェクトの平均取得遅延時間は 500 ミリ秒を超えている．しかし CDN を用いることで，平均取得遅延時間が 500 ミリ秒を超える Web サイトの比率を，夜間は約 5%から 20%程度に，日中は約 2%から 40%程度に削減できることが確認される．

CDN オブジェクトの平均取得遅延時間は，Games といった娯楽系の Web サイトは夜間に大きくなる傾向があり，Business や Reference といった仕事系のサイトは日中に大きくなる傾向がある．kids は主に日中，活動的な傾向があるため，Kids & teens のサイトの CDN オブジェクトの平均取得遅延時間も日中に大きい，各 URL カテゴリの Web サイトの日中の需要量は夜間の需要量とは異なるため，キャッシュオブジェクトの置換といったキャッシュ制御において URL カテゴリを元に優先度を変えることは有効と思われる．例えば夜間は娯楽系のサイトに，日中は仕事系のサイトに各々，高い優先度を与えることが望ましい．

## 5. ま と め

本稿では，PlanetLab を用いて世界の 12 の拠点から，アクセス頻度の多い約 1,000 の Web サイトにアクセスしたときに発生する通信パタンを測定し，サーバ距離，遅延時間，オブジェクト数といった各種特性値の地域的な傾向について分析し，オリジナルオブジェクトの地域的な配置傾向や，CDN の利用状況に関する傾向を明らかにした．得られた主な知見を以下にまとめる．

- 各 Web サイトを構成するオブジェクトの中で CDN を用いて配信されている CDN オブジェクトの占める割合は，アクセス地点とは独立であるが，URL カテゴリによって大きく異なり，0.2 から 0.5 程度の割合である．
- Society，Health，Adult，Recreation，Sports といったサイトのオリジナルオブジェクトの多くは北米で提供されており，これらカテゴリのコンテンツは地域性が低く，北米で提供される同一のコンテンツが世界中のユーザから閲覧される傾向がある．対照的に Home や Shopping といったカテゴリの Web サイトは地域性が高く，これらサイトのオリジナルオブジェクトは各地域に固有のものが各々の地域で提供される傾向がある．Computers，News，Reference，Science，Arts，Games，Kids & teens といったカテゴリのオリジナルオブジェクトの地域的な配置傾向は，これら二つの極端な場合の中間的な傾向を示しており，オリジナルオブジェクトは北米，欧州，アジアで提供される傾向がある．
- CDN のキャッシュサーバの地理的配置パタンは同一ではなく，様々な配置パタンが存在する．一番目の配置パタンはキャッシュサーバを北米，欧州，アジア，オセアニアといった様々な地域に配置しているのに対して，二番目の配置パタンでは北米と欧州に集中して配置されており，また三番目の配置パタンではオセアニアにのみ配置されている．Adult を除く全てのカテゴリにおいて，80%以上の Web サイトは一番目の配置パタンの CDN を用いており，様々な地域に配置されたキャッシュサーバからオブジェクトが取得される．Adult の Web サイトは二番目の配置パタンの CDN を利用する傾向が高い．
- 各 URL カテゴリのサイトの日中の需要量は夜間の需要量とは異なり，オブジェクト取得遅延時間は時間帯によって異なる．そのためキャッシュオブジェクトの置換といったキャッシュ制御において，URL カテゴリ別に優先制御を行うことが有効と思われる．例えば夜間は娯楽系のサイトに対して，日中は仕事系のサイトに対して各々，キャッシュ制御において高い優先度を与えることが望ましい．

今回行った Web トラヒック測定実験では，人気の高い Web サイトのホームページのみを分析対象としたが，ホームページから辿れるページに対してもオリジナルオブジェクトや CDN キャッシュの地理的配置について分析する予定である．また一日や一週間といったスパンにわたり，Web トラヒックの通信構造がどのように変化するかについても分析する予定である．

## 文 献


[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, Englewood Cli?s, NJ Prentice-Hall, 1988.
[14] 上山, 中野, 塩本, 長谷川, 村田, 宮原, Web トラヒックの地域的な傾向分析, 信学技報 NS2014-20.
[15] M. Karlsson and M. Mahalingam, Do We Need Replica Placement Algorithms in Content Delivery Networks?, WCW 2002.
[16] C. Labovitz, S. Iekel-Johnson, J. Oberheide, and F. Jahanian, Internet Inter-Domain Traffic, ACM SIGCOMM 2010.
[17] MaxMind, GeoIP Downloadable Databases, http://dev.maxmind.com/geoip/downloadable.
[18] J. Mickens, Silo: Exploiting JavaScript and DOM Storage for Faster Page Loads, USENIX WebApps 2010.
[19] E. Nygren, R. Sitaraman, and J. Sun, The Akamai Network: A Platform for High-Performance Internet Applications, ACM SIGOPS 2010.
[20] J. Odvarko, HAR Viewer, Software is hard, http://www.softwareishard.com/blog/har-viewer.
[21] J. Ott, M. Sanchez, J. Rula, F. Bustamante, Content Delivery and the Natural Evolution of DNS, ACM IMC 2012.
[22] PlanetLab, https://www.planet-lab.org/
[23] G. Podjarny, Not as SPDY as You Thought, http://www.guypo.com/technical/not-as-spdy-as-you-thought/
[24] Quantcast, http://www.quantcast.com/top-sites-1.
[25] 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.
[26] F. Schneider, S. Agarwal, T. Alpcan, and A. Feldmann, The new web: characterizing AJAX traffic, ACM PAM 2008.
[27] A. Sharma, A. Venkataramani, and R. Sitaraman, Distributing Content Simplifies ISP Traffic Engineering, ACM SIGMETRICS 2013.
[28] S. Sivasubramanian, et al., Analysis of Caching and Replication Strategies for Web Applications, IEEE Internet Computing, 11(1), pp.60-66, 2007.
[29] S. Souders, High Performance Web Sites: Essential Knowledge for Front-End Engineers, O' Reilly Media, 2007.
[30] SPDY: An experimental protocol for a faster web, http://www.chromium.org/spdy/spdy-whitepaper.
[31] Squid cache replacement policy, http://www.squid-cache.org/Doc/config/cache_replacement_policy/.
[32] 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.
[33] S. Sundaresan, N. Feamster, R. Teixeira, and N. Magharei, Characterizing and Mitigating Web Performance Bottlenecks in Broadband Access Networks, ACM IMC 2013.
[34] X. Wang, A. Balasubramanianm A. Krishnamurthy, and D. Wetherall, Demystifying Page Load Performance with WProf, NSDI 2013.
[35] X. Wang, A. Balasubramanianm A. Krishnamurthy, and D. Wetherall, How speedy is SPDY?, NSDI 2014.