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.

エッジ配信によるWeb応答時間削減効果の分析

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 ページは Ajax 等による動的生成オブジェクトを様々な配信サーバから取得する傾向が高く，Web ページ閲覧時に発生する処理や通信パタンが複雑化しており，Web ページが表示されるまでの待ち時間 (Web 応答時間) が増大している．Web 応答時間を低減するにはユーザ近くから動的オブジェクトを含めて配信することが望ましく，エッジノードでオブジェクトを生成・配信するエッジ・コンピューティング等のエッジ配信が有効である．エッジ配信の Web 応答時間削減効果はオブジェクトの地理的配置状況等に依存するため，ISP や CDN 事業者のエッジ配信導入に際しては，Web 応答時間削減効果を簡易な方法で推定できることが望ましい．そこで本稿では Web ページの閲覧時に発生するオブジェクト取得フローをモデル化し，エッジ配信で得られる Web 応答時間削減量の簡易な推定式を導出する．そして PlanetLab を用いて世界の 12 の地点から約 1,000 の高人気 Web ページを閲覧することで得られた測定データを用いて，ニュースやスポーツなどの Web ページのカテゴリごとに，エッジ配信の Web 応答時間削減効果を分析する．
キーワード　Web，応答時間，エッジ配信

Analyzing Effect of Edge Computing on Reduction of Web Response Time

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, so we face the increase of web response time. To reduce the web response time, edge computing in which dynamic objects are generated and delivered from edge node is effective. For ISPs and CDN providers, it is desirable to estimate the reduction effect of web response time when introducing the edge computing, so in this paper, we derive a simple formula estimating the reduction effect of web response time by modeling the flows obtaining objects of web pages. We investigate the reduction effect of web response time by edge computing in each website category, e.g., news and sports, using data measured by browsing about 1,000 popular web pages from 12 locations in the world on the PlanetLab.
Key words　Web, response time, edge computing

1. はじめに

近年，Web ページ閲覧はインターネットの主要サービスの一つとなっている．しかし 6 割以上のユーザは頻繁に Web ページ閲覧時の待ち時間の長さを意識している [4]．本稿では，ユーザがブラウザ上で目的 Web ページのリンクをクリックした瞬間から，目的 Web ページを構成する全てのオブジェクトが取得され目的 Web ページの表示が完了するまでの時間を Web 応答時間と定義する．17%のユーザは待ち時間が 5 秒を超えると閲覧を諦めるといった報告や [7]，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%低減することが報告されている [23]．また Web 応答時間が 0.1 秒低減するごとに Amazon の利益が 1%増加することや [11]，高速に表示される Web ページはユーザが購買を完了する回数が 15%も多く，1 ページだけ閲覧した後にページから離脱する回数が 9%少ないことが報告されている [15]．そのため Web 応答時間を低減することが，Web 閲覧サービスのユーザ体感品質向上やコンテンツプロバイダの収益増加に重要である．

Web 応答時間の低減技術として CDN (contents delivery networks) が一般的であり [16] [22]，アクセス数上位 1,000 のページの中では 74%が CDN を利用している [16]．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 [12]，CDN の提供形態が多様化している．従来の Web ページ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．これら静的オブジェクトを配信する上では，遠方に存在するオリジンサーバの代わりにユーザ端末近くに存在するキャッシュサーバから配信する CDN は，Web 応答時間を低減する上で効果的である．しかし近年，ユーザ端末からのリクエスト受信時に，Servlet や JSP (Java server pages) のプログラムを配信サーバ側で実行するか，JavaScript で書かれた Ajax によるプログラムをユーザ端末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 [4] [19]．動的オブジェクトはユーザの属性情報などに基づきページ閲覧時に動的に生成されることから，CDN を用いた配信が困難である．

動的オブジェクトを低遅延でユーザ端末に配信するためには，動的オブジェクトの生成など Web サービス提供のために必要となる各種処理を，ネットワーク上の遠方やコアに存在する配信サーバやクラウドで実行する代わりに，ユーザ端末の近くに存在するエッジノードで実行することが有効であり，AT&T

が ACDN という名称で [18]，また Akamai が EdgeComputing という名称で提唱している [5]．そして Sivasubramanian らは EdgeComputing の様々なアーキテクチャをデータの一貫性保持や遅延の観点から比較している [20]．さらに近年，クラウドと比較してよりユーザの近くに存在するエッジノードで演算を行うという意味で，Fog Computing という名称で同様のアーキテクチャが提案されている [3] [27] [29]．またスマホなど，処理能力や電力消費量に制約のあるデバイス上での処理をエッジノードで代替するという観点で，同様のアーキテクチャが提唱されている [24]．本稿ではこれらエッジノードで動的オブジェクトの生成などの Web サービスための各種処理を行うアーキテクチャを総称してエッジ配信と呼ぶ．エッジ配信を用いることで，静的オブジェクトに加えて動的オブジェクトもユーザ端末の近くから配信することが可能となり，Web 応答時間の低減が期待される．しかしエッジ配信を行うためにはネットワーク上に多数存在するエッジノードに高度な演算機能を用意し，複数エッジノード間でデータの一貫性を保つ処理を行うなど，導入には高いコストが伴う．そのため ISP や CDN 事業者がエッジ配信導入の是非を判断するには，Web 応答時間の削減効果を簡易な方法で推定できることが望ましい．

そこで本稿では，Web ページの閲覧時に発生するオブジェクト取得フローをモデル化し，エッジ配信の Web 応答時間削減量の簡易な推定式を導出する．そして PlanetLab [17] を用いて世界の 12 の地点から約 1,000 の高人気 Web ページを閲覧し得られた測定データを用いて，ニュースやスポーツなどの Web ページのカテゴリごとに，エッジ配信の Web 応答時間削減効果を分析する．以後，2 節で Web 応答時間分析に関する既存研究をまとめた後，3 節にてエッジ配信の Web 応答時間削減量の推定式を導出する．そして 4 節にて PlanetLab を用いた Web 閲覧実験方法について述べ，5 節にてエッジ配信の Web 応答時間削減量の数値結果を考察し，最後に 6 節にて全体をまとめる．

2. 関 連 研 究

本節では，Web ページ閲覧時の応答時間の分析に関連する既存研究をまとめる．まず分析対象の Web ページを実際に閲覧して発生したトラヒックを測定するアクティブ計測により Web トラヒックの各種指標を分析する取組として，例えば Butkiewicz らはランダムに選択した約 1,700 の Web ページを世界の 4 つの測定地点から 9 週間にわたり周期的にアクセスし，得られた HAR (HTTP archive record) ファイルを分析することで Web ページのカテゴリごとに構成オブジェクト数やアクセスサーバ数等の各種指標を収集し，これら指標から Web 応答時間を推定する回帰式を導出している [4]．また Zaki らはアフリカのガーナから Wifi や 3G による無線アクセスによる Web 品質を HAR ファイルより分析し，DNS 解決時間，HTTP リダイレクション，TLS/SSL ハンドシェイクなどの各処理が Web 応答時間に占める割合を分析している [28]．これらの研究では多数の Web ページを閲覧する実験を行うことで，各カテゴリの応答時間に与える影響の大きな指標を明らかにしているが，あくまでも傾向の分析に留まっており，オブジェクトをエッジから配信することで得られる応答時間の削減効果を推定することはできない．Sundaresan らは，Web ページの静的オブジェクトのみを取得するプログラム (Mirage) を世界中の多数のユーザ端末に実装し，静的オブジェクトをホームネットワーク内でキャッシュした場合としない場合での 9 つの Web ページにアクセスした際の静的オブジェクトの取得時間を比較することで，静的オブジェクトをホームネットワーク内でキャッシュする効果を分析している [23]．しかし分析対象が少数の Web ページに限定されており，カテゴリ間の傾向の差異については分析されていない．

一方，端末等でキャプチャしたパケットトレースデータを分析することで，各種方策の Web 応答時間低減効果を分析するパッシブ測定による取組も見られる．例えば Li らは，Web サービス利用時のパケットデータからターゲット Web サービスを構成するオブジェクトの依存関係を DAG (directed acyclic graph) で抽出し，RTT やサーバ応答時間や DNS 解決時間といった特定のオブジェクト取得時の品質を変化させた場合の PLT (page load time) を推定する WebProphet を提案している [13]．また Dhawan らは，ブラウザにインストールして Web 通信の品質をパッシブおよびアクティブ測定するツール Fathom を提案し実装したものをオープンソースとして公開している [6]．また Tariq らは，CDN を用いた Web サービスの応答時間改善のために，サーバ選択法，サーバ配置，容量増設，等の方策を実施したときの Web 応答時間低減効果を，What-if シナリオとしてオペレータが入力することで定量的に予測するシステム WISE を提案している [25]．さらに Wang らは，Web 閲覧時にブラウザで実行される各種機能の関係をモデル化し，各種機能の依存関係を測定するためのブラウザプラグインツール WProf を提案し，様々な Web ページ閲覧時の表示待ち時間に影響を与えるボトルネックを分析している [26]．しかしこれらの方式は Web 応答時間推定に要する処理負荷が高く，多数の測定地点から多数の Web ページに対する網羅的な分析を行うことは困難である．パッシブ測定パケットデータを用いた網羅的な Web 品質の分析に関する研究も見られ，例えば He らは IaaS クラウドを利用しているページ URL のリストとパケットキャプチャデータから人気上位ページのホスティングサービスの利用実態を調べ，ダイナミックに複数のエリアのキャッシュを利用することでスループットや遅延性能が改善することを示している [9]．しかし各オブジェクトの配信遅延時間やスループットの分析に留まっており，Web 応答時間の改善効果については評価がなされていない．

3. エッジ配信の Web 応答時間削減効果の推定法

Web ページをブラウザに描写する際には，HTML の解析，各種オブジェクトの取得，JavaScript によるユーザ端末での動的オブジェクト生成，DOM (document object model) tree の生成，DOM tree から Web ページの描写処理 (rendering) などの，様々な処理が発生する [26]．エッジ配信により，これらの処理のうち，オブジェクトを遠方に存在する配信サーバから取得する際に生じる遅延時間を削減することが可能となる．本節では，ユーザ端末と配信サーバとの間で発生するオブジェクト取得フローをモデル化し，続いて，エッジ配信による Web 応答時間削減量の推定式を導出する．

3.1 オブジェクト取得フローのモデル化

近年の Web ページの多くは，40 個から 100 個程度の多数のデータオブジェクトから構成されている [10]．ユーザが Web ブラウザ上で目的 Web ページへのリンクをクリックすると，ブラウザは最初に HTML ファイルをダウンロードして解析し，図 1(a) に示すように，各構成オブジェクトを TCP コネクション上に張られた HTTP セッションにより配信サーバから取得することで(注1)，目的 Web ページを描写する [26]．図 1(a) の例では，サーバ A から 3 個の，サーバ B から 1 個のオブジェクトを各々取得している．各オブジェクトの取得は，ユーザ端末が HTTP request を配信サーバに送付し，配信サーバが HTTP response としてオブジェクトをユーザ端末に送付することで行われる．Java Servlet や JSP (Java server pages) により動的オブジェクトを生成する際の処理時間など，HTTP request 受信から HTTP response 送信の間には，配信サーバにおいてサーバ応答時間 (server response time) が発生する．そのため各オブジェクトの取得には，1 RTT (round-trip time) にサーバ応答時間を加えた遅延時間が生じる．

Web ページを構成する多数のオブジェクトを単一の TCP コネクションを用いて取得すると Web 応答時間が大きくなることから，多くのブラウザでは通常，複数の TCP コネクションを各配信サーバとの間に確立し，複数のオブジェクトを並行して同時に取得する．しかし多数の TCP コネクションを同時に同一の配信サーバとの間に確立すると，配信サーバが高負荷となり性能が劣化することから，多くのブラウザでは同一のサーバとの間に確立可能な TCP コネクション数に上限 P を設けている [21]．例えば HTTP/1.1 では最大数として $P = 2$ が奨励されており，Sarafi 3 や Opera 9 では $P = 4$，Internet Explore 8 や Firefox 3 では $P = 6$ に，各々上限が設定されている [21]．図 1(b) に，図 1(a) のオブジェクト取得例において $P = 2$ の場合の配信フローを例示する(注2)．サーバ A からは 3 個のオブジェクトを取得するため $P = 2$ 本の TCP コネクションを確立し，オブジェクト a と b が並列にダウンロードされる．そしてオブジェクト a の取得が完了した後で，残るオブジェクト c がオブジェクト a を取得する際に用いたのと同一の TCP コネクションを用いてダウンロードされる．一方，サーバ B からは 1

(注1)：コンテンツの提供者が CDN を利用している場合，CDN のキャッシュサーバから配信される．また同一の Web ページを短期間内に閲覧した場合には，ユーザ端末内でキャッシュされており配信フローが発生しない．

(注2)：各 TCP コネクションの確立の際に，ユーザ端末と配信サーバとの間で three-way handshaking が各々必要となるが，図では省略している．

個のオブジェクト d のみが取得されるため，1 本の TCP コネクションのみが確立される．

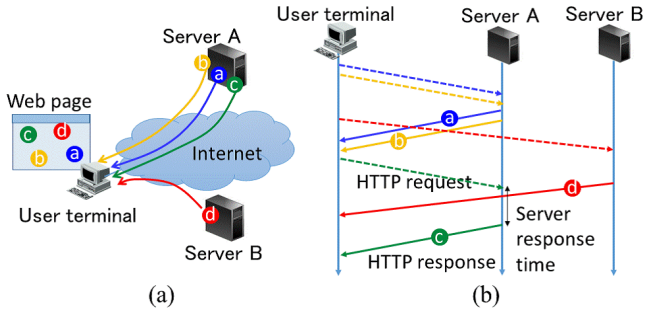

図 1　Example of sequence of objects fetching

3.2　エッジ配信による Web 応答時間削減量の導出

本節では，エッジ配信の Web 応答時間削減量を与える簡易な推定式を導出する．前節で述べたようにオブジェクトの取得に要する時間は，各配信サーバから取得されるオブジェクト数，各配信サーバの応答時間，各配信サーバとの RTT など，様々な要因が影響する．さらに DOM tree の一貫性を保つなどの理由により，あるオブジェクトの取得や生成処理が完了しないと，他のオブジェクトの取得が開始できないなど，多くのオブジェクト間には依存関係が存在する [26]．エッジ配信による Web 応答時間の削減量を厳密に算出するためには，これらオブジェクト間の依存関係を考慮する必要があるが，本稿ではエッジ配信による Web 応答時間削減量の下限値の簡易な推定式を導出することを目的として，オブジェクト間の依存関係は考慮せず各 TCP コネクション上にはユーザ端末上での待ち時間が発生することなく連続してオブジェクトが取得されることを想定する．そのため配信サーバ s の平均応答時間を v_s，サーバ s から取得する各オブジェクトの平均転送時間を u_s，ユーザ端末と配信サーバ s との間の平均 RTT を r_s，TCP コネクション q 上でサーバ s から取得するオブジェクト数を $\xi_{s,q}$ とすると，TCP コネクション q 上でサーバ s からのオブジェクト取得に要する総時間 $y_{s,q}$ は，$y_{s,q} = r_s + \xi_{s,q}(r_s + v_s + u_s)$ で与えられる．ただし TCP コネクション確立のための three-way handshaking に要するサーバ s の処理時間はゼロと考え，その処理時間としては 1 RTT r_s のみを考慮する．

配信サーバ s に対して確立された最大 P 本の TCP コネクションは，全て同時に確立され同時にオブジェクトの取得が開始されることを想定する．このとき，配信サーバ s からのオブジェクト取得に要する総時間は，最大 P 本の TCP コネクションの中で取得オブジェクト数 $\xi_{s,q}$ が最大の TCP コネクション q における総遅延時間 $y_{s,q}$ となる．そのため配信サーバ s から m_s 個のオブジェクトを取得するとき，配信サーバ s からのオブジェクト取得に要する総時間 Y_s が最小となるのは，最大 P 本の並列 TCP コネクション間で取得オブジェクト数の偏りが最小となる場合であり，このとき各 TCP コネクション上で取得されるオブジェクト数の最大値は $\lceil m_s/P \rceil$ となる．ただし $\lceil x \rceil$ は x を下回らない最小の整数である．さらに配信サーバ s との TCP コネクションの確立に先立ち，配信サーバ s の IP アドレスを DNS により解決する必要があり，そのための処理時間(注3)の平均値を d とすると Y_s は次式で得られる．

$$Y_s = d + r_s + \left\lceil \frac{m_s}{P} \right\rceil (r_s + v_s + u_s) \qquad (1)$$

さらに Web ページ x のオブジェクトを提供する配信サーバの集合 $\boldsymbol{S}_x$ の全配信サーバに対して同時に TCP コネクションが確立されオブジェクトの取得が開始されることを想定する．そのためエッジ配信を用いない場合に，Web ページ x の全てのオブジェクトを取得するために要する時間 w_x は，$w_x = \max_{s \in \boldsymbol{S}_x} Y_s$ で与えられる．

次に Web ページ x の構成オブジェクトを全て，エッジノードから配信した場合に要する時間 w'_x を考える．この場合にも，DNS 解決処理時間，サーバ応答時間，オブジェクト転送時間が発生するが，エッジノードから配信しない場合と同じ処理時間，d，v_s，u_s がエッジ配信の場合も発生するものと見なす．ただしエッジ配信の場合には r_s がゼロを想定すると，配信サーバ s から取得していたオブジェクトを全てエッジノードから配信した場合の総時間 Y'_s は次式で得られる．

$$Y'_s = d + \left\lceil \frac{m_s}{P} \right\rceil (v_s + u_s) \qquad (2)$$

Web ページ x の全オブジェクトをエッジノードから取得した場合に要する時間 w'_x は，やはり w_x と同様 $w'_x = \max_{s \in \boldsymbol{S}_x} Y'_s$ で得られるが，世界の多くの地点において Web ページを閲覧した際の v_s は r_s より一桁程度大きいことから [10]，同一の配信サーバ s が Y_s と Y'_s の最大値を与えるとみなせる．よって Web ページ x の全てのオブジェクトをエッジノードから配信することで得られる Web 応答時間の削減量 e_x は次式で得られる．

$$e_x = w_x - w'_x = \max_{s \in \boldsymbol{S}_x} \left[\left\{ \left\lceil \frac{m_s}{P} \right\rceil + 1 \right\} r_s \right] \qquad (3)$$

4.　多地点からの Web 閲覧実験

式 (3) を用いてエッジ配信による Web ページ x の応答時間削減量 e_x を推定するためには，Web ページ x のオブジェクト配信サーバの集合 $\boldsymbol{S}_x$，各配信サーバ s の配信オブジェクト数 m_s と RTT r_s に関するデータが必要となる．同一の Web ページ x を閲覧した場合も，閲覧場所によってオブジェクトの配信サーバの位置が異なる [1] [10]．そのためエッジ配信の効果を分析するためには，世界の様々な地点から閲覧した場合のデータを取得する必要がある．そこで PlanetLab [17] を用いることで世界の 12 の地点から約 1,000 の高人気 Web ページを閲覧することで得られた測定データを用いて，e_x の数値評価を行う．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) 評価 Web ページの選択，(ii)PlanetLab 上での測定環境の構築と測定地点の選択，(iii) 各測定地点から各分析 Web ページを閲覧したときの HAR (HTTP archive record) ファイルの取得，(iv) RTT の測定，の 4 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．

4.1　評価 Web ページの選択

エッジ配信の Web 応答時間削減効果の傾向を分析するためには，閲覧数の多い高人気の Web ページを分析対象とすることが望ましい．そこで Alexa の Web ページ [2] 上で公開されているアクセスランキングを元に，表 1 に示す 16 の各カテゴリから，最もアクセス数の多い上位 300 の Web ページを測定対象として選択した．

表 1　Properties of webpages of each category evaluated

ID	Category	N_c	$\overline{O}_c$	$\overline{S}_c$	$\overline{M}_c$	$\overline{R}_c$
C1	Business	59	60.7	13.2	6.3	84.3
C2	Computers	112	45.8	9.7	5.9	87.4
C3	News	39	67.4	13.2	6.2	98.0
C4	Reference	112	41.2	6.7	8.3	118.2
C5	Regional	80	51.5	9.4	6.6	86.3
C6	Science	95	52.2	10.0	7.0	86.3
C7	Society	79	65.1	11.6	7.6	86.7
C8	Health	86	57.3	10.8	7.0	77.1
C9	Home	85	63.3	13.5	5.7	69.0
C10	Shopping	69	68.6	14.0	6.1	64.0
C11	Adult	112	50.3	6.4	9.8	111.6
C12	Arts	55	55.2	12.2	6.2	84.9
C13	Games	87	55.3	11.4	6.1	87.2
C14	Kids&teens	106	58.1	11.1	6.4	81.8
C15	Recreation	86	61.0	11.0	7.0	72.8
C16	Sports	38	78.6	15.8	7.3	73.6

4.2　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 (iii) と (iv) の処理を 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 (i) で選択した Web ページを閲覧する．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

(注3)：ローカル DNS サーバへの問合せと，ローカル DNS でキャッシュされていない場合にはルート DNS サーバを起点とする再帰的な全ての問合せ処理を含む処理時間の合計．

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 2 にまとめる．ただし L12 の Reunion はアフリカ南東部のマダガスカル島の東沖に存在するフランス領の島である．

表 2　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

4.3　各評価 Web ページ閲覧時の HAR ファイルの取得

各評価 Web ページに対して，各測定用 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 record) ファイルとして取得した [8]．HAR ファイルは，ユーザ端末と配信サーバ間で転送される HTTP データのヘッダ情報から，ユーザ端末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，評価できるページ数が限られたものとなる．そこで多数のページ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 [14] を用いることで，多数のページ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配信サーバから取得した．そして取得された各 HAR ファイルから配信サーバ名に対応する value を抽出することで，各測定用 PlanetLab ノードにて各評価 Web ページの各配信サーバ名を取得し，各配信サーバ s の配信オブジェクト数 m_s を算出する．

Web ページ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) より取得した現地時刻が 0:00 に各測定地点において評価 Web ページの閲覧を開始し，評価 Web ページリストに従い順次閲覧を行った．ただし配信サーバが 30 秒以上応答しない場合には，次の評価 Web ページの閲覧を開始した．その結果，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 1,124 の Web ページを分析対象とした．表 1 に各カテゴリ c の分析対象 Web ページ数 N_c をまとめる．

4.4　RTT の測定

PlanetLab の各測定ホストにおいて，各評価 Web ページを閲覧して HAR ファイルを取得し，各配信サーバ名を HAR ファイルより抽出した直後に，各配信サーバに対して PlanetLab 測定ホストから ping コマンドを送ることで，測定ホストと配信サーバとの間の RTT を測定した．

5.　数 値 評 価

本節では，4. 節で述べた Web ページの多地点閲覧実験により取得された r_s と m_s を用いて，式 (3) から推定される Web 応答時間削減量 e_x の数値評価結果を分析する．e_x は，各配信サーバ s の RTT r_s と配信オブジェクト数 m_s から定まることから，まずこれら指標に関して各カテゴリの傾向を分析する．そしてこれら測定データを用いて e_x を算出し，世界の各地域における各カテゴリのエッジ配信による Web 応答時間削減効果の傾向を明らかにする．

5.1　各 Web カテゴリの RTT の傾向

表 1 に示す各カテゴリ c の評価分析対象 Web ページの集合 $\boldsymbol{X}_c$ に対して，測定地点 a にて閲覧した際のカテゴリ c の各 Web ページの平均 RTT $R_c^{(a)}$ を次式で定義する．

$$R_c^{(a)} = \frac{1}{N_c} \sum_{x \in \boldsymbol{X}_c} \sum_{s \in \boldsymbol{S}_x^{(a)}} m_s^{(a)} r_s^{(a)} / o_x^{(a)} \tag{4}$$

ただし $\boldsymbol{S}_x^{(a)}$，$m_s^{(a)}$，$r_s^{(a)}$ は各々，測定地点 a から Web ページ x を閲覧した際の，$\boldsymbol{S}_x$，m_s，r_s の測定結果であり，$o_x^{(a)}$ は測定地点 a から Web ページ x 閲覧時の取得オブジェクト数である ($o_x^{(a)} = \sum_{s \in \boldsymbol{S}_x^{(a)}} m_s^{(a)}$)．表 1 には 16 の各カテゴリ c に対して，各 Web ページの構成オブジェクト数の平均値 $\overline{O}_c$，平均配信サーバ数の平均値 $\overline{S}_c$，各配信サーバからの取得オブジェクト数の平均値の平均値 $\overline{M}_c$，平均 RTT の平均値 $\overline{R}_c$ をまとめる．

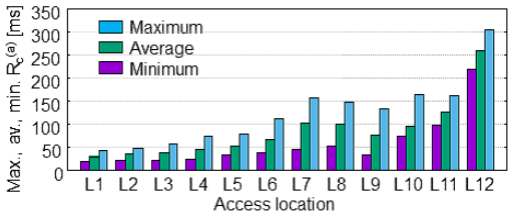

図 2　Maximum, average, and minimum $R_c^{(a)}$, average RTT of each webpage of each category measured at each location

表 3　Categories with top and bottom four ranks in $R_c^{(a)}$

Rank	All	California	Japan	Reunion
R1	Reference	Reference	Adult	Regional
R2	Adult	News	Reference	Business
R3	News	Adult	News	Shopping
R4	Games	Society	Science	Reference
R13	Sports	Recreation	Regional	Kids&teens
R14	Recreation	Arts	Home	Science
R15	Home	Home	Recreation	Home
R16	Shopping	Shopping	Shopping	Health

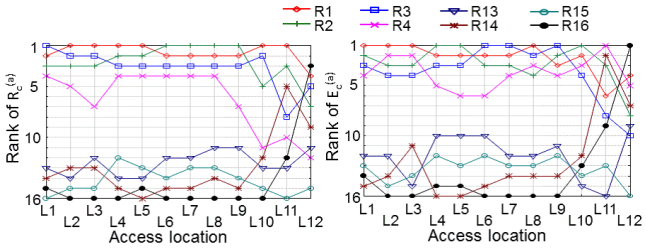

図 3　Ranking of average RTT and average reduction effect of web response time at each location of top and bottom four categories in world

図 2 に，各測定地点における $R_c^{(a)}$ の 16 の全カテゴリにおける最大値，平均値，最小値を各々プロットする．北米 (L1, L2, L3) と欧州 (L4, L5) は $R_c^{(a)}$ の平均値とカテゴリ間の最大格差が共に 25ms〜50ms と小さく，どのカテゴリのオブジェクトも近くから取得される傾向が確認できる．一方，ロシア (L6)，オセアニア (L7, L8)，日本 (L9)，エクアドル (L10) は $R_c^{(a)}$ の平均と最大格差は伴に 75ms〜100ms あり，カテゴリによるオブジェクト配信サーバまでの距離の差異が大きい．さらにアルゼンチン (L11)，アフリカ (L12) は $R_c^{(a)}$ の平均が 125ms〜250ms と大きい反面，カテゴリ間の最大格差は 50ms〜100ms と平均値と比較して小さく，どのカテゴリの配信サーバも遠方に存在する傾向が確認できる．

次に，各測定地点において $R_c^{(a)}$ が大きいカテゴリと小さいカテゴリの傾向の差異を確認するため，表 3 に California, Japan, Reunion の各々における $R_c^{(a)}$ の上位 4 つのカテゴリと下位 4 つのカテゴリをまとめる．また表には，12 の各測定地点における $R_c^{(a)}$ の順位の平均が上位 4 つと下位 4 つのカテゴリについて同様にまとめる (All)．また図 3(a) に，これら全測定地点の平均順位が上位 4 つのカテゴリ (R1〜R4) と下位 4 つのカテゴリ (R13〜R16) の各々の，12 の各測定地点における順

位をプロットする．南米とアフリカ以外の地域は $R_c^{(a)}$ のカテゴリ間の順位に大きな差異がなく世界全体の順位と同等である．

Stack Overflow，Yahoo Answers，Internet Archive など全世界のユーザが共通に使用する情報共有ページが Reference に，Reddit (UGC news ページ)，CNN，Yahoo News など世界規模でニュースを提供しているニュースページが News に各々分類されている [2]．図 2 で見たように北米における $R_c^{(a)}$ が特に小さいことから，これら同一のコンテンツが全世界のユーザに対して提供される傾向の高いカテゴリの Web ページは ICT 環境が良好で多くのコンテンツ事業者が存在する北米にオブジェクト配信サーバが集中する結果，世界の多くの地域で $R_c^{(a)}$ が大きくなる傾向が確認できる．一方，Amazon，Ebay，Netflix など各国別に商店のラインナップが提供されるオンラインショッピングページが Shopping に，Yahoo finance，Yelp (地域情報共有ページ)，Groupon．など各地域に特化した情報が提供される地域情報ページが Home に各々分類されており [2]，これら地域固有の情報を提供する傾向の高いカテゴリの Web ページは同一の URL にアクセスした場合も各地域で異なるコンテンツが提供される傾向があり，世界の多くの地域で $R_c^{(a)}$ が小さくなる傾向が確認できる．

しかし南米とアフリカから閲覧した場合，Home や Shopping など地域固有の情報が提供される傾向のある Web ページであっても，これら地域に特化したコンテンツは提供されない場合が多く，世界の他の地域と比較してカテゴリ間の $R_c^{(a)}$ の順位が大きく異なる．

5.2 配信サーバからの取得オブジェクト数の傾向

測定地点 a にて閲覧した際のカテゴリ c の各 Web ページにおいて，各配信サーバからの平均取得オブジェクト数 $M_c^{(a)}$ を次式で定義する．

$$M_c^{(a)} = \frac{1}{N_c} \sum_{x \in \boldsymbol{X}_c} \sum_{s \in \boldsymbol{S}_x^{(a)}} m_s^{(a)} / s_x^{(a)} \tag{5}$$

ただし $s_x^{(a)}$ は測定地点 a から Web ページ x を閲覧した際にオブジェクトが取得される配信サーバ数である．図 4 に，各測定地点における $M_c^{(a)}$ の 16 の全カテゴリにおける最大値，平均値，最小値を各々プロットする．全世界で各カテゴリの $M_c^{(a)}$ は同様の傾向で，表 1 に示すように配信サーバ数の少ない Adult や Reference といったカテゴリが大きく，配信サーバ数の多い Home や構成オブジェクト数の少ない Computers といったカテゴリが小さい．世界のどの地域においても，$M_c^{(a)}$ のカテゴリ間の最大格差は平均値の半分程度あり，$R_c^{(a)}$ と比較してカテゴリ間の格差が小さい．

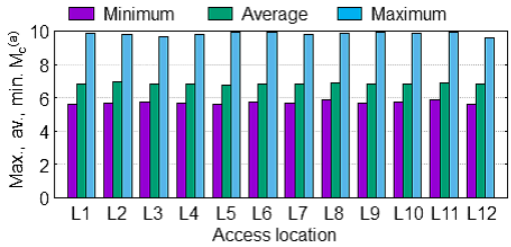

図 4 Maximum, average, and minimum $M_c^{(a)}$, average number of objects obtained from each server

5.3 エッジ配信による Web 応答時間削減効果の傾向

本節では，測定地点 a における r_s と m_s の測定データを用いて，$P = 2$ として式 (3) から推定される Web 応答時間削減量 e_x のカテゴリ c の平均値 $E_c^{(a)}$ を分析し，世界の各地域における各カテゴリのエッジ配信による Web 応答時間削減効果の傾向を明らかにする．測定地点 a から Web ページ x を閲覧した際の，各オブジェクトの平均 RTT を $R_x^{(a)} (= \sum_{s \in \boldsymbol{S}_x^{(a)}} m_s^{(a)} r_s^{(a)} / o_x^{(a)})$，各配信サーバの平均配信オブジェクト数を $M_x^{(a)} (= o_x^{(a)} / s_x^{(a)})$，式 (3) より算出した e_x の値を $e_x^{(a)}$ とする．各測定地点におけるエッジ配信の Web 応答時間削減効果の支配要因を調べるため，図 5 に各測定地点 a における，$e_x^{(a)}$ と $R_x^{(a)}$ との相関係数 (cc: correlation coefficient) と，$e_x^{(a)}$ と $M_x^{(a)}$ との相関係数を各々プロットする．図 2 と 4 で見たように，南米とアフリカ以外の地域は RTT のカテゴリ間の差異が大きいのに対して各サーバの配信オブジェクト数のカテゴリ間の差異が小さい．そのためこれら二つの e_x 決定要因のうち $R_x^{(a)}$ の影響が大きく，$e_x^{(a)}$ と $R_x^{(a)}$ との相関係数が $e_x^{(a)}$ と $M_x^{(a)}$ との相関係数と比較して遥かに大きい．一方，南米とアフリカにおいては，RTT のカテゴリ間の格差が小さいことから，$e_x^{(a)}$ と $R_x^{(a)}$ との相関の強さは $e_x^{(a)}$ と $M_x^{(a)}$ との相関の強さと同等である．

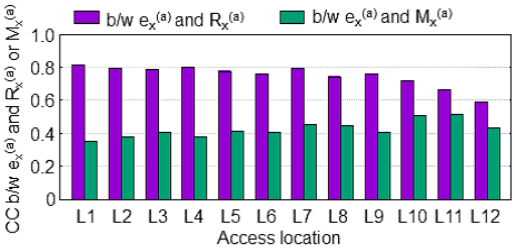

図 5 Correlation coefficient between average reduction of web response time and average RTT or average object count delivered from each server at each measurement location

表 4 Categories with top and bottom four ranks in $E_c^{(a)}$

Rank	All	California	Japan	Reunion
R1	News	News	Adult	Shopping
R2	Reference	Society	Reference	Business
R3	Adult	Reference	News	Sports
R4	Society	Adult	Society	News
R13	Home	Home	Business	Home
R14	Recreation	Computers	Recreation	Science
R15	Computers	Arts	Regional	Games
R16	Shopping	Shopping	Shopping	Computers

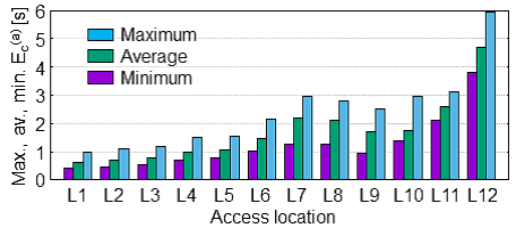

図 6 Maximum, average, and minimum $E_c^{(a)}$, average reduction of web response time

測定地点 a における各カテゴリ c の e_x の平均値を $E_c^{(a)}$ とする．表 4 に California，Japan，Reunion の各々における $E_c^{(a)}$ の上位 4 つのカテゴリと下位 4 つのカテゴリをまとめる．また表には，12 の各測定地点における $E_c^{(a)}$ の順位の平均が上位 4 つと下位 4 つのカテゴリについて同様にまとめる (All)．また図 3(b) に，これら全測定地点の平均順位が上位 4 つのカテゴリ (R1〜R4) と下位 4 つのカテゴリ (R13〜R16) の各々の，12 の各測定地点における順位をプロットする．RTT の場合と同様，南米とアフリカ以外の地域は $E_c^{(a)}$ のカテゴリ間の順位に大きな差異がなく世界全体の順位と同等である．図 5 で見たように，これら地域においては RTT が $E_c^{(a)}$ の決定要因であり $R_c^{(a)}$ のカテゴリ順位と同様の傾向が確認できる．一方，南米とアフリカは $R_c^{(a)}$ と $M_c^{(a)}$ の両方が $E_c^{(a)}$ に影響を与え，他の地域とは異なる固有の傾向が見られる．

図 6 に，各測定地点における $E_c^{(a)}$ の 16 の全カテゴリにおける最大値，平均値，最小値を各々プロットする．エッジ配信による Web 応答時間の削減量のカテゴリ別の平均値は，北米 (L1, L2, L3) は 0.4〜1 秒，欧州 (L4, L5) は 0.6〜1.5 秒程度と小さいが，ロシア・オセアニア・日本・南米 (L6, L7, L8, L9, L10, L11) は 1 秒〜3 秒程度，アフリカ (L12) は 4 秒〜6 秒程度の削減効果が期待できる．図 2 で見たように，北米・欧州以外の地域は多くのカテゴリで配信サーバまでの RTT が大きく，特にアフリカは RTT が大きい．これら RTT の大きな地域で，

エッジ配信は特に効果的である．またアフリカとアルゼンチン以外の地域は $E_c^{(a)}$ のカテゴリ間格差が平均と同程度あり，これら RTT のカテゴリ間格差の大きな多くの地域では，カテゴリごとにエッジ配信の効果が異なることが確認できる．図 7 に，日本とアフリカの各々における，$E_c^{(a)}$ の上位 4 つのカテゴリ (R1〜R4) と下位 4 つのカテゴリ (R13〜R16) の，各 Web ページの応答時間削減量 e_x の累積補分布をプロットする．様々な Web ページにわたり，カテゴリ間でエッジ配信による応答時間削減効果の差異を確認できる．

最後に図 8 に，日本とアフリカにおける各カテゴリの，エッジ配信を用いない場合と用いる場合の各々の平均応答時間をプロットする．ただし各 Web ページの応答時間の実測値をエッジ配信なしの場合の応答時間とし，そこから e_x を減じた値をエッジ配信時の応答時間とした．カテゴリごとに応答時間の削減効果は異なり，日本においては 25%〜54%程度の，アフリカにおいては 43%〜59%程度の応答時間削減効果が期待できる．

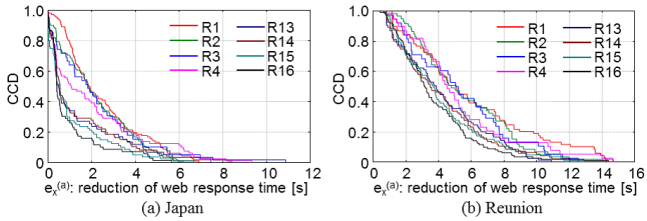

図 7　Complementary cumulative distribution of reduction of web response time at two locations

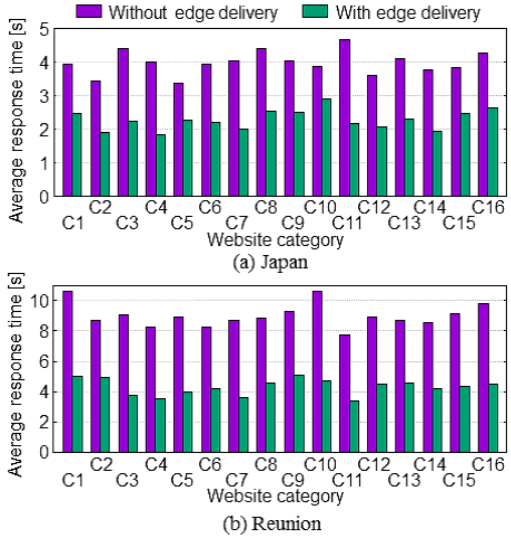

図 8　Average response time without and with edge delivery

6. ま と め

本稿では，Web オブジェクトをエッジノードから配信することで得られる Web 応答時間の削減量を見積もる簡易な推定式を導出し，PlanetLab を用いて世界の 12 の拠点から約 1,000 の Web ページを閲覧することで得られた RTT と配信サーバごとの配信オブジェクト数のデータを用いた数値評価を行うことで，ニュースやスポーツなどの Web ページのカテゴリごとにエッジ配信で得られる Web 応答時間削減効果を分析した．その結果，主に以下の知見を得た．

- Web オブジェクト配信サーバまでの RTT は，ロシア・オセアニア・日本・エクアドルはカテゴリ間で 100ms 弱程度と，全カテゴリの平均 RTT と同程度の差異が見られるが，アフリカとアルゼンチンはカテゴリ間の格差が平均 RTT と比較して小さく，どのカテゴリのオブジェクトも遠方から取得している．カテゴリ間の平均 RTT の順位はアフリカと南米は固有の傾向が見られる．
- Web 応答時間削減効果を決める要因として，アフリカと南米は RTT と配信サーバあたりの平均配信オブジェクト数が影響しており，カテゴリ間の効果の順位も固有であるが，それ以外の地域は RTT のみが支配要因であり，世界中で共通の傾向が見られ，Reference や News など全世界で共通のコンテンツが共有される傾向の高いカテゴリは北米に配信サーバが集中するためエッジ配信の効果が高い反面，Shopping や Home など地域固有のコンテンツが各地域で提供される傾向の高いカテゴリはエッジ配信の効果が低い．
- 北米・欧州以外の地域ではエッジ配信の効果が高く，全カテゴリの平均応答時間削減量は，ロシア・オセアニア・日本・南米が 1.5 秒〜2.5 秒，アフリカが 4.5 秒程度，期待することができる．応答時間全体に占める削減率で考えると，例えば日本においては 25%〜54%程度の，アフリカにおいては 43%〜59%程度の削減効果が期待できる．

文　献

[1] B. Ager, W. Muhlbauer, G. Smaragdakis, and S. Uhlig, Web Content Cartography, ACM IMC 2011.
[2] Alexa, http://www.alexa.com/topsites/category.
[3] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, Fog computing and its role in the internet of things, ACM MCC 2012.
[4] M. Butkiewicz, H. V. Madhyastha, and V. Sekar, Understanding Website Complexity: Measurements, Metrics, and Implications, IMC 2011.
[5] A. Davis, J. Parikh, and W. Weihl, EdgeComputing: Extending Enterprise Applications to the Edge of the Internet, WWW 2004.
[6] M. Dhawan, J. Samuel, R. Teixeira, C. Kreibich, M. Allman, N. Weaver, and V. Paxson, Fathom: A Browser-based Network Measurement Platform, IMC 2012.
[7] When seconds count. http://www.gomez.com/wp-content/downloads/GomezWebSpeedSurvey.pdf.
[8] J. Odvarko, HAR Viewer, Software is hard, http://www.softwareishard.com/blog/har-viewer.
[9] K. He, A. Fisher, L. Wang, A. Gember, A. Akella, and T. Ristenpart, Next Stop, the Cloud: Understanding Modern Web Service Deployment in EC2 and Azure, IMC 2013.
[10] N. Kamiyama, Y. Nakano, K. Shiomoto, G. Hasegawa, M. Murata, and H. Miyahara, Investigating Structure of Modern Web Traffic, IEEE HPSR 2015.
[11] R. Kohavi and R. Longbotham, Online Experiments: Lessons Learned, IEEE Computer, 40(9), pp.103-105, Sep. 2007.
[12] C. Labovitz, S. Iekel-Johnson, J. Oberheide, and F. Jahanian, Internet Inter-Domain Traffic, ACM SIGCOMM 2010.
[13] Z. Li, M. Zhang, Z. Zhu, Y. Chen, A. Greenberg, and Y. Wang, WebProphet: Automating Performance Prediction for Web Services, NSDI 2010.
[14] GitHub, Network Monitoring, http://github.com/ariya/phantomjs/wiki/Network-Monitoring
[15] E. Nygren, R. Sitaraman, and J. Sun, The Akamai Network: A Platform for High-Performance Internet Applications, ACM SIGOPS 2010.
[16] J. Ott, M. Sanchez, J. Rula, F. Bustamante, Content Delivery and the Natural Evolution of DNS, ACM IMC 2012.
[17] PlanetLab, https://www.planet-lab.org/
[18] M. Rabinovich, Z. Xiao, and A. Aggarwal, Computing on the Edge: A Platform for Replicating Internet Applications, WCW 2003.
[19] 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.
[20] 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.
[21] S. Souders, High Performance Web Sites: Essential Knowledge for Front-End Engineers, O' Reilly Media, 2007.
[22] 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.
[23] S. Sundaresan, N. Feamster, R. Teixeira, and N. Magharei, Characterizing and Mitigating Web Performance Bottlenecks in Broadband Access Networks, IMC 2013.
[24] N. Takahashi, H. Tanaka, and R. Kawamura, Analysis of Process Assignment in Multi-tier mobile Cloud Computing and Application to Edge Accelerated Web Browsing, IEEE Mobile-Cloud 2015.
[25] M. Tariq, A. Zeitoun, V. Valancius, N. Feamster, and M. Ammar, Answering "What-If" Deployment and Configuration Questions with WISE, SIGCOMM 2008
[26] X. Wang, A. Balasubramanianm A. Krishnamurthy, and D. Wetherall, Demystifying Page Load Performance with WProf, NSDI 2013.
[27] S. Yi, Z. Hao, Z. Qin, and Q. Li, Fog Computing: Platform and Applications, IEEE HotWeb 2015.
[28] Y. Zaki, J. Chen, T. Potsch, and T. Ahmad, Dissecting Web Latency in Ghana, IMC 2014
[29] J. Zhu, D. Chan, M. Prabhu, P. Natarajan, H. Hu, and F. Bonomi, Improving Web Sites Performance Using Edge Servers in Fog Computing Architecture, IEEE SOSE 2015.