NTT

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

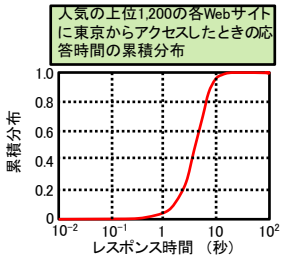
Analyzing Geographical Usage of CDN and Original Content of Website

上山憲昭[1)2)], 中野雄介[1)2)], 塩本 公平[1)]
長谷川剛[3)], 村田正幸[2)], 宮 原秀 夫[2)]

[1)]NTTネットワーク基盤技術研究所
[2)]大阪大学大学院情報科学研究科
[3)]大阪大学サイバーメディアセンター

2015. 5. 21

Web応答時間の増大

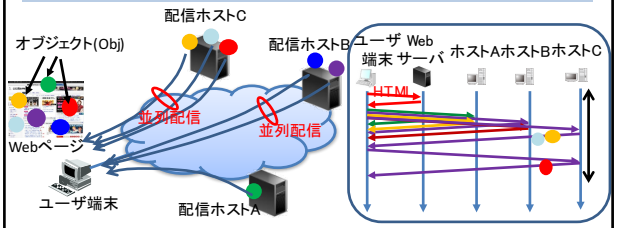


- 50%のサイトは4秒以上，10%のサイトは9秒以上の応答時間が発生
- Web応答時間の改善が重要課題

Web閲覧の応答時間増大の要因

- 【要因1】同一の配信ホストに確立可能なHTTPセッション数に上限があり，直列HTTPセッション数が性能のボトルネック
- 【要因2】特定の配信ホストが高負荷で応答性が低下
- 【要因3】ユーザ端末と特定の配信ホスト間のNWの輻輳
- 【要因4】ユーザ端末のレンダリング処理やHTML解析処理の遅延

HTTP応答時間

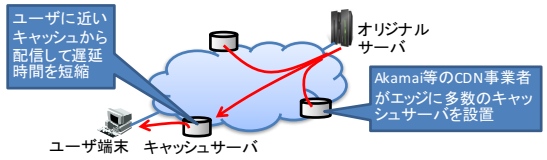


要因2,3による応答時間(HTTP応答時間)を改善するためにCDNが広く普及

CDN: Webオブジェクトの配信プラットフォーム

- 1,000のTopサイトの74%がCDNを利用しており*，CDNはHTTP応答時間改善のための最も一般的な技術

*J. Ott, et al., Content Delivery and the Natural Evolution of DNS, ACM IMC 2012



しかしHTTP応答時間は，配信サーバの位置や性能に大きく依存し，その改善には様々な施策が考えられるが，各施策の効果に関する定量的評価は未検討

HTTP応答時間改善の施策

I. **CP(content provider)がCDN使用オブジェクト(CDN-Obj)の比率を増加**
現状，CDNを用いずに配信しているObj(non-CDN-Obj)をCDNで配信し，CDN-Objの比率を増加させることでHTTP応答時間を低減

II. **CPが使用CDNを変更**
CDN事業者のキャッシュの地理的な配置，キャッシュサーバ能力，NWスループット，キャッシュ制御ポリシィ，等が異なることから，よりHTTP応答時間の低減効果が見込めるCDN事業者にCPが契約先を切替

III. **CDN事業者がObjのキャッシュ位置や配信サーバ選択を適正化**
CDN事業者がキャッシュサーバ配置場所，キャッシュサーバ制御ポリシィ(置換方式，キャッシュ判断方式)，配信キャッシュサーバ選択を適正化し，ユーザに近いキャッシュサーバから配信することでHTTP応答時間を低減

本発表：広域アクティブ測定により，CDN-ObjとNon-CDN-Objの地理的な配置傾向を分析し，施策&IIIのHTTP応答時間改善のポテンシャルを評価

分析のアプローチ

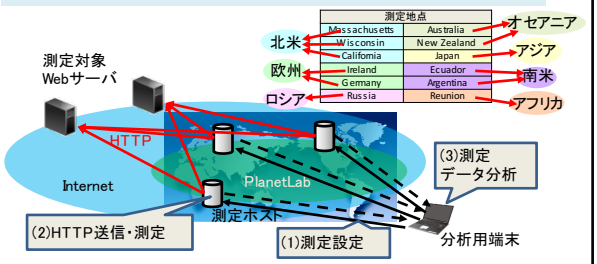
I. 世界の12の地点から約1,000のWebサイトにアクセスした際の，配信サーバ距離，RTT，HTTP応答時間等の各種通信特性を測定

II. 測定データを，CDN-ObjとNon-CDN-Objとに分離して，さらにWebサイトのジャンルごとに，平均値や累積分布を算出

I. 12の各測定地点における各特性値に基づきWebサイトをクラスタ分析することで，各WebサイトジャンルのCDN-ObjとNon-CDN-Objの地理的な配置傾向を分析

広域測定実験の手順

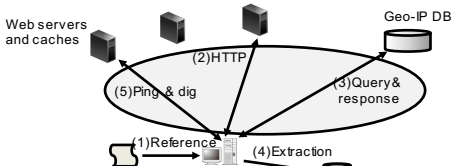
1. 12のPlanetLabホストを測定ホストとして選択
2. 測定プログラムを各測定ホストで指定時刻に実行して多数のWebサイトにバッチ処理でアクセスし，各種通信特性値を測定
3. 収集測定データを分析用端末に集積



PlanetLab: インターネット上に構築された実験用オーバレイNWで，世界中に存在する約500のホスト上で様々なプログラムを実行可能

測定プログラムの動作概要

1. アクセスURLリストを作成
2. 指定時刻(0:00 or 12:00)に各PlanetLabホストはURLリストに従いWebページにアクセスし，発生した通信の各種情報を含むHAR(HTTP Archive)ファイルを取得
3. HARファイル中の各オブジェクトのURL情報から，MaxMindのGeo IP-DBを参照し，各配信サーバの位置座標や都市名を取得
4. 取得HARファイルから各種特性値データを抽出
5. pingを用いて各配信サーバまでのRTTを測定し，digを用いて各配信サーバのドメイン名を取得



アクセスURLリスト

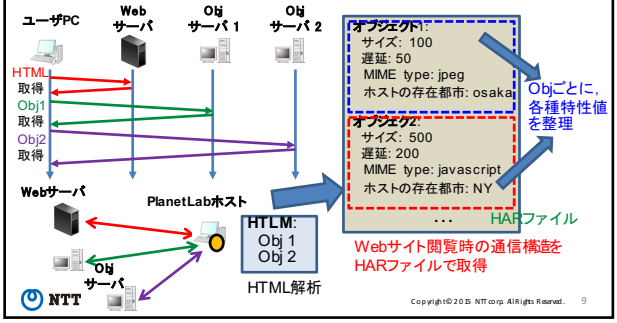
- Alexaのサイト*で公開されているランキング情報をもとに，16の各サイトジャンルから閲覧数上位300のサイトを選択
- 12の全ての測定地点でHARファイルが正しく取得できた927サイトを分析対象に選定

*http://www.alexa.com/topsites

ジャンル	#sites	ジャンル	#sites
Business	40	Home	47
Computer	91	Shopping	68
News	27	Adult	102
Reference	109	Arts	60
Regional	73	Games	58
Science	86	Kids & teens	64
Society	83	Recreation	52
Health	52	Sports	53

HARファイルの取得

- 最初にHTMLが取得され，その中に埋込まれているObjを個別に取得
- HAR(HTTP Archive)ファイル: HTTPデータのヘッダ情報から各Objの各種通信特性値を算出しJSON (JavaScript Object Notation)形式で出力したもの



取得データ

- 各受信オブジェクトに対して，HARファイルから以下の情報を抽出(GeoIPのAPIを用いてホスト名から都市名と座標を取得)

データ項目名	Key
HTTP 送信先ホスト名	"request"-"url"
ホストの存在する国名	GeoIP: "country_name"
ホストの存在する都市名	GeoIP: "city"
ホストの緯度	GeoIP: "latitude"
ホストの経度	GeoIP: "longitude"
サイズ (byte)	"response"-"content" -"size"
総遅延時間 (ms)	"time"
コネクション接続時間 (ms)	"timings"-"blocked"
DNS 名前解決時間 (ms)	"timings"-"dns"
TCP コネクション確立時間 (ms)	"timings"-"connect"
HTTP リクエスト転送時間 (ms)	"timings"-"send"
サーバ応答待ち時間 (ms)	"timings"-"wait"
レスポンス転送時間 (ms)	"timings"-"receive"
SSL/TLS 時間 (ms)	"timings"-"ssl"
MIME Type	"response"-"content" -"mimeType"

- digコマンドを用いて，実際に各オブジェクトを配信したサーバのドメイン名を取得し，さらにpingを送付してRTTを計測

CDNの利用有無によるオブジェクトの分類

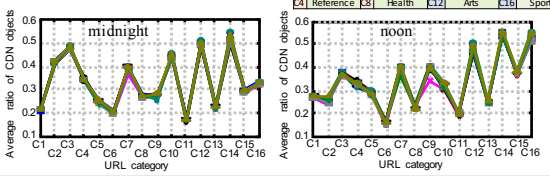
- 以下の手順で各オブジェクトについてCDNの利用の有無を推定
  - "CDN一覧"等の検索結果で得られるCDN事業者の使用ドメイン名をWikiを用いて確認し，CDN使用時の2ndレベルドメイン名リストを作成
  - HARファイルから抽出された各オブジェクトのURL(www.google.com/xxx/yyy.jpgなど)を引数にして，digコマンドを用いて配信サーバのURL(www.akamai.com/qqq/rrrなど)を取得
  - CDNリストと部分一致したURLからの配信をCDN-Objに，それ以外をnon-CDN-Objに分類

CDNの2ndレベルドメインリスト

profile.ak.fbcdn.net	cloudfront.net	akamaihd.net	edgesuite.net
static.ak.fbcdn.net	vo.msecnd.net	edgesuite.net	cloudfront.net
r.ddmcdn.com	edgecastcdn.net	edgekey.net	vo.msecnd.net
s.cdn-care.com	cdngc.net	srip.net	edgecastcdn.net
cmscdn.staticcache.org	bootstrapcdn.com	akamaitechnologies.com	cdngc.net
g-ecx.images-amazon.com	example.com	akamaitechnologies.fr	push-11.cdnsun.com
max.blurtitcdn.com	akadns.net	akamaitech.net	ve14.fr3.atl1.llnw.net
a.espncdn.com	akam.net	akadns.net	hs-9.cdn77.com
ecx.images-amazon.com	akamaiedge.net	akam.net	nyud.net
edgekey.net	akamai.net	akamaistream.net	CloudFlare
edgesuite.net	akamaiedge.net	edgekey.net	Incapsula

CDNオブジェクトの占める割合

ID	ジャンル名	ID	ジャンル名	ID	ジャンル名	ID	ジャンル名
C1	Business	C5	Regional	C9	Home	C13	Games
C2	Computers	C6	Science	C10	Shopping	C14	Kids & teens
C3	News	C7	Society	C11	Adult	C15	Recreation
C4	Reference	C8	Health	C12	Arts	C16	Sports



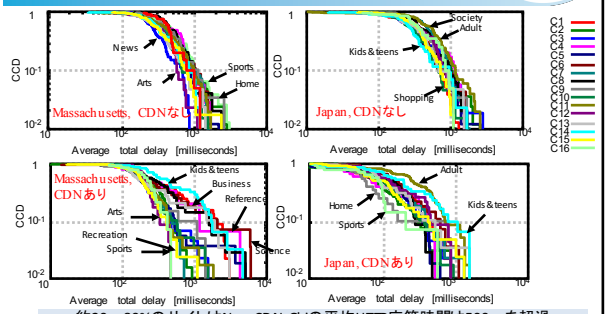
- ジャンルによるCDN利用傾向には大きな差異:
  - Computers, News, Society, Shopping, Arts, Kids & teens等のジャンルのサイトはCDNを用いた配信が多い傾向
  - Business, Regional, Science, Adult, Games等のサイトはCDNを用いた配信が少ない傾向

CDN使用比率増加の余地はジャンル間で差異 (non-CDN-Obj比率は50%〜80%)

NTT

 12

平均WaitのCCDを比較(12:00)



- 約20〜80%のサイトはNon-CDN-Objの平均HTTP応答時間は500msを超過
- CDN-Objの平均HTTP応答時間が500msを超過するサイトは約2%〜40%

CDN使用によるHTTP応答時間削減効果を確認

NTT

 13

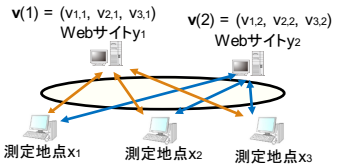
各特性値の地理的傾向に基づくサイトクラスタ分析

- 同一のWebサイトでも，発生通信パタンはアクセス地点に依存
- 12の各測定地点の12の各通信特性値(下左表)に基づきWebサイトをクラスタ分析し，サイトジャンルによる傾向の差異を分析
- 測定地点XからWebサイトYにアクセスしたときの特性値v_{xy}を要素とするベクトル$\mathbf{v}(y)$を元にk-means法で各サイトをクラスタ分析(JD法*を用いて最適クラスタ数kを選択し，初期クラスタをk-means++法**で設定)

平均サーバ距離
サーバ距離の分散
平均サーバRTT
サーバRTTの分散
平均オブジェクトサイズ
オブジェクトサイズの分散
平均オブジェクト取得遅延時間
オブジェクト取得遅延時間の分散
オブジェクト数
アクセスホスト数
アクセス都市数
レスポンス時間

*A. L. Jain and R. C. Dubes, Algorithms for Clustering Data, Englewood Cliffs, NJ Prentice-Hall, 1988

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



 14

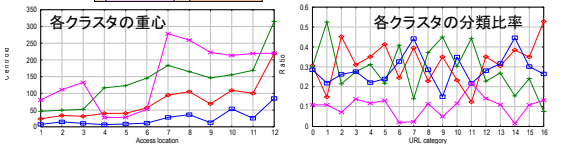
RTTに基づくクラスタリング結果(CDNなし12:00)

Non-CDN-Objの配信パタンにより，オリジナルの配置傾向を確認

Cluster 1
Cluster 2
Cluster 3
Cluster 4

ID	測定地点	ID	測定地点
L1	Massachusetts	L7	Australia
L2	Wisconsin	L8	New Zealand
L3	California	L9	Japan
L4	Ireland	L10	Ecuador
L5	Germany	L11	Argentina
L6	Russia	L12	Reunion

ID	ジャンル名	ID	ジャンル名	ID	ジャンル名	ID	ジャンル名
C1	Business	C5	Regional	C9	Home	C13	Games
C2	Computers	C6	Science	C10	Shopping	C14	Kids & teens
C3	News	C7	Society	C11	Adult	C15	Recreation
C4	Reference	C8	Health	C12	Arts	C16	Sports



- クラスタ1&3はアフリカ以外の全地域でRTTが小さく，各地域に固有のコンテンツが提供⇒Society，Arts，Games，Kids&teens，Recreation，Sports等の，娯楽性の高いサイトや，Regional等の比率が大(地域性が高いジャンル)
- クラスタ2は北米で提供されるコンテンツの比率が大⇒Business，Science，Health，Adult等のサイトのコンテンツは北米に偏重(地域性が低いジャンル)

ジャンル間でCDNを用いること で得られるHTTP応答 時間 削減効 果に差 異

 15

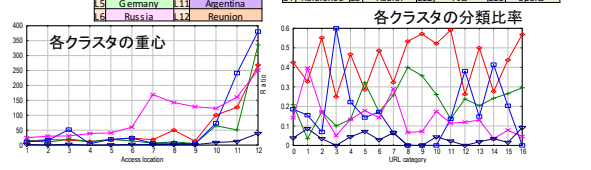
RTTに基づくクラスタリング結果(CDNあり12:00)

Non-CDN-Objの配信パタンにより，CDNの配信元位置の傾向を確認

Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5

ID	測定地点	ID	測定地点
L1	Massachusetts	L7	Australia
L2	Wisconsin	L8	New Zealand
L3	California	L9	Japan
L4	Ireland	L10	Ecuador
L5	Germany	L11	Argentina
L6	Russia	L12	Reunion

ID	ジャンル名	ID	ジャンル名	ID	ジャンル名	ID	ジャンル名
C1	Business	C5	Regional	C9	Home	C13	Games
C2	Computers	C6	Science	C10	Shopping	C14	Kids & teens
C3	News	C7	Society	C11	Adult	C15	Recreation
C4	Reference	C8	Health	C12	Arts	C16	Sports



- クラスタ1&2&3は南米とアフリカ以外の地域から配信⇒News，Reference，Health，Home，Kinds&teens，Recreation，Sports等の分類比率が大
- クラスタ4は北米と欧州から配信⇒Business，Societyの分類比率が大

ジャンル間でCDNの配信元位 置の傾 向が 異なるため，CDNのキャッシュ配置やサーバ選択等の適正化による効 果に差異

NTT

 16

まとめ

- CDNの利用やCDN運用の適正化によるWeb応答時間改善のポテンシャルを，広域アクティブ測定実験により評価
- 各Webサイトを構成するオブジェクトのCDN使用比率は約20%〜50%であり，サイトのジャンル間で差異
- 約20〜80%のサイトはNon-CDN-Objの平均HTTP応答時間は500msを超過しているが，CDN-Objの平均HTTP応答時間が500msを超過するサイトは約2%〜40%
- オリジナルオブジェクトの配置傾向はジャンルによって異なり，CDNを用いることで得られるHTTP応答時間削減効果に差異

  ⇒ CPがCDN使用オブジェクトの比率を増加させることでWeb応答時間が低減する余地あり
- ジャンルによってCDNの配信元位置の傾向が異なり，CDNのキャッシュ配置やサーバ選択等の適正化による効果に差異

  ⇒ CDN事業者がObjのキャッシュ位置や配信サーバ選択を適正化することでWeb応答時間が低減する余地あり

NTT

 17

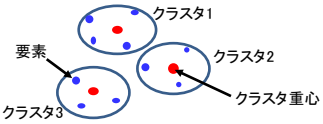
HARファイルの例

- phantomJS(ブラウザの機能を提供)+netsniff(HARファイルを抽出)を用いることで，バッチ処理で多数のサイトのHARファイルを取得

HARファイルの例(www.google.com)

クラスタリング手法

- k-means法: 非階層型クラスタリング手法の一つで，クラスタの重心を用いて，各要素を k 個のクラスタに分類
  - 各要素を重心の距離が最も近いクラスタに分類する処理をクラスタが収束するまで反復

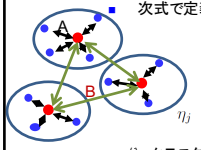


- k-means++法: 距離の離れた要素を初期クラスタの重心に設定することで，分類精度を向上
  - ランダムに一つの要素を選び，クラスタ重心に設定
  - 各要素 x に関して，その最近傍重心との距離D(x)を計算
  - D(x)²に比例する確率に従い，新しいクラスタ重心としてランダムに一つ要素を選択
  - k 個のクラスタ重心が選択されるまで上記処理を反復
  - 以後はk-means法を用いてクラスタを生成

NTT

クラスタ数 k の最適選定

- Jain-Dubes法*を用いて最適なクラスタ数 k を設定
  - 要素数が n のときに，$2 \le k \le 1+\log_2 n$ の範囲で各クラスタ数 k のクラスタリングを実施
  - 次式で定義されるコストp(m)が最小となるkを選択



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

$$\eta_j = \frac{1}{n_j}\sum_{i=1}^{n_j} D\left(\boldsymbol{x}_i^{(j)}, \boldsymbol{m}_j\right) \qquad \xi_{ij} = D(\boldsymbol{m}_i, \boldsymbol{m}_j)$$

$x_i^{(j)}$: クラスタ j 内の i 番目の要素, n_j: クラスタ j の要素数
m_j: クラスタ j の重心, D(**a**,**b**): ベクトル**a**と**b**間の距離

- 各クラスタに属する要素のクラスタ重心に対する距離Aの平均値の，二つのクラスタの重心間の距離Bに対する比率を，最小化することに相当

*A. K. Jain and R. C. Dubes, Algorithms for clustering data, Prentice-Hall, 1988

NTT

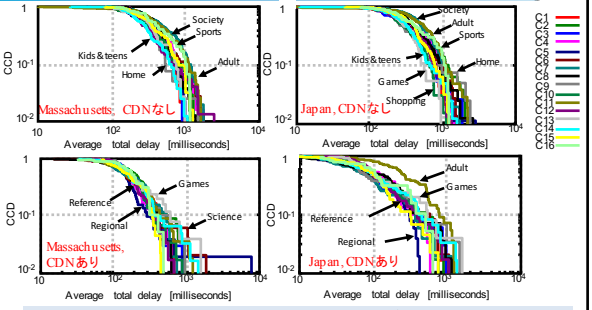
基本特性

ID	Category	Website count 0:00	Website count 12:00	Object size (kbytes)	Object count	Total size (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

- 娯楽系サイト(Arts, Shopping, Sportなど)は，総データ量とオブジェクト数が多い傾向
- 情報収集系サイト(Business, Computers, Health, Referenceなど)は，総データ量とオブジェクト数が少ない傾向

NTT

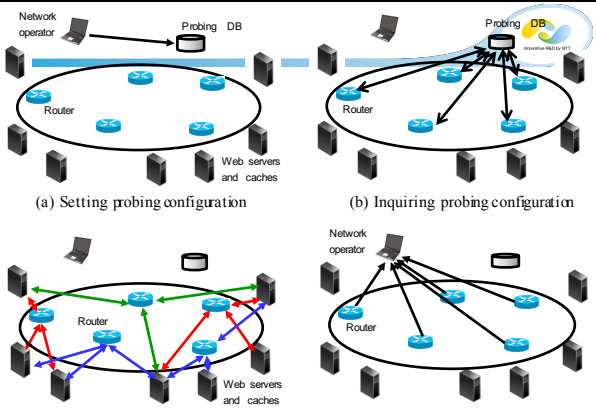
平均WaitのCCDを比較(midnight)



- 約20～80%のサイトはNon-CDN-Objの平均HTTP応答時間は500msを超過
- CDN-Objの平均HTTP応答時間が500msを超過するサイトは約5%～20%

CDN使用によるHTTP応答時間削減 効果を確認

NTT



NTT

 entire_system