Analyzing Effect of Edge Computing on Reduction of Web Response Time

Noriaki Kamiyama[1] Yuusuke Nakano[1]
Kohei Shiomoto[1] Go Hasegawa[2]
Masayuki Murata[2] Hideo Miyahara[2]

[1]NTT Network Technology Laboratories
[2]Osaka University

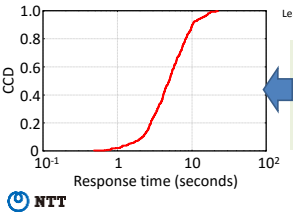
2016. 12. 5

Increase of Web Response Time

- Web response time*: longer than 5 seconds in 50% webpages, and longer than 10 seconds in 10% webpages
- Amazon increased revenue 1% for every 0.1 second reduction in web response time.**
- Need to reducing web response time

*Web response time: waiting time after clicking hyperlink until entire part of webpage is shown

**R. Kohavi and R. Longbotham, Online Experiments: Lessons Learned, IEEE Computer, Vol.40, No. 9, pp.103-105, Sep. 2007.

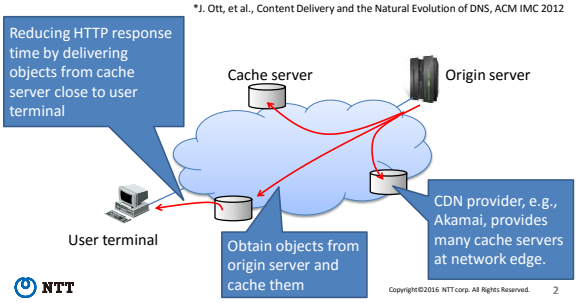


Complementary cumulative distribution (CCD) of web response time of most popular 1,000 websites when accessing from Tokyo, Japan, in June 2015

CDN: Platform Delivering Web Objects

- 74% of 1,000 most popular websites use CDN*, and CDN is most common technique for reducing HTTP response time.

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



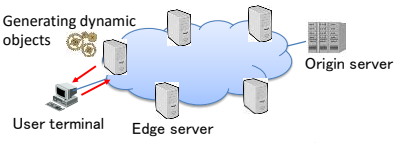
Edge Computing

- Many objects are dynamically generated in modern webpages.
- Edge computing* is effective to deliver dynamic objects efficiently.

*M. Rabinovich, et al., Computing on the Edge: A Platform for Replicating Internet Applications," WCW 2003.
A. Davis, et al., EdgeComputing: Extending Enterprise Applications to the Edge of the Internet, WWW 2004.

Edge servers located at edge nodes

- Cache application codes for generating dynamic objects
- Dynamically generate objects for user requests



Effect of Edge Computing

Geographical deployment pattern may differ among webpage categories, e.g., Sports and News, and effect of edge computing will depend on webpage categories.

Stack Overflow, Yahoo Answers, Internet Archive, ...

Identical content from North America

Yelp, Groupon, ...

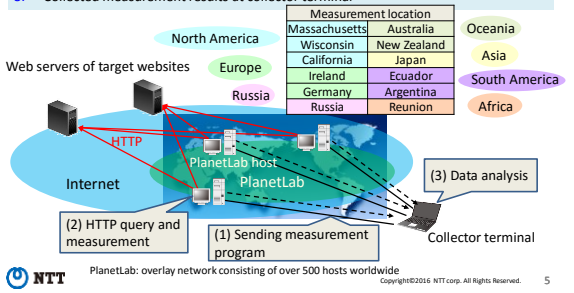
Unique content at each region

Contribution of this work

1. Roughly derives equation estimating reduction effect of web response time by edge computing
2. Evaluates effect of edge computing using active measurement data obtained from 12 locations in world for 1,000 most popular webpages

Measurement Procedure

1. Selected 12 PlanetLab hosts as measurement terminals accessing various webpages
2. Measured various properties, e.g., object count obtained and RTT, by executing program at each PlanetLab host to access various webpages sequentially
3. Collected measurement results at collector terminal



PlanetLab: overlay network consisting of over 500 hosts worldwide

URL List of Measurement Target

- Selected 300 most popular websites in each of 16 categories based on public information of Quantcast*
- Totally Selected 1,124 webpages from which measurement data were successfully obtained at all 12 measurement locations

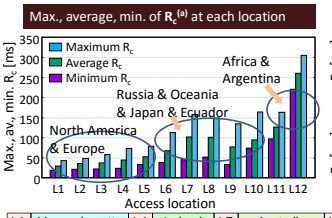
*http://www.quantcast.com/top-pages

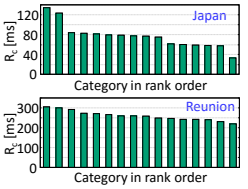
Category	#pages	Category	#pages
Business	59	Home	85
Computer	112	Shopping	69
News	39	Adult	112
Reference	112	Arts	55
Regional	80	Games	87
Science	95	Kids&teens	106
Society	79	Recreation	86
Health	86	Sports	38

Average RTT to Servers in Each Web Category

R_c: average RTT in webpages of category c

Max., average, min. of $R_c^{(a)}$ at each location





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

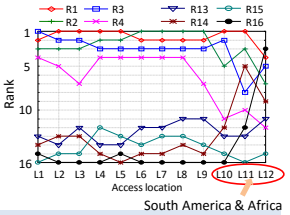
- Africa & Argentina: large RTT in all categories
- Other areas: large difference in RTT among categories, i.e., max. R_c was more than two times larger than min. R_c

Rank of Average RTT among Web Categories

Category Rank in all areas (all) and 3 areas

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
R5	Computers	Business	Computers	News
R6	Science	Science	Society	Arts
R7	Society	Games	Games	Adult
R8	Regional	Kids&teens	Kids&teens	Computers
R9	Arts	Computers	Arts	Recreation
R10	Business	Regional	Health	Society
R11	Kids&teens	Health	Sports	Sports
R12	Health	Sports	Business	Games
R13	Sports	Recreation	Regional	Kids&teens
R14	Recreation	Arts	Home	Science
R15	Home	Home	Recreation	Home
R16	Shopping	Shopping	Shopping	Health

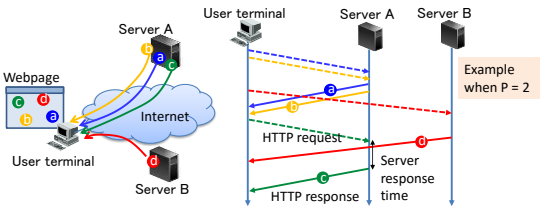
Rank of top and bottom 4 categories at each location



- Category rank of average RTT is common in all areas except South America and Africa
  - Objects of universal webpages, e.g., Reference (Stack overflow, Yahoo Answers, etc) and News (CNN, Yahoo News, etc) concentrate in North America. ⇒ Large RTT
  - Objects of webpages with high locality, e.g., Shopping (Amazon, Ebay, etc) and Home (Yelp, Groupon, etc) are unique in each area. ⇒ Small RTT
- In South America and Africa, objects of all categories exist remote location, and category rank of average rank is unique.

Objects Acquisition at Web Browsing

Web browser displays webpage after downloading many objects from multiple servers at various locations.



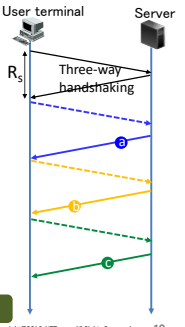
- Number of parallel sessions established with one server is limited below P
  - P = 2 (suggested in HTTP/1.1 specification)
  - P = 4 (Safari 3, Opera 9)
  - P = 6 (Explore 8, Firefox 3)

Roughly Estimating Reduction of Web Response Time

Assumption

- Starts obtaining objects on all TCP co. with all servers
- Fairly obtains objects over all TCP co. with each server
- Continuously receives objects on each TCP connection
- Zero RTT between edge servers and user terminals
- Unchanged DNS resolving, server response, and object sending times

Flow sequence on TCP connection



Estimated time reduced by delivering objects of webpage x from edge servers

$$\max_{s \in \boldsymbol{S}_x} \left[\left\{ \left\lceil \frac{m_s}{P} \right\rceil + 1 \right\} r_s \right]$$

- $\boldsymbol{S}_x$: set of servers sending objects of webpage x
- m_s: number of objects obtained from server s
- r_s: average RTT b/w user terminal and sever s

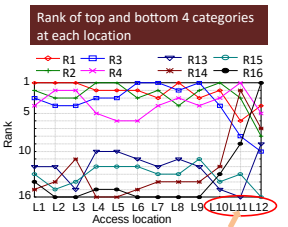
Apply measured value

Rank of Average Reduction of Web Response Time among Web Categories

Category Rank in all areas (all) and 3 areas

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
R5	Sports	Kids&teens	Health	Society
R6	Games	Business	Kids&teens	Regional
R7	Health	Games	Games	Recreation
R8	Business	Health	Science	Reference
R9	Kids&teens	Science	Sports	Arts
R10	Science	Sports	Home	Adult
R11	Regional	Recreation	Arts	Kids&teens
R12	Arts	Regional	Computers	Health
R13	Home	Home	Business	Home
R14	Recreation	Computers	Recreation	Science
R15	Computers	Arts	Regional	Games
R16	Shopping	Shopping	Shopping	Computers

Rank of top and bottom 4 categories at each location

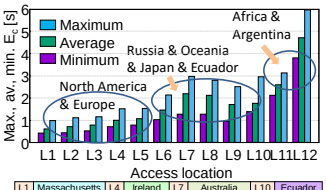


- In South America and Africa: large reduction effect in categories with large average RTT and many objects per server
- In other areas: identical tendency in category rank because E_c is large in categories with large average RTT

Average Reduction of Web Response Time in Each Category

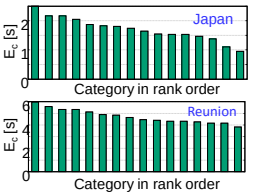
E_c: average reduction of web response time in category c P = 2

Max., av., min. of E_c at each location



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

E_c of categories in descending order



- In areas except North America and Europe:
  - large reduction effect of web response time by edge computing, i.e., 1-3 seconds in Japan and South America and 4-6 seconds in Africa
  - 1-2 seconds difference in E_c among categories

Confirm difference in effect of caching content among web categories

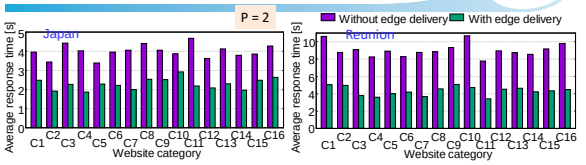
Conclusion

- Derived rough estimate of reduction effect of web response time by edge computing
- Evaluated effect of edge computing using measured data obtained by browsing about 1,000 most popular webpages from 12 locations

- In all areas except Africa and South America, average RTT was largely different among categories: large RTT in global categories, e.g., Reference and News, whereas small RTT in local categories, e.g., Home and Shopping
- Identical category tendencies in effect of edge computing in all areas except Africa and South America, and large effect in global categories
- In Africa and South America, unique tendency of category rank in effect of edge computing

Average Web Response Time in Each Category

P = 2



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

Reduction ratio of web response time:
- Japan: 25% - 54%
- Reunion: 43% - 59%

In Japan and Africa, edge computing can reduce web response time by 0.2 to 0.6 in all categories, and effect depends on categories.