On the Evolution of Locality in Web Traffic

Joachim Charzinski
joachim.charzinski@nsn.com

Nokia Siemens Networks, Munich, Germany

Outline

- Evolution of Web applications and mash-ups
- Experiments and traces
- Results
- Conclusions

Web 0.5, 1.0, 1.5

Web 0.5 (before the Web)

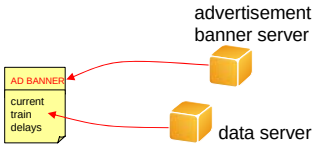
- single servers and isolated documents
- clear client-server relationship

Web 1.0

- hyperlinks
- images from same server

Web 1.5

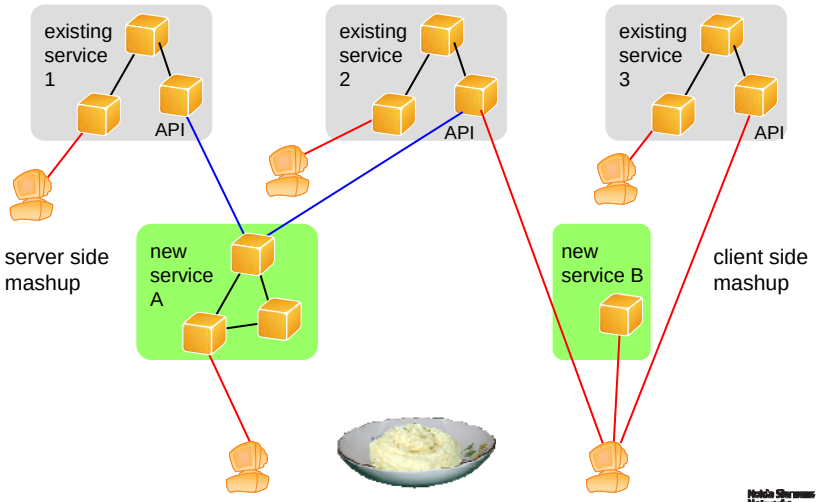
- CDN
- some elements from other servers included on pages
- dynamic documents



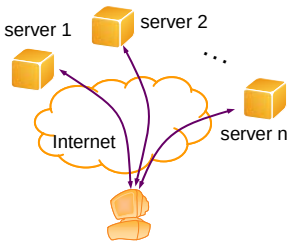
Web 2.0 Mashups

image source: flickr via http://metaatem.net/words/mashup

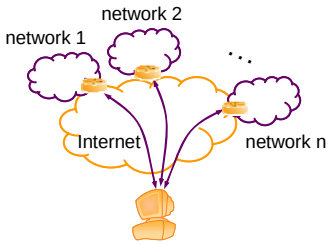
- re-mixing of existing services into new services



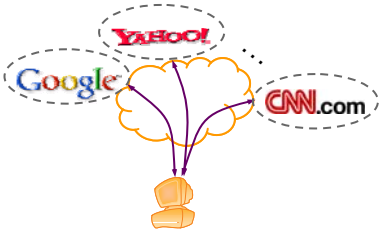
Different Notions of Locality



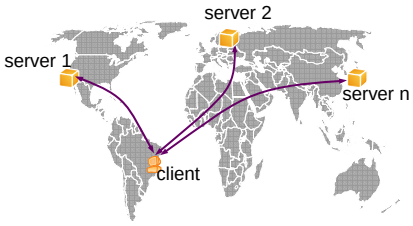
Hosts



Routing Domains



Organizations



Geographic Locations

Data sets

Trace A: Uni Münster 1998

Trace B: Uni Napoli, 2004

Trace C: top mashups, 2008
according to programmableweb.com

Trace D: US Web ~top 25, 2008
according to alexa.com

- Reflect usage by real users
  - many different Web sites
  - reflecting local preferences
- Anonymized client side traces
  - all client side traffic captured
  - all IP addresses obfuscated
  - no API traffic between Web servers
  - heuristic needed to distinguish clicks or site visits

- Directed experiments
  - select targets from top sites lists
  - use selected site for 2-5 min
  - avoid links pointing away from site
- One trace file per site visit
  - all client side traffic captured
  - real addresses seen
  - no API traffic between Web servers

- Both sets of traces: proxy based load balancing etc (multiple hosts hiding behind the same IP address) cannot be distinguished from single host

Evolution of Locality

Trace	year	$P\{N_{Server/click}>1\}$	$E[N_{Server/Click}]$
A	1998	0.22 ±0.02	1.4 ±0.05
B	2001	0.42 ±0.03	2.2 ±0.15
C	2008	0.60	4.7
D	2008	0.65	11.2

- number of servers contacted increased significantly over time

Top and bottom users of external services
(Trace D – Web top 25)

Site	Hosts	RDs	DDs	mHpRD	Conn.
photobucket.com	89	37	31	11	426
blogger.com	79	57	37	5	222
www.myspace.com	62	18	16	20	193
weather.com	59	28	23	9	485
aol.com	50	19	13	11	145
aim.com	49	21	17	11	148
wordpress.com	48	33	27	4	139
www.google.com (full)	43	24	8	7	126
msn.com	39	24	14	6	130
www.yahoo.com	37	17	7	7	139
...			...		
www.google.com (search only)	6	6	1	1	10
facebook.com	5	4	3	2	7
craigslist.org	5	5	4	1	42
wikipedia.org	2	1	1	2	7

Top external services used

… by Web top 25 sites (Trace D)

Service	Sites using this	Routing domains seen
google.com	19	22
doubleclick.net	13	7
rackspace.com	10	4
yahoo.com	9	18
pnap.net	9	7

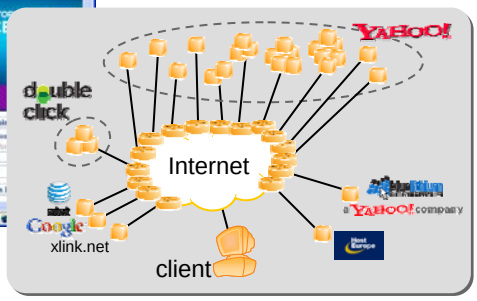
advertisement
hosting
CDN

… by top mashup sites (Trace D)

Service	Sites using this	Routing domains seen
google.com	14	13
yahoo.com	6	8
rackspace.com	3	2
aspadmin.com	3	3

hosting
hosting

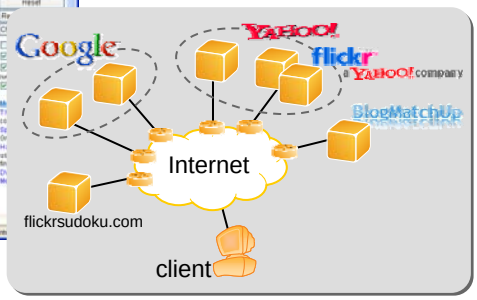
Locality example for top Web sites



http://www.yahoo.com/

Locality example for dedicated mashup site

http://www.flickrsudoku.com/



Conclusions

Observations

- steady growth of number of servers contacted over the past 10 years
- most Web sites are offering highly distributed services now
- many routing domains even for single service providers (google, yahoo)

Consequences

- IP edge routers definitely needed
- explicit QoS support (reservations) for Internet Web services is not feasible
- connection oriented approaches mostly for (semi) static backbone interconnections
- hard to determine required security settings for employing Internet Web services in enterprise environments
- availability challenges or benefits from high degree of distribution?

only considering Web traffic! There are other important applications around