Revisiting Cacheability in Times of User Generated Content

Bernhard Ager Fabian Schneider Juhoon Kim Anja Feldmann
{bernhard|fabian|jkim|anja}@net.t-labs.tu-berlin.de

Technische Universtität Berlin
Deutsche Telekom Laboratories

Global Internet Symposium 2010

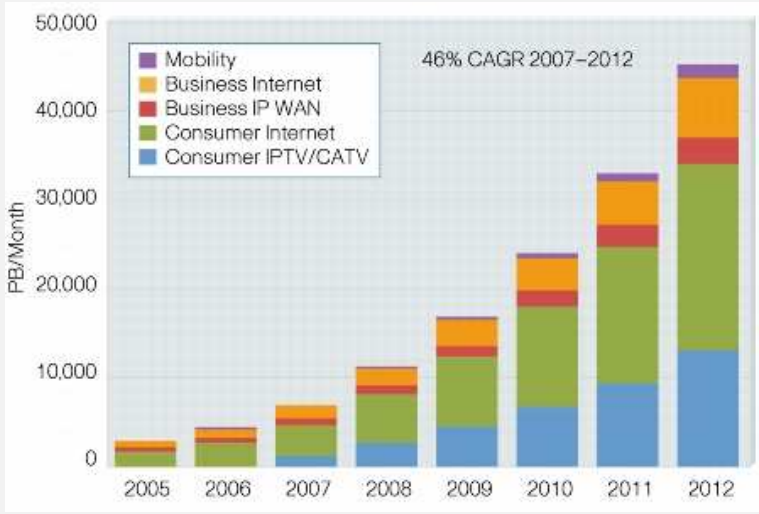
Motivation

- Internet traffic growing rapidly
- Providers must decide to
  - continuously upgrade infrastructure
  - search for possibilities to cut down traffic
- Caching can be used to cut down traffic w/o harming users.

⇒ Want to understand cacheability:

- Which applications?
- What is the potential?

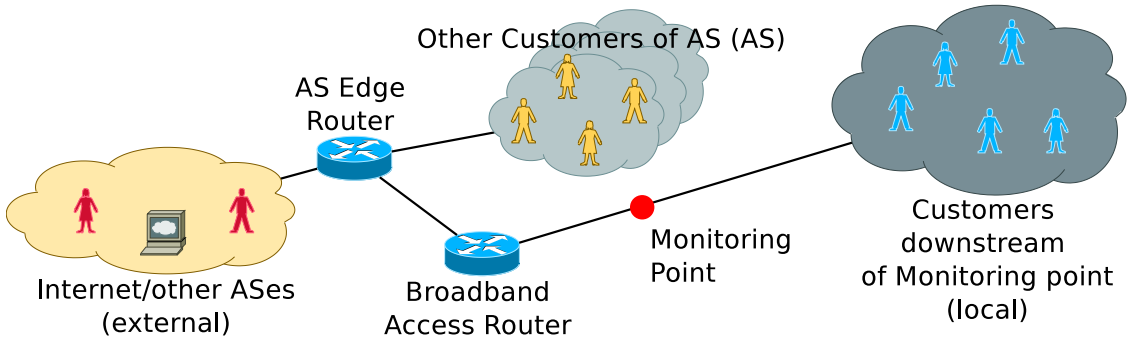
Global traffic growth



Source: Cisco Visual Networking Index: Forecast and Methodology, 2008-2013

Outline

Data



Trace properties

- anonymized header logs generated with Bro IDS
- 20.000 DSL lines
- Each protocol log covers two weeks
- Logs recorded within two month period

Trace statistics

Protocol	Log file	Volume
HTTP	$>$ 800 GB	$>$ 40 TB
NNTP	$>$ 2 GB	$>$ 2 TB
BitTorrent	$>$ 80 GB	$>$ 5 TB

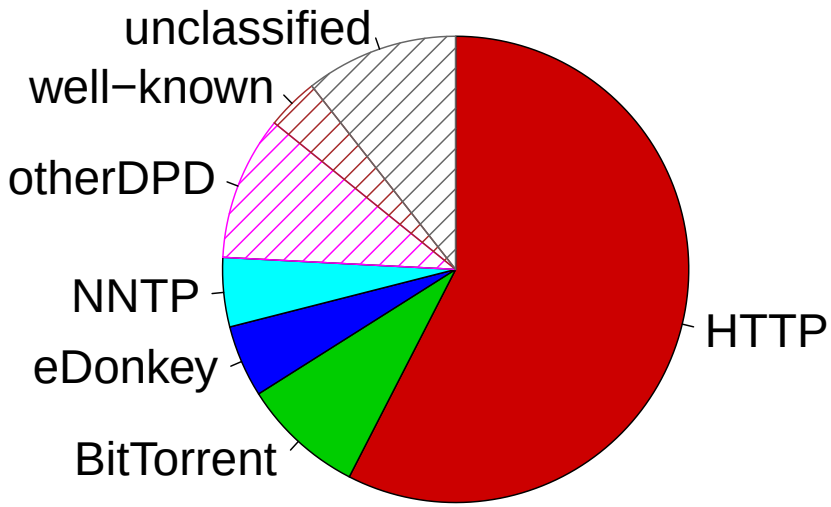
eDonkey: several 24 h and 48 h packet level traces

What can we cache?

Which protocols are worthwhile caching?

Heavy hitters [IMC'09]

- HTTP dominates: 58 %
- BitTorrent 9 %
- EDonkey 6 %
- NNTP 2 to 5 %
  (→ Fabian's talk)
- Each other protocol < 2 %
- Unclassified traffic < 11 %



Application mix in September 2008

[IMC'09] Gregor Maier, Anja Feldmann, Vern Paxson, and Mark Allman. On dominant characteristics of residential broadband internet traffic. In Proc. ACM Internet Measurement Conference, Nov 2009.

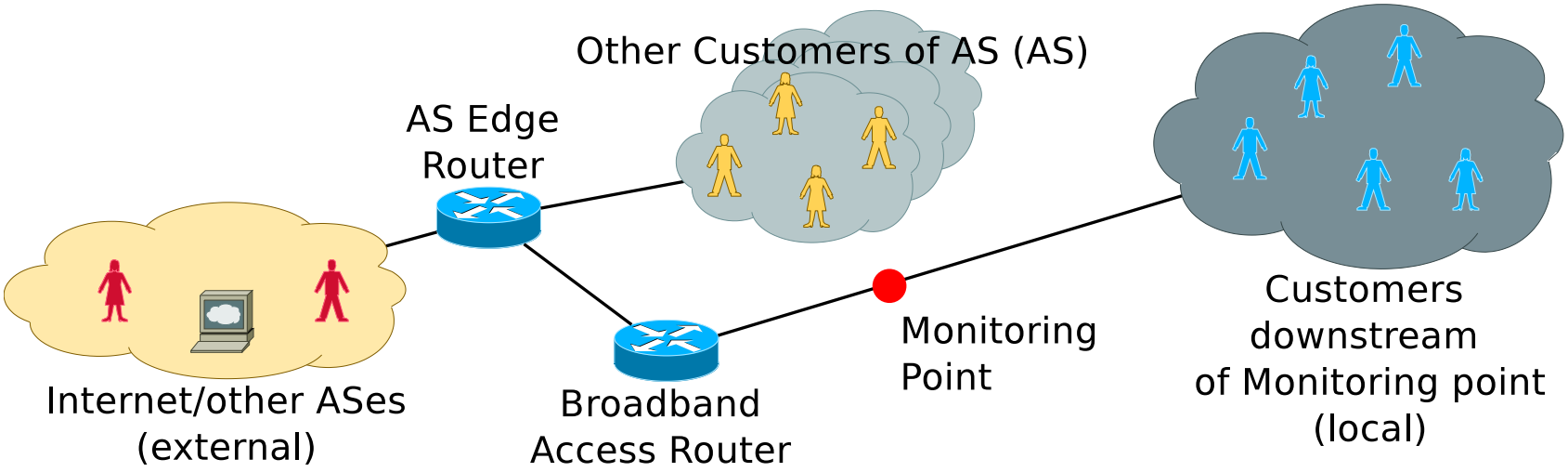
How to estimate the caching potential?

Caching assumptions

Goal: find upper limit of cacheability

- Cacheability: Saved download volume
- Unlimited disk space
- If item is downloaded n times, $n-1$ times are cacheable
  - For HTTP: also considering cache-control
- Possible deployment scenarios
  - Dedicated caches
  - P2P: traffic redirection
- We do not look at caching strategies!

Terminology

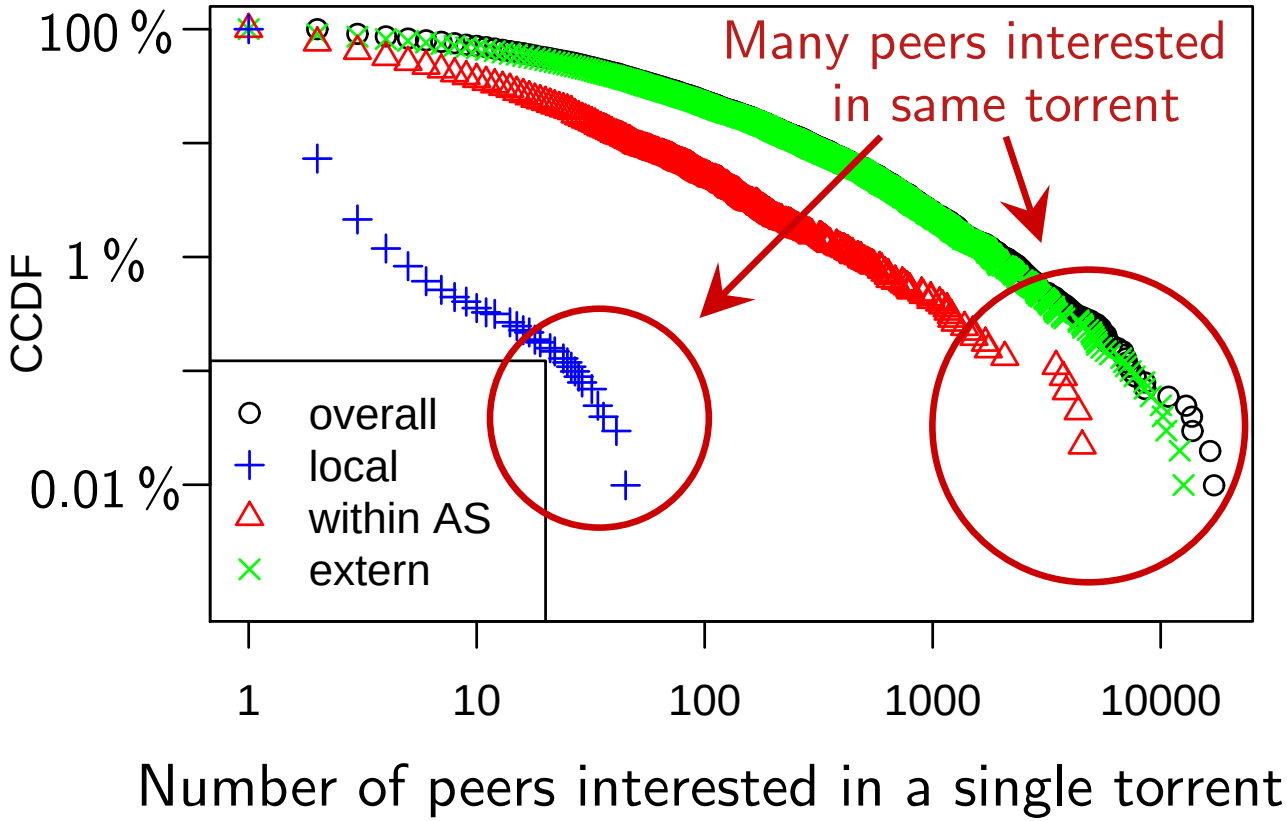


Locations

- Local: the local customer base
- AS: within the AS excluding Local
- External: outside AS

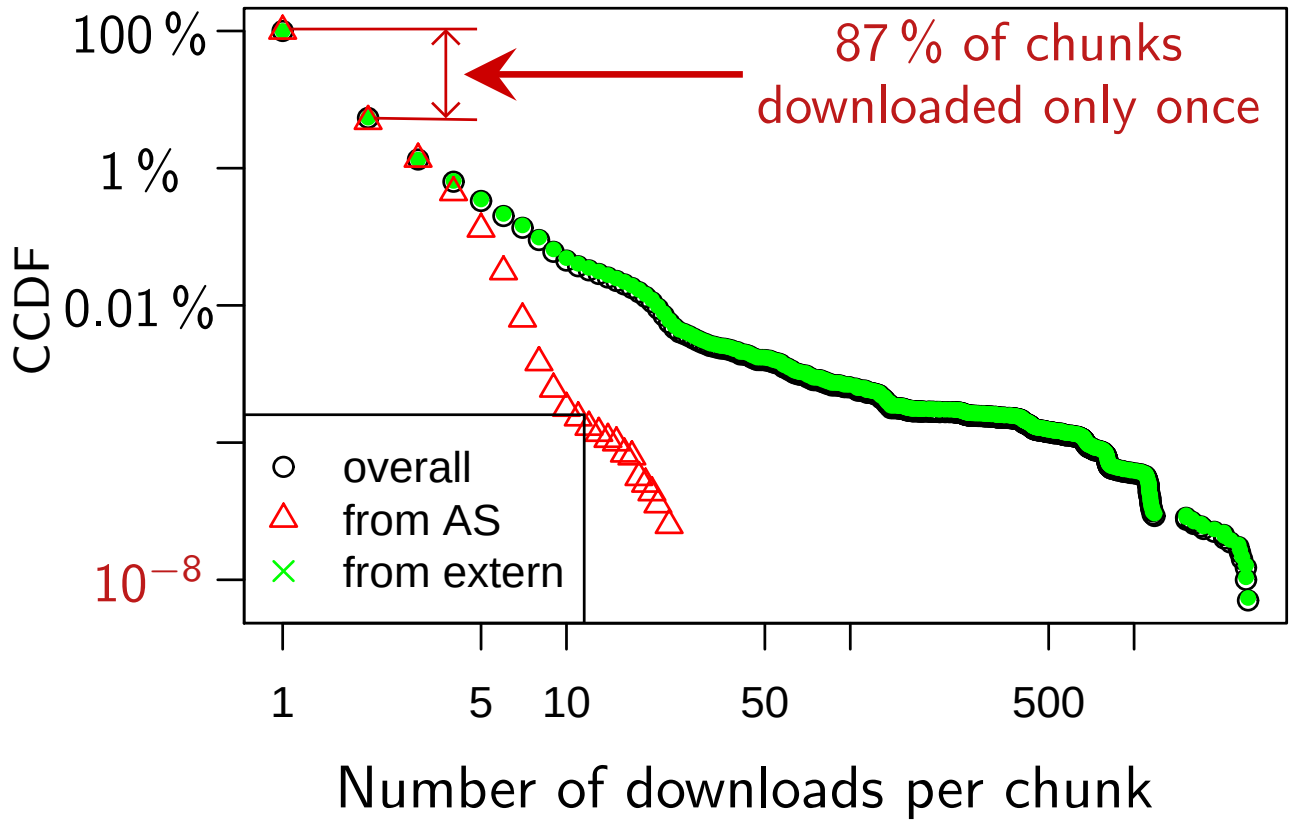
Outline

BitTorrent: Peers interested in a torrent



- High number of downloaders very promising
- 97 % cacheability AS wide, 27 % cacheability for Local customers

BitTorrent: Number of downloads per chunk



- Cacheability drops to 8 %
- But: many Local/AS peers available for seeding
- Prime cache with seeding peers: 85 % cacheability

HTTP and NNTP

Metrics and overall cacheability

Five caching metrics

scenario	object ID	HTTP method	return code	host	path	cache control
ideal	✔					
domain	✔			sld		
complete	✔	✔	✔	✔	✔	
full	✔	✔	✔	✔	✔	✔
realistic		✔	✔	✔	✔	✔

HTTP and NNTP

Metrics and overall cacheability

Five caching metrics

scenario	object ID	HTTP method	return code	host	path	cache control
ideal	✔					
domain	✔			sld		
complete	✔	✔	✔	✔	✔	
full	✔	✔	✔	✔	✔	✔
realistic		✔	✔	✔	✔	✔

Overall: NNTP

	ideal
Bytes	<7 %

Overall: HTTP

	ideal	domain	complete	full	realistic
Bytes	28 %	27 %	21 %	9.5 %	22 %

HTTP

Service types and optimizations

Caching potential by service (domain)

type of service	UGC	fraction	realistic
OCH1	✔	12.6 %	1.5 %
Video1	✔	10.8 %	3.8 %
Software1		2.8 %	64.8 %
CDN1	?	1.5 %	25.4 %
Search	?	1.0 %	32.7 %

Cacheability observations

- User Generated Content (UGC) appears hardly cacheable
- Software very good
- CDN1 and Search in between: Mixed content?

HTTP

Reasons for non-cacheability

Reasons for non-cacheability

- Load balancing over different host names
- URL parameters
  - session management
  - website dynamics
  - URL validity timeouts
- Cache-control has a severe impact
- Lack of cache control $\Rightarrow$ not cacheable

HTTP

Service types and optimizations

Caching potential by service (domain)

type of service	UGC	fraction	realistic	optimized
OCH1	✔	12.6 %	1.5 %	+18.4 %
Video1	✔	10.8 %	3.8 %	+24.6 %
CDN1	?	1.5 %	25.4 %	+23.7 %

Cache optimization heuristics

Allow violations of strict caching policies:

- Video1: strip personalization and DNS load balancing
- OCH1, CDN1: opportunistic expiries (here: ∞)
  - OCH1: misconfigured download accelerators
  - Overall effect: 4 %

Summary

Conclusions

- P2P caching
  - Very promising: $> 95\,\%$
  - However limited when only considering current downloads: $< 10\,\%$
  - Proper mechanisms allow caching-rates up to $85\,\%$
- HTTP caching not necessarily beneficial, contrary to recent work
  - Overall cacheability: $22\,\%$
  - Software downloads very good cacheable: $> 60\,\%$
  - User generated content very hard to cache: $< 10\,\%$
  - Opportunistic heuristics may help

Summary

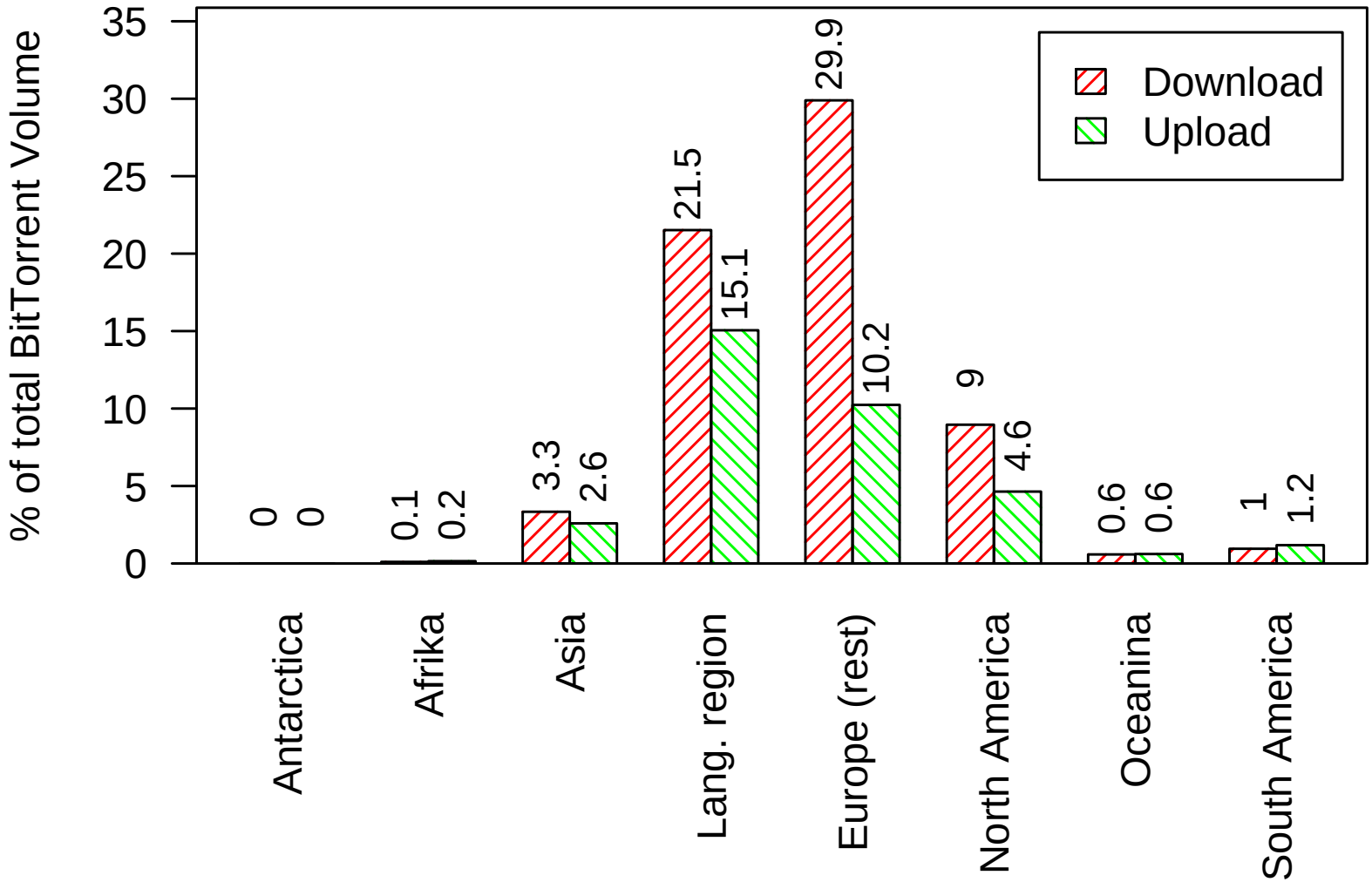
Conclusions

- P2P caching
  - Very promising: $> 95\,\%$
  - However limited when only considering current downloads: $< 10\,\%$
  - Proper mechanisms allow caching-rates up to $85\,\%$
- HTTP caching not necessarily beneficial, contrary to recent work
  - Overall cacheability: $22\,\%$
  - Software downloads very good cacheable: $> 60\,\%$
  - User generated content very hard to cache: $< 10\,\%$
  - Opportunistic heuristics may help

Questions?

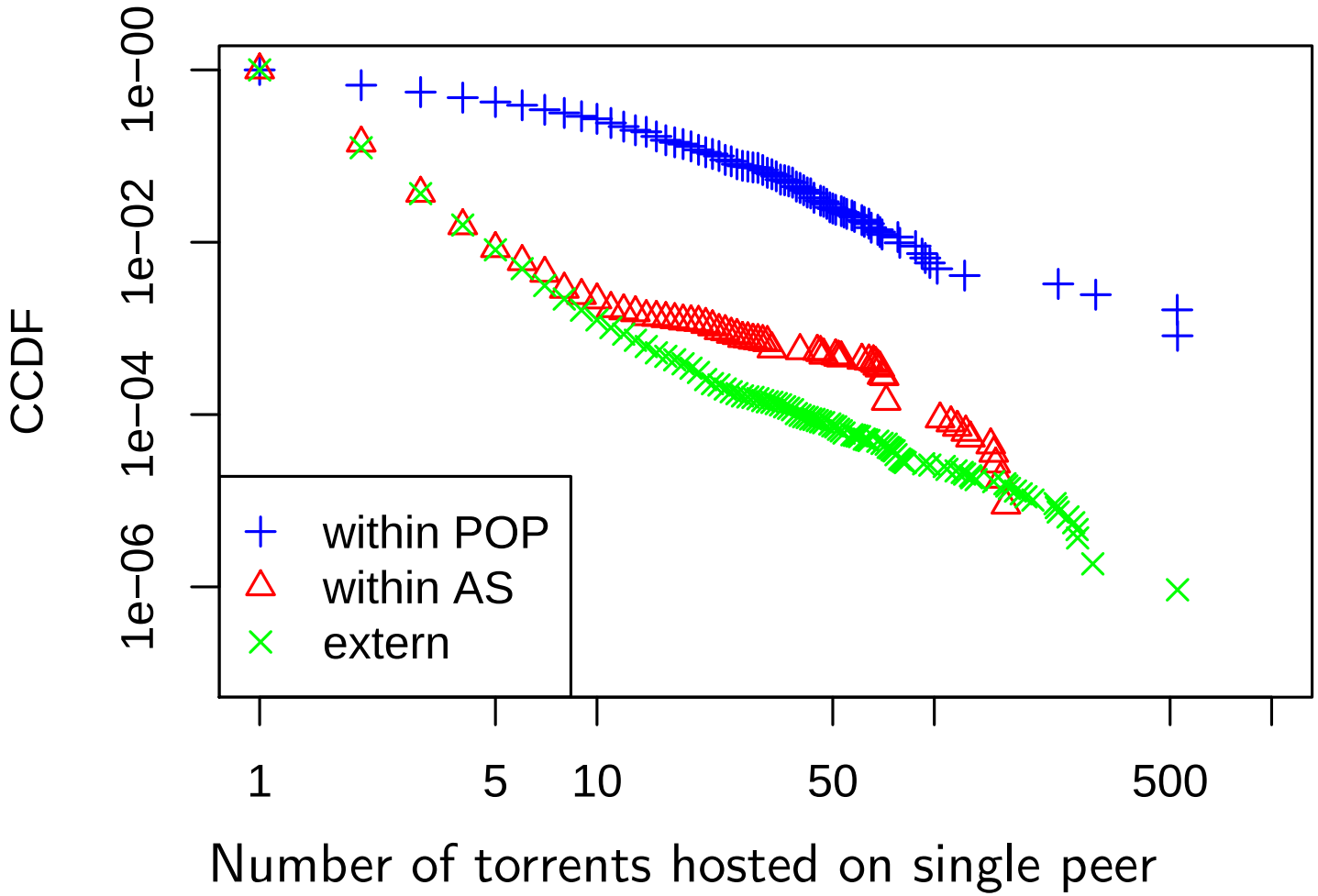
BitTorrent

Traffic prefers to stay in Europe



BitTorrent

Some peers host many torrents



HTTP

Influence of population size on cacheability