Chadi Barakat
Ian Pratt (Eds.)

LNCS 3015

Passive and Active Network Measurement

5th International Workshop, PAM 2004
Antibes Juan-les-Pins, France, April 2004
Proceedings

Springer

Chadi Barakat Ian Pratt (Eds.)

Passive and Active Network Measurement

5th International Workshop, PAM 2004
Antibes Juan-les-Pins, France, April 19-20, 2004
Proceedings

Springer

Volume Editors

Chadi Barakat
INRIA Sophia Antipolis - Planète groupe
2004, route des lucioles, 06902 Sophia Antipolis, France
E-mail: chadi.barakat@inria.fr

Ian Pratt
University of Cambridge, Computer Laboratory
JJ Thomson Avenue, Cambridge CB3 0FD, UK
E-mail: ian.pratt@cl.cam.ac.uk

Library of Congress Control Number: 2004103365

CR Subject Classification (1998): C.2, C.4

ISSN 0302-9743
ISBN 3-540-21492-5 Springer-Verlag Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by PTP-Berlin, Protago-TeX-Production GmbH
Printed on acid-free paper SPIN: 10997468 06/3142 5 4 3 2 1 0

Preface

PAM 2004 was the 5th International Workshop on Passive and Active Measurement, held in Juan-les-Pins on the French Riviera, co-organized by the University of Cambridge and INRIA-Sophia Antipolis, with financial support from Intel and Cisco Systems.

This year we received a record number of submissions (184), reflecting the growth of the field and the critical role it plays in maintaining the network infrastructure on which we all rely. From the two-page abstracts submitted, the programme committee selected 29 papers whose authors were invited to submit full papers to appear in these proceedings. Particular emphasis was placed on selecting work that we felt was fresh and exciting, so as to encourage a dynamic and interactive workshop that provided a first public presentation of research that will go on to appear in other, more formal conferences and journals. The programme committee was greatly impressed with the strength and depth of submissions received, which bodes well for the future of the subject area.

This workshop took place during April 19–20, 2004 in Juan-les-Pins. Located between the Alps and the Mediterranean, and close to Nice, Cannes and Monaco, Juan-les-Pins is one of the most beautiful sites on the French Riviera. Juan-les-Pins is also close to Sophia Antipolis, the French telecom valley.

The workshop could not have succeeded without the support of many people whom we would like thank. First, we thank the members of the programme committee for donating a considerable amount of time to review the unexpectedly large number of submissions, while working to a very tight deadline.

Second, we would like to thank the University of Cambridge and INRIA Sophia Antipolis for their support. In particular, special thanks go to Tim Granger for running the conference Web site, and writing the tools used to process the paper submissions and collate reviews. Dany Sergeant from INRIA kindly took care of the registration and financial issues. Gianluca Iannaccone chaired the student travel support panel, and Andrew Moore collated the published proceedings. Joerg Micheel of Endace ran the e-mail MTA for the pam2004.org domain.

We are very grateful to Cisco Systems and Intel for their generous financial support that enabled us to assist 15 graduate students in attending the workshop.

Finally, our thanks go to the authors of the submitted papers, to the speakers, and to all the participants for making PAM 2004 a success!

We hope you enjoyed the PAM 2004 workshop, and had a pleasant stay on the French Riviera.

April 2004

Chadi Barakat (General Chair)
Ian Pratt (PC Chair)

Organization

PAM 2004 was organized by the University of Cambridge Computer Laboratory and INRIA Sophia Antipolis.

Program Committee

Chadi Barakat (INRIA, France)
Andre Broido (CAIDA, USA)
Nevil Brownlee (University of Auckland, New Zealand)
Randy Bush (IIJ, Japan)
Avi Freedman (Akamai, USA)
Ian Graham (University of Waikato, New Zealand)
Geoff Huston (Telstra, Australia)
Gianluca Iannaccone (Intel Research, UK)
Peter Key (Microsoft Research, UK)
Yasuichi Kitamura (APAN, Japan)
Joerg Micheel (NLANR/Endace, USA)
Sue Moon (KAIST, Korea)
Andrew Moore (University of Cambridge, UK)
Ian Pratt (University of Cambridge, UK)
Ronn Ritke (NLANR/MNA, USA)
Nicolas Simar (DANTE, UK)
Alex Tudor (Agilent, USA)
Daryl Veitch (University of Melbourne, Australia)

Additional Reviewers

M. Crovella
J. Crowcroft
C. Filsfils
T. Griffin
A. Al Hamra
V. Kanevsky
J. Li
V. Ramos
L. Rizzo

Sponsoring Institutions

Intel Research, Cambridge, UK
Cisco Systems, San Jose, California, USA

Table of Contents

Session 1: P2P and Overlay

Session 2: Network Optimization

Session 3: Traffic Analysis

Session 4: Protocol and System Measurement

Session 5: Tools

Session 6: Miscellaneous

Session 7: Network Measurement

Session 8: BGP and Routing

Dissecting BitTorrent: Five Months in a Torrent's Lifetime

M. Izal, G. Urvoy-Keller, E.W. Biersack, P.A. Felber, A. Al Hamra, and L. Garcés-Erice

Institut Eurecom, 2229, route des Crêtes, 06904 Sophia-Antipolis, France
{izal,urvoy,erbi,felber,alhamra,garces}@eurecom.fr

Abstract. Popular content such as software updates is requested by a large number of users. Traditionally, to satisfy a large number of requests, lager server farms or mirroring are used, both of which are expensive. An inexpensive alternative are peer-to-peer based replication systems, where users who retrieve the file, act simultaneously as clients and servers. In this paper, we study BitTorrent, a new and already very popular peer-to-peer application that allows distribution of very large contents to a large set of hosts. Our analysis of BitTorrent is based on measurements collected on a five months long period that involved thousands of peers. We assess the performance of the algorithms used in BitTorrent through several metrics. Our conclusions indicate that BitTorrent is a realistic and inexpensive alternative to the classical server-based content distribution.

1 Introduction

BitTorrent [4] is a file distribution system based on the peer-to-peer (P2P) paradigm. BitTorrent has quickly emerged as a viable and popular alternative to file mirroring for the distribution of large content, as testified by the numerous Web sites that host active "torrents" (e.g., `http://f.scarywater.net/`).

We have conducted a comprehensive analysis of BitTorrent to assess its performance. To that end, we have used two sources of information. First, we have obtained the "tracker" log of arguably the most popular torrent (BitTorrent session) so far—the 1.77GB Linux Redhat 9 distribution—for its first 5 months of activity. The log contains statistics for more than 180,000 clients, and most interestingly, it clearly exhibits an initial flash-crowd period with more than 50,000 clients initiating a download in the first five days. This first source of information allows us to estimate the global efficiency of BitTorrent, the macroscopic behavior of clients, and the scalability of a P2P application under flash-crowd conditions. Our second source of information consists of data collected with a modified client that participated to the same torrent downloading Redhat 9. This second log allows us to study the direct interactions between the clients. The remaining of the paper is organized as follows. In Section 2, we present the main features of BitTorrent. In Section 3, we review the related work. In Section 4, we present the results obtained from the tacker log and in Section 5 the conclusions obtained from our client log. We conclude with future directions in Section 6.

C. Barakat and I. Pratt (Eds.): PAM 2004, LNCS 3015, pp. 1–11, 2004.

2 BitTorrent

BitTorrent is a P2P application that capitalizes the *resources* (access bandwidth and disk storage) of peer nodes to efficiently distribute large contents. There is a separate torrent for each file that is distributed. Unlike other well-known P2P applications, such as Gnutella or Kazaa, which strive to quickly *locate* hosts that hold a given file, the sole objective of BitTorrent is to quickly *replicate* a single large file to a set of clients. The challenge is thus to maximize the speed of replication.

A torrent consists of a central component, called *tracker* and all the currently active peers. BitTorrent distinguishes between two kinds of peers depending on their download status: clients that have already a complete copy of the file and continue to serve other peers are called *seeds*; clients that are still downloading the file are called *leechers*. The tracker is the only centralized component of the system. The tracker is not involved in the actual distribution of the file; instead, it keeps meta-information about the peers that are currently active and acts as a rendez-vous point for all the clients of the torrent.

A user joins an existing torrent by downloading a *torrent* file (usually from a Web server), which contains the IP address of the tracker. To initiate a new torrent, one thus needs at least a Web server that allows to discover the tracker and an initial seed with a complete copy of the file. To update the tracker's global view of the system, active clients periodically (every 30 minutes) report their state to the tracker or when joining or leaving the torrent. Upon joining the torrent, a new client receives from the tracker a list of active peers to connect to. Typically, the tracker provides 50 peers chosen at random among active peers while the client seeks to maintain connections to 20−40 peers. If ever a client fails to maintain at least 20 connections, it recontacts the tracker to obtain additional peers. The set of peers to which a client is connected is called its *peer set*.

The clients involved in a torrent cooperate to replicate the file among each other using *swarming* techniques: the file is broken into equal size chunks (typically 256kB each) and the clients in a peer set exchange chunks with one another. The swarming technique allows the implementation of parallel download [7] where different chunks are simultaneously downloaded from different clients. Each time a client obtains a new chunk, it informs all the peers it is connected with. Interactions between clients are primarily guided by two principles. First, a peer preferentially sends data to peers that reciprocally sent data to him. This "tit-for-tat" strategy is used to encourage cooperation and ban "free-riding" [1]. Second, a peer limits the number of peers being served simultaneously to 4 peers and continuously looks for the 4 best downloaders (in terms of the rate achieved) if it is a seed or the 4 best uploaders if it is a leecher.

BitTorrent implements these two principles, using a "choke/unchoke" policy. "Choking" is a temporary refusal to upload to a peer. However, the connection is not closed and the other party might still upload data. A leecher services the 4 best uploaders and chokes the other peers. Every 10 seconds, a peer re-evaluates the upload rates for all the peers that transfer data to him. There might be more than 4 peers uploading to him since first, choking is not necessarily reciprocal

and second, peers are not synchronized. [1] He then chokes the peer, among the current top 4, with the smallest upload rate if another peer offered a better upload rate. Also, every 3 rounds, that is every 30 seconds, a peer performs an *optimistic unchoke*, and unchokes a peer regardless of the upload rate offered. This allows to discover peers that might offer a better service (upload rate). Seeds essentially apply the same strategy, but based solely on download rates. Thus, seeds always serve the peers to which the download rate is highest.

Another important feature of BitTorrent is the chunk selection algorithm. The main objective is to consistently maximize the entropy of each chunk in the torrent. The heuristic used to achieve this goal is that a peer always seeks to upload the chunk the least duplicated among the chunks it needs in its peer set (keep in mind that peers only have a local view of the torrent). This policy is called the *rarest first policy*. There exists an exception to the rarest first policy when a peer joins a torrent and has no chunks. Since this peer needs to quickly obtain a first chunk (through optimistic unckoke), it should not ask for the rarest chunk because few peers hold this chunk. Instead, a newcomer uses a random first policy for the first chunk and then turns to the rarest first policy for the next ones.

3 Previous Work

Approaches to replicate contents to a large set of clients can be classified as client side and server side approaches. The first client-side approach was to cache contents already downloaded by clients of a given network. A symmetric approach on the server side is to transparently redirect clients to a set of mirror sites run by a content provider (e.g. Akamai).

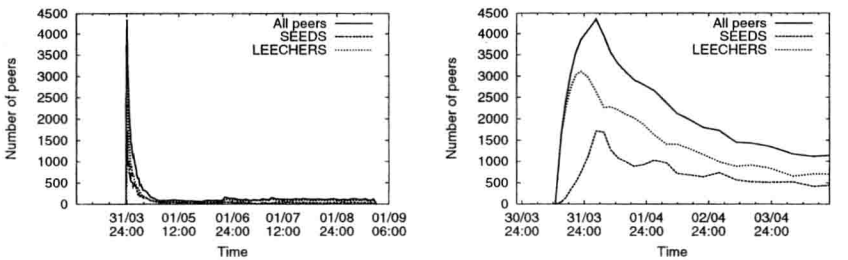
The peer-to-peer paradigm has been applied to obtain client side and server side solutions. On the server side, one finds proposals like [5] where overlay nodes dynamically organize themselves so as to form a tree with maximum throughput. FastReplica [2] is designed to efficiently replicate large contents to a set of well-known and stable overlay nodes.

On the client side, one finds many proposals to build application-layer multicast services [8,6,3]. A solution similar to BitTorrent is Slurpie [9]. Slurpie aims at enforcing cooperation among clients to alleviate the load of a Web server that is the primary source of the content. The algorithms of Slurpie are much more complex than the ones of BitTorrent and require to estimate the number of peers in the Slurpie network. The expected improvements over BitTorrent is less load on the topology server (equivalent to the BitTorrent tracker) and on the primary source (original seed) through a back-off algorithm. Results are promising since Slurpie is able to outperform BitTorrent in a controlled environment. Still, the actual performance of Slurpie in case of flash crowds and for a large number of clients is unknown.

[1] For instance, a peer that was offering a very good upload rate has decided to choke this connection just 1 second before the reevaluation on the other side and thus he might still be in the top 4 list for the next 10 seconds and received service.

4 Tracker Log Analysis

The tracker log covers a period of 5 months from April to August 2003. The corresponding torrent has as content the 1.77 GB Linux Redhat 9 distribution. 180,000 clients participated to this torrent with a peak of 51,000 clients during the first five days (see Figures 1(a) and 1(b)). These first five days clearly exhibits a flash-crowd. As clients periodically report to the tracker their current state, along with the amount of bytes they have uploaded and downloaded, the tracker log allows us to observe the global evolution of the file replication process among peers.



(a) Complete trace

(b) Zoom on the first five days

Fig. 1. Number of active peers over time

4.1 Global Performance

Analyzing the tracker log, our first finding is that BitTorrent clients are altruistic in the sense that they actively send data to other clients, both as leechers and as seeds. Altruism is enforced during the download phase by the tit-for-tat policy, as a selfish client will be served with a very low priority. Once they become seed, the peers remain connected for another six and a half hours on average. This "social" behavior can be explained by two factors: first, the client must be explicitly terminated after completion of the download, which might well happen while the user is not at his computer, e.g., overnight; second, as the content being replicated is perfectly legal, the user has no particular incentive to quickly disconnect from the torrent. In fact, the presence of seeds is a key feature, since it greatly enhances the upload capacity of this torrent and the ability to scale to large client populations. Over the 5 months period covered by the log file, we observed that the seeds have contributed more than twice the amount of data sent by leechers (see Figure 2). We also observed that the proportion of seeds is consistently higher than 20%, with a peak at 40% during the first 5 days

(see Figure 3). This last figure clearly illustrates that BitTorrent can sustain a high flash-crowd since it quickly creates new seeds. To put it differently, in situations where new peers arrive at a high rate, the resources of the system are not divided evenly between clients, which would result, like in a processor sharing queue under overload, in no peers completing the download. On the contrary, older peers have a higher priority since they hold more chunks than younger peers, which gives them more chance to complete the download and become seeds for newcomers. Obviously, this strategy benefits from the cooperation of users that let their clients stay as seeds for long periods of time.

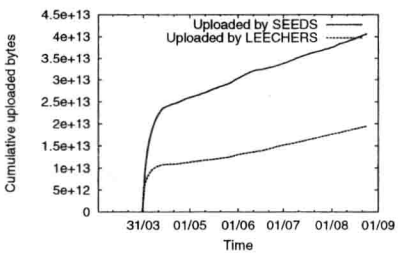


Fig. 2. Volumes uploaded by seeds and leechers

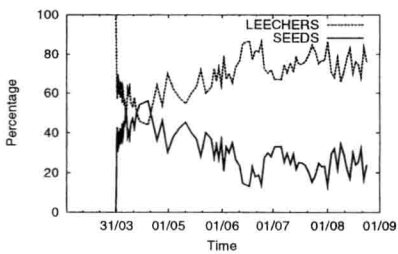


Fig. 3. Proportions of seeds and leechers

4.2 Individual Session Performance

A fundamental performance metric for BitTorrent is the average download rate of leechers. Over the 5 months period covered by the tracker log, we observed an average download rate consistently above 500kb/s. This impressive figure indicates that most of the BitTorrent clients have good connectivity (at least ADSL), which is not surprising given the total size of the file. Moreover, BitTorrent exhibits good scalability: during the initial flash-crowd, the average download rate was close to 600kb/s and the aggregate download rate of all simultaneously active clients was more than 800Mb/s.

BitTorrent is thus able to capitalize the bandwidth of end system hosts to achieve a very high aggregate throughput. Still, the final objective of BitTorrent is to replicate a given content over all the peers in a torrent. We thus need to know how many peers eventually completed the download and the fraction of bytes that the peers that did no complete the transfer downloaded. Since BitTorrent allows users to suspend and resume the download at any time, a peer might complete the download after one or several sessions. In the tracker log, a session id identifies the session (hash of the IP address of the host and its current time), along with the IP address of the host as seen by the tracker. Thus, identifying single session downloads is easy based on the session id. To reconstruct multi-sessions downloads, we assumed that the IP address of the host does not change

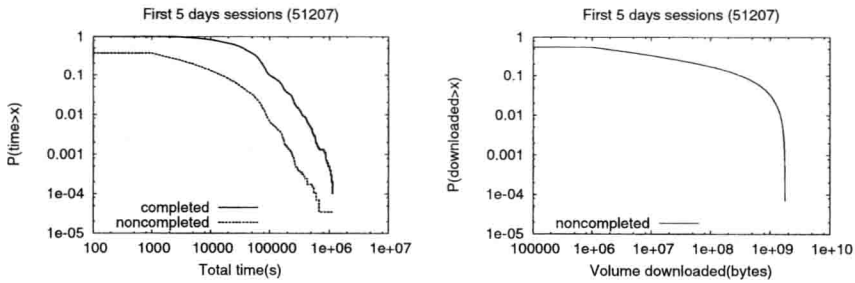
from one session to another. NATs often used port numbers to disambiguate connexions and not IP addresses, so our assumption can hold even in this case. Of course, if two or more hosts behind a firewall simultaneously participate to the same torrent, it might be difficult to disambiguate them. In addition, since a peer provides the amount of bytes it has downloaded in each of its reports (sent every 30 minutes), we mated two sessions with the same IP if the amount of bytes of the last report for the first session is close to the amount of bytes of the first report of the second session. These two values are not necessarily the same if for instance, the user disconnected the BitTorrent client improperly which results in the latter not sending its disconnection report with its current amount of bytes downloaded. A set of sessions form a multi-sessions download if the download of the file is completed during the last session (on average, we found that a multi-session download consists of 3.9 sessions). We also observed some sessions that started with already 100% of the file downloaded. This is the case if the user is kind enough to rejoin the torrent after the completion of the download to act as a seed. We thus categorize the sessions into four categories : single session download, multi-session download, seed sessions and session that never complete (incomplete downloads). We provide statistics for these sessions in Table 1.

Table 1. Sessions types and their characteristics

Type	Number of sessions	Total down. (TB)	Total up. (TB)	Down. /session (MB)	Up. /session (MB)	Down. rate (kb/s)	Up. rate (kb/s)	Down. Duration (hours)
Single session downloads	20584	34.7	28.5	1765.7	1450.2	1331.1	325. 2	8.1
Multi-session downloads	14285	6.2	4.3	455.2	319.2	390.6	145.0	19.2
Incomplete downloads	138423	11.2	9.1	84.5	69.1	114.7	50.5	-
Seed sessions	8555	-	12.7	-	1556.6	-	223.6	-
Total	181847	52.1	54.6	-	-	-	-	-

Overall, Table 1 indicates that only 19% of the sessions are part of a transfer that eventually completed. However, the 81% of sessions that did not complete the transfer represent only 21.5% of the total amount of downloaded bytes. Moreover, the incomplete sessions uploaded almost as much bytes as they downloaded. It is difficult to assess the reason behind the abortion of a transfer. It might be because the user eventually decides he is not interested in the file or because of a poor download performance. To investigate this issue, we look at the statistics (durations and volumes) for sessions that start with 0% of the file during the first five days. We classify these sessions into two categories denoted as "completed" and "non-completed". The completed category corresponds to the single session downloads defined previously while the non-completed category corresponds to transfers that never complete or the first session of transfers that will eventually complete. We present on Figure 4 the complementary cumulative distribution functions for the download volumes and session durations

for the completed and non-completed sessions. We can observe from this figure that 90% of the non-completed sessions remain less than 10,000 seconds in the torrent (and 60% less than 1000 seconds) and retrieve less than 10% of the file. Most of the non-completed sessions are thus short and little data is downloaded.



(a) Duration

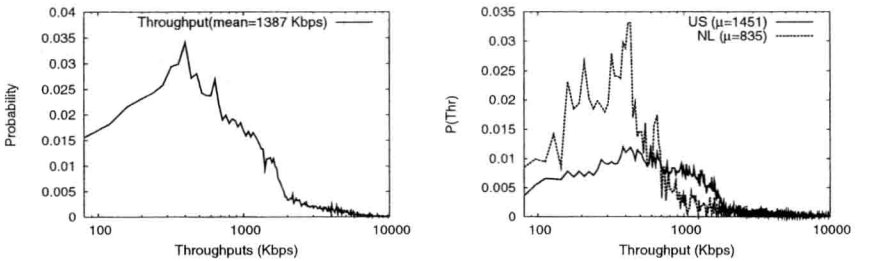
(b) Download volume

Fig. 4. Single sessions statistics for sessions seen during the first 5 days

Let us now further focus on the performance achieved by single session downloads. The average download rate of these 20,584 sessions is close to 1.3Mb/s over the whole trace, which is larger than the download rate averaged over all sessions (500kb/s). The average download time for single session download is about 29,000 seconds overall. Such values reveal a high variability in the download rates achieved by these sessions since if they had all achieved a download rate of 1.3Mb/s, the average download time would be $\frac{1.77\text{GB}}{1.3\text{Mb/s}} \sim 10,000$ seconds. This is confirmed by the distribution of these download throughputs (see Figure 5(a)) that clearly exhibits a peak around 400kb/s (a typical value for ADSL lines), a value significantly smaller than the mean.

4.3 Geographical Analysis

The tracker log provides the IP addresses of the clients. Based on these addresses, we have used NetGeo (`http://www.caida.org/tools/utilities/netgeo/`) to obtain an estimation of the origin country of the peers that participated to the torrent. The estimation might be imprecise as NetGeo is not maintained any more. Overall, we were not able to obtain information for around 10% of the hosts. In Table 2 we indicate the top five countries for the first 5 days, the first 4 weeks and the complete trace. We can observe that this set of countries, as well as the ranking is consistently the same for all three time scales. Most of the clients are US clients while Europe is represented only through the Netherlands. Still, for Netherlands, 50% of the hosts originate from ripe.net (an information



(a) Global throughput distribution

(b) Throughput distribution per cluster

Fig. 5. Throughputs for single session downloads

also provided by NetGeo), which acts as the Regional Internet Registry for the whole of Europe, Middle East, Africa and part of Asia also. Thus we can guess that most of these peers are spread all over Europe.

Table 2. Geographical origins of peers

Countries	First week	First four weeks	Complete trace
United States	44.6%	44.3%	32%
Netherlands (Europe)	14.9%	15.3%	23.9%
Australia	12.7%	12.6%	17.8%
Canada	5.7%	5.8%	4.9%
% of total hosts	77.9%	78%	78.6%

We next investigate the relative performance of the four clusters identified above (US, NL, AU and CN). To do so, we consider again the 20,584 peers that complete the download in a single session. 17,000 peers out of the 20,584 initial peers, are in the four clusters (11498 in US cluster, 2114 in NL, 1995 in AU and 1491 in CM). We plot in Figure 5(b) the distribution of download throughputs achieved by the peers in the NL and US clusters (the AU and CN clusters are highly similar to the US cluster and not depicted for sake of clarity). Figure 5(b) reveals that the download throughput of the hosts in the NL cluster is significantly smaller than the throughput of the hosts in the US cluster (where more mass is on the right side of the curve). This can indicate that clients in the US have, in general, better access links than in Europe.

5 Client Log Analysis

To better observe the individual behavior of a BitTorrent peer, we have run an instrumented client behind a 10Mb/s campus network access link. Our client joined the torrent approximatively in the middle of the 5 months period (i.e., far after the initial flash crowd). We experienced a transfer time of approximately 4,500 seconds (i.e., much lower than the average download times previously mentioned) and our client remained connected as a seed for 13 hours. Our client logged detailed information about the upload or download of each individual chunk. In Figure 6, we represent the number of peers with whom our client is trading. At time 0, our client knows around 40 peers whose addresses have been provided by the tracker. We next continuously discover new peers (peers that contacted us) up to the end of the download (at time 4,500 seconds) where we observe a sudden decrease of the number of peers, most probably because the seeds we where connected to have closed their connection to us as soon as we have completed the download. After the download, we stay connected as seed to between 80 and 95 leechers (4 of them being served while the others are choked).

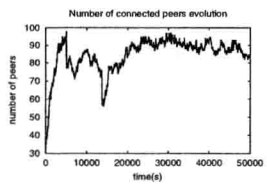


Fig. 6. Number of peers during and after download

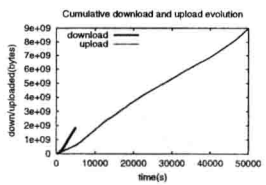


Fig. 7. Complete torrent

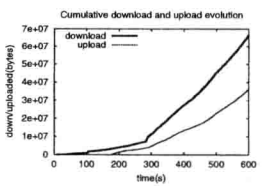


Fig. 8. First 10 minutes of download

Figures 7 and 8 show the amount of bytes downloaded and uploaded with respect to time for the complete trace and the first 10 minutes of the trace respectively. From these figures, we can draw several conclusions:

- There is a warm-up period (around 100 seconds) to obtain some first chunks. But as soon as the client has obtained a few chunks, it is able to start uploading to others peers. This means that the rarest first policy works well, otherwise our client would maybe not find anyone interested in its first chunks.
- The download and upload rates are (positively) correlated, which indicates that the tit-for-tat policy works. It also indicates that we always find peers interested in our chunks and peers from which we can download chunks. Otherwise, we would observe some periods where the curves are flat, indicating that our client is stalled. Also, the way peer sets are built with "old" and "new" peers mixed together must have a positive impact on the efficiency of BitTorrent. Indeed, a torrent can be viewed as a collection of interconnected sets of peers. A peer that joins a torrent will obviously be the youngest peer in its initial set, but it may later be contacted by younger peers. It may also