Design and Study of a BitTorrent-like File Sharing System

Yuzhou Wang
School of Electrical and Computer Engineering
University of Waterloo
Waterloo, ON, Canada
Email: y2345wan@uwaterloo.ca

Sainan He
School of Electrical and Computer Engineering
University of Waterloo
Waterloo, ON, Canada
Email: s66he@uwaterloo.ca

***Abstract*—The abstract goes here.**

I. Introduction

This demo file is intended to serve as a "starter file" for IEEE conference papers produced under LaTeX using IEEEtran.cls version 1.8b and later. I wish you the best of success.

mds

August 26, 2015

II. Terminology overview

A. BitTorrent

BitTorrent is a peer-to-peer(P2P) file sharing mechanism, which is designed to facilitate large file transfers and make it more efficient among multiple peers. A large file is divided into blocks with same size except the last block. Based on this assumption, when peers download blocks from the server, they also upload to each other concurrently. Thus, the system achieves selfscaling as its serving capability grows with peers increasing.

The most significant role in the system is called tracker[1], which keeps track of peers to maintain lists of alive peers, those are corresponding to a certain file, either download or upload it. Since peers send messages periodically to and new one also contact with the tracker, the lists will always keep updated. The contact information of the tracker and detail of a large file is stored in a related metainfo file known as .torrent. A peer is any BitTorrent client which takes part in a download task. It can play two roles in the system, seeder and leecher. Any peer which has and upload a complete file to those seek to download is a seeder. In contrast, those who send a request to the tracker for downloading and get a response with the list mentioned before to contact with other peers is a leecher. However, it will upload the blocks of file it owns to other leechers.

There are three main policies in this scalable file distribution protocol, Local Rarest First(LRF), Tit-for-Tat(TFT) and Choking policy. As a result of that LRF policy enables peers to choose rarest blocks preferentially rather than at random to download, the system avoids a circumstance where it is likely to take some time for peers to find blocks with few replicas and slow down the download rate. TFT policy is proposed to solve free-riding problem in which some leechers are selfish by only downloading without serving others. The aim of it is to achieve a cooperation that upload bandwidth is exchange with download bandwidth. This strategy is complied by fair trading, whereby peers prefer allocating upload slots within threshold to those also send data at a fast rate in return. As to Choking policy, it is raised to assist TFT to control the number of active connections among peers. Download rate and upload rate are used respectively to determine which remote peer to be choked related to whether the peer has a complete copy of the file or not. These strategies allow BitTorrent to use bandwidth between peers (i.e., perpendicular bandwidth) effectively[2] and handle flash crowds well.

B. ZooKeeper

ZooKeeper is a centralized service deployed to assist large-scale distributed system in process coordination across unreliable networks. It provides naming registry, maintaining configuration and synchronization services to alleviate management complexity.

It is developed to have wait-free property related to shared registers along with driven-by-event one, which improves communication performance of the system. The study of Hunt et al. shows that ZooKeeper is able to handle tens to hundreds of thousands of transactions per second for the target workloads, 2:1 to 100:1 read to write ratio[3], and keeps their linearizability with a lock mechanism at the same time. Therefore, it has been proved to have high throughput and low latency.

As the architecture of ZooKeeper is essentially a shared distributed hierarchical key-value store, the mechanism is fairly robust. Distributed processes contact with each other by obtain information from the shared namespace of registers, which is known as znodes[4]. Zookeeper nodes can be read from and write to by processes, the associated data of which can includes configuration, status information, location information etc.

One of its aims is to make large-scale distributed system achieve fault-tolerant control. This means that if some processes fails, others will be notified if they have set a watcher on znodes. The watch feature also helps to update group membership and start a leader election when old processes leave or new ones join.

III. Related works

In order to increase system's availability, some BitTorrent protocol extensions have been proposed and deployed. These approaches consider different mechanisms for peers to discover other peers including: multi-tracker, Distributed Hash Table (DHT) protocol and Peer Exchange (PEX) protocol.

The first approach is multi-tracker. To avoid overloading trackers and have backup trackers against failures, it allows two or more trackers to track one same torrent instead only one tracker. Every peer that participates in sharing a file can be tracked by one tracker and is a member of one swarm. Multiple swarms tracked by multiple trackers which are associated with one file can coexist in parallel. Multiple trackers improve availability, but the improvement largely comes from a single highly available tracker. The performance of small swarms is sensitive to fluctuations in peer participation. Measurements and analysis have shown that peers in small (less popular) swarms achieve lower throughput on average[5]. In [6], the authors studied the availability of multi-tracker observe the correlated failures of different trackers can reduce the potential improvement from multi-tracker. Besides, the use of multiple trackers can significantly reduce the connectivity of BitTorrent overlay.

It is obvious that limited tracker bandwidth will make an effect on upload/download rate of peers. In order to reduce this impact and accelerate building connection between peers, Andrew and Arvid[7] exploit a trackerless structure using Distributed Hash Table(DHT) protocol. In this protocol, each peer has a DHT node, which maintains contact information of closest nodes/peers in a routing table. In another word, a peer can play a role of tracker, and thus locating those peers related with its requesting file by itself, which simplifies the bootstrapping in the original mechanism. For determining the closeness, they use a XOR-based distance metric in Kademlia[8]. As this novel topology uses parallel, asynchronous queries to update routing table, nodes have enough information so that flexible route queries are guaranteed. When in a fault-prone situation, this system has provable great performance in consistency.

The PEX approach makes use of the communication among peers to share the contact information they have with each other periodically. Though multiple versions of PEX have been implemented, their main idea is that peers keep their neighbors informed about their current contact list. With its decentralized nature, PEX can help the swarm survive much longer in case of tracker failures, thus increasing the fault tolerance of the system. Unfortunately, using PEX does not eliminate the need for a tracker because the peer need to request the tracker to know at least one other peer. According to the experiments study in [9], PEX could improve the download performance - the average reduction of the download time was measured to be around 7%. As the peer needs to send messages containing contact lists to every other neighbor, a trade-off on the frequency of messages sent must be considered. [9] shows that over 80% of PEX messages have a freshness ratio greater than 0.5, but there exists a large degree of redundancy in PEX messages.

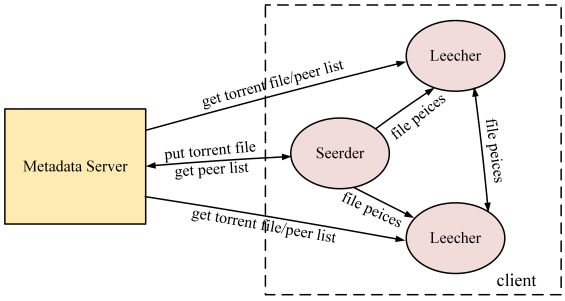

Fig. 1: System Architecture

IV. Design and implementation

The BitTorrent protocol generally requires the components including tracker, metainfo file containing information about the torrent, original seeder that has the whole content and end user downloaders. Figure 1 illustrates the architecture of our system. The system consists of two main components: metadata server and client.

In our design, we use Apache ZooKeeper to act as a metadata server which provides the information of peer lists of each swarm and metainfo of each file. The top level directories in ZooKeeper consist of /peer and /file nodes. The /peer node is used to hold the hostname and port number information of each peer. The /file nodes maintains the files and swarms. Each file is associated with a descendant node under /file, we store the torrent content in the file nodes data. All peers within one file swarm are registered under that specific file node.

A client can be started either as a seeder or as a leecher with a unique peer ID. If the user has a file to share, he can start the client as a seeder. The client will create a metafile and advertise the file to the ZooKeeper. A node with filename as nodes name is created under /file and the contents of the torrent information include filename, file size, pieces length will be stored in the node. Meanwhile, the peer is registered under both the /peer node and /file/filename node with the peer ID as the nodes name and hostname and port number as the nodes data. If the user wants to download a file, he needs to initiate the client as a leecher with the files name. The client will retrieve the metafile and peer lists associated with the file from the ZooKeeper. The peer is also added to the lists. Then the client can connect to those peers in the returning list to start exchange file pieces. Downloading or uploading multiple files simply involves running multiple client instances.

A. Connection state

A peer must maintain state information for each connection that it has with a remote peer. A peer usually play the roles of both downloader and uploader, Figure 2 shows two finite state machine for these two actions.

B. Deal with churn

The dynamics of peer participation, or churn, are an inherent property of Peer-to-Peer systems and critical for design. In

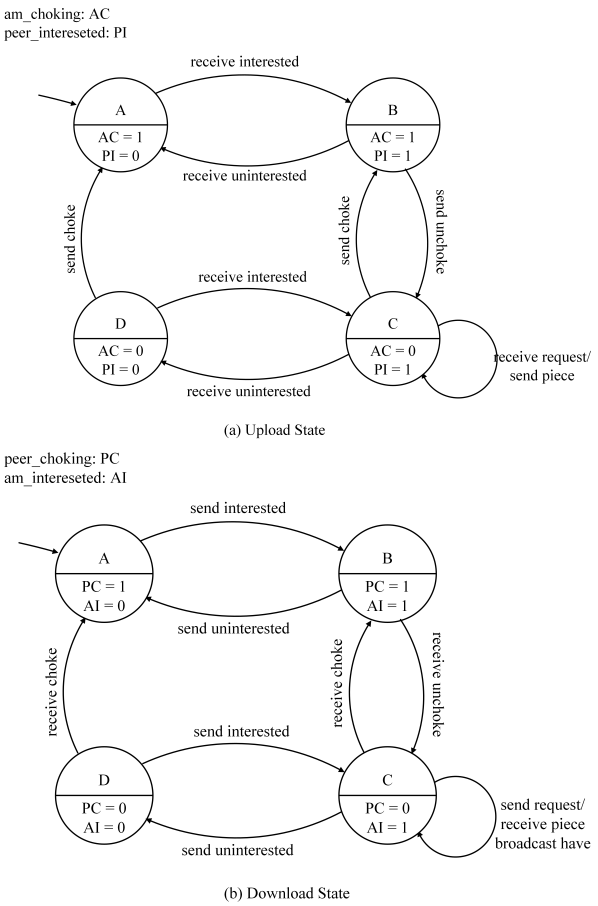

Fig. 2: Connection State

our system, each client spawns a listening thread to accept new connections from peers joining the swarm later. If receive handshake message successfully, current peer will add the new peer into its peer list. To deal with dropping offline peers, as every node is registered in the metadata server, we set a watch for each peer to monitor the file's node they involve in. If any node in the swarm changes, the peer will be informed and pull the updated peer list from the metadata server with dead peers removed.

C. Choking mechanism

Each peer always unchokes a fixed number of other peers which allows TCPs built-in congestion control to reliably saturate upload capacity[1]. This is decided periodically. In our implementation, a scheduled task is executed at a fixed rate to decide the unchoked sets. For simplifying, we calculate the download rates of all other peers that the current peer provide uploadings to, picks k peers who has the fastest rate. If a peer in the unchoked set is choked previously, then the current peer will send unchoke message to that peer.

D. Piece selection

Selecting pieces to download in a good order is very important for transfer efficiency. For example, if all the peers start to download the first piece from the first bit of the bitfield, it results in all peers having same bits of pieces, and can not exchange files with each other. In our design, we adopted the random first piece algorithm. The peer will check its own bitfield, pick the bit with zero randomly and request that index of piece from other peer.

V. Experiments and result

Our system is tested on ecelinux[1-3].uwaterloo.ca. The ZooKeeper service is on snorkel.uwaterloo.ca. The measurements focus on fault tolerance and evaluation of performance. We ran the system with three sizes-small, medium, large- of files of various types. There are 6 peers in this experiment, one is the seeder and the other 5 are leechers. In each test, a peer as seeder is started at the beginning, then different numbers of leechers ran simultaneously or non-simultaneously.

A. Fault tolerance

For fault tolerance, as the centralized tracker is already eliminated in our design, we introduced failures to peers. After killing several peers, new peers were started and connect to existing alive peers. When the new peers finish downloading, we examined the downloaded file by comparing with the original file to check the correctness.

NONE

B. Evaluation of performance

Performance refers to the transfer rate and download duration in terms of ranging swarm size and file size. files are split into fixed-size pieces which are all the same length of 256K except for possibly the last one which may be truncated. Unchoke interval for the choke task is 1s, unchoke slot is set to contain 4 peers. All the tests are run 10 times and all results are averages from 10 runs.

1) Performance as a Function of the Swarm Size(number of seeders): This experiment aims to show that BitTorrent performance improves with increasing seeder size. The data points are collected through running experiments with swarm sizes 1 to 5. At each run, one new peer joined a swarm until completing download. The swarm was initially created with one seeder then the rest 5 peers entered the swarm one by one after the preivous peer finishing downloading. Thus, peers that obtain the complete file can be seen as seeders. That is n seeders-one leecher.

Table 1 shows the measured download time of newly joined peer of 10 runs for the tests. The file size is medium. Figure 3 and 4 represent average download time and average download rates for newly-joined peers. The performance increases as seeder size changed from 1 to 5.

The results show that larger swarm performs better than smaller one. We see that the the performance of average download time and download rate for all 3 sizes files increases as the swarm size increases. It is obvious that the larger sizes the swarms are, the more peers that can provide file pieces to the new peer so that the download can speed up. We also notice that the decrease of download time is dramatical form

TABLE I: Download Time of Different Seeder Sizes

Test #	size 1	size 2	size 3	size 4	size 5
1	17874	12086	6817	6802	7904
2	14305	11394	5278	7288	6897
3	19312	9793	9747	7229	7901
4	25102	9825	8751	8734	6208
5	23135	10915	11600	6765	5856
6	24998	12018	7847	7138	6299
7	20161	9402	6862	8708	6102
8	25185	15401	9814	7876	7778
9	24632	10248	10264	7834	7323
10	21093	12885	8194	6737	6789
Average(ms)	21579.7	11396.7	8517.4	7511.1	6905.7

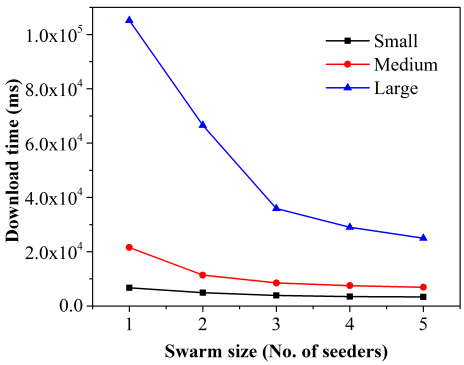

Fig. 3: Download Time vs. Seeder size

size 1 to size 3 and then tends to flat. One possible reason for this tendency is that with more providers, the leecher may send redundant requests and messages which affects the efficiency.

2) Performance as a Function of the Leecher Size(number of leechers): In this experiment, the swarm is initiated with one seeder and then we start different number of leecher peers in the same time making their download simultaneous. This is one seeder-n leecher. The measurement of performance is for 1, 3, and 5 newly leechers respectively. The download time

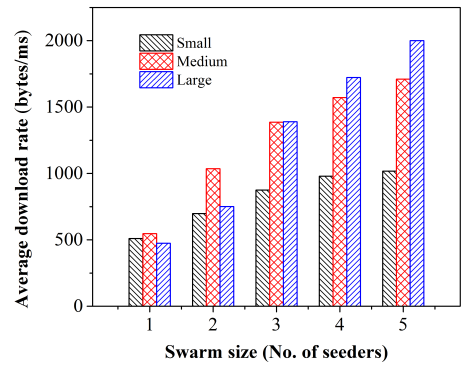

Fig. 4: Average Download Rate vs. Seeder size

TABLE II: Average Download Time of 1 Leecher

file size	medium(ms)	large(ms)
leecher_1	21579.7	105223.1

TABLE III: Average Download Time of 3 Leechers

file size	medium(ms)	large(ms)
leecher_1	10291.7	35817.2
leecher_2	1066.5	34736.4
leecher_3	10467.5	33409.6
Average	10275.2	34654.4

and average transfer rate is shown in table 2-5.

Comparing the results in these tables, we see that the average download time of every leecher peer decreases as the leecher number increases and the average rate increase accordingly. All leechers are started concurrently and they are downloading the same file, the improvement of their performance is due to the advantages of Bittorrent protocol that peers can share pieces with each other during downloading. Another reason is that we adopted the random piece selection algorithm so that the peers are very likely to obtain different pieces at the very beginning, thus they can upload to and download from others simultaneously. This balances the load from the only seeder to all peers.

3) Estimated Instantaneous Rate: The estimated instantaneous rate here is defined as a "windowing" average rate which is the bytes length per piece received divided by the time difference between current piece received and last piece received. We plot the instantaneous rate with time from the download established to download finish in Figure 5. We fitted the curve with polynomial fitting.

We can see that instantaneous rate slows down after about 81% completion. There are two reasons for this tendency. One is that the peer select piece it does not own randomly. In

TABLE IV: Average Download Time of 5 Leechers

file size	medium(ms)	large(ms)
leecher_1	9348.7	28307.3
leecher_2	10477.0	27830.9
leecher_3	10561.6	26701.2
leecher_4	9757.1	26273.5
leecher_5	9252.	27462.6
Average	9879.4	27315.1

TABLE V: Average Download Rate of Different Leecher Sizes

file size	size_1	size_3	size_5
medium(byte/ms)	547.17	1149.1	1195.2
large(byte/ms)	474.68	1441.3	1828.6

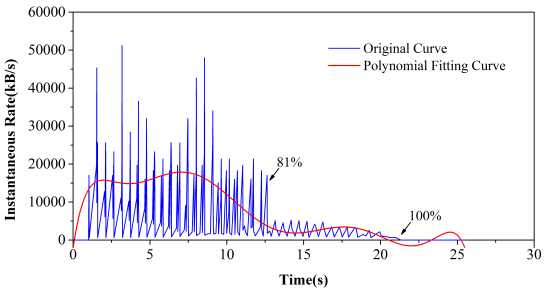

Fig. 5: Estimated Instantaneous Rate vs. Time

our implementation, we generate a random index number and check this index with local peers bitfield to decide whether the peer need to request. Thus, when most of the bitfield is already set, it requires more runs to generate an index of bitfield which is not set yet. The other reason is that the peer may send multiple requests for same piece to remote peers and receive redundant content. To speed this up, the peer can send requests for all of its missing pieces to all remote peers and enable cancellation every time a piece arrives.

VI. Conclusion and Future Works

Acknowledgment

The authors would like to thank...

References

[1] Cohen B. The BitTorrent protocol specification[J]. 2008.
[2] Bharambe A R, Herley C, Padmanabhan V N. Analyzing and improving BitTorrent performance[J]. Microsoft Research, Microsoft Corporation One Microsoft Way Redmond, WA, 2005, 98052: 2005-03.
[3] Hunt P, Konar M, Junqueira F P, et al. ZooKeeper: Wait-free Coordination for Internet-scale Systems[C]//USENIX Annual Technical Conference. 2010, 8: 9.
[4] Carlos D. Morales. Apache ZooKeeper Description[J]. 2008.
[5] D. Menasche, A. Rocha, B. Li, D. Towsley, and A. Venkataramani. *Content Availability and Bundling in Swarming Systems.* in Proc. ACM CoNEXT, Dec. 2009.
[6] G. Neglia, G. Reina, H. Zhang, D. Towsley, A. Venkataramani, and J. Danaher. *Availability in BitTorrent Systems.* in Proc. IEEE INFOCOM, May 2007.
[7] Loewenstern A, Norberg A. DHT protocol[J]. 2008.
[8] Maymounkov P, Mazieres D. Kademlia: A peer-to-peer information system based on the xor metric[C]//International Workshop on Peer-to-Peer Systems. Springer Berlin Heidelberg, 2002: 53-65.
[9] Wu, Di, Prithula Dhungel, Xiaojun Hei, Chao Zhang, and Keith W. Ross. *Understanding Peer Exchange in BitTorrent Systems.* in Proc. of IEEE Peer-to-Peer Computing (P2P), 2010.