Peer-to-Peer Technology for File Sharing

S. H. Kwok
California State University, Long Beach, USA

Y. M. Cheung
The Hong Kong University of Science and Technology, Hong Kong

K.Y. Chan
The Hong Kong University of Science and Technology, Hong Kong

INTRODUCTION

A recent survey revealed that 18 millions American Internet users, or approximately 14% of total American Internet population have peer-to-peer (P2P) file-sharing applications running on their computers (Rainie & Madden, 2004). Not surprisingly, P2P applications have become common tools for information sharing and distribution since the appearance of Napster (Napster, 2003) in 1999.

P2P systems are the distributed systems in which all nodes are equal in terms of functionality and able to directly communicate with each other without the coordination of a powerful server. Anonymity, scalability, fault resilience, decentralization and self-organization are the distinct characteristics of P2P computing (Milojicic et al., 2002) compared with the traditional client-server computing. P2P computing is believed to be capable of overcoming limitations of the computing environment placed by the client-server computing model. Milojicic et al. (2002), for example, suggested that P2P computing is capable of providing improved scalability by eliminating the limiting factor, the centralized server existing in the client-server computing.

In the past few years, P2P computing and its promised characteristics have caught the attention of researchers who have studied the existing P2P networks, and the advantages and disadvantage of P2P systems. Important findings include the excessive network traffic caused by flooding-based searching mechanism that must be tackled in order to fully utilize the improved scalability of P2P systems (Matei, Iamnitchi, & Foster, 2002; Portmann & Seneviratne, 2002). There were proposed efficient searching techniques targeted for both structured and unstructured P2P systems. Other research projects were conducted to study, and were intended to complement, the drawbacks brought by distinct characteristics of P2P systems. For example, the P2P users' free-riding behavior is generally attributed to the anonymity of such form of communication (Adar & Huberman, 2000). Recent research projects have shifted to a new line of investigation of P2P networks from the economic perspective and applications of P2P systems in workplaces (Kwok & Gao, 2004; Tiwana, 2003).

BACKGROUND

P2P systems, contrary to traditional client-server (C/S) systems, can be best described as distributed systems in which all nodes are considered to be equal in their capacity for sharing resources and information with each other. Presently, the P2P computing could be found in supporting two classes of applications, highly parallel computing and file sharing.

In highly parallel computing, peers share processing cycles. Parallel applications run on available nodes (peers) such that these computing-intensive applications are distributed to idle peers, and are free from the dependency on a sophisticated but costly central server. One famous example is SETI@Home that analyzes gigantic amount of scientific data by utilizing the extra processing cycles of millions of computers in home and office environments (Anderson, Cobb, Korpela, Lebofsky, & Werthimer, 2002; SETI, 2003). Prior researchers devoted their efforts to investigate the similarities and differences between P2P and grid computing models, followed by the discussions on the synergy and convergence between the two computing models (Foster & Iamnitchi, 2003; Talia & Trunflo, 2003). One useful study was by Talia and Trunflo (2003) who suggested implementing grid tools and services with the P2P approach.

P2P systems are well-known to the public as the popular systems for file sharing. P2P communities make use of the file-sharing P2P systems, such as Gnutella (Gnutella, 2003) and KaZaA (KaZaA, 2003), to share files, mostly multi-media files, with other users. Previous research in this area has focused on the issues concerning searching capability and efficiency, free-riding and reputation management systems.

Before looking at the academic discussions concerning P2P systems, we first review the existing file sharing systems that are widely used by the P2P communities. These file sharing P2P systems could be classified by topologies and the most common topologies are decentralized, centralized and hybrid (Minar, 2003).

Decentralized

In the decentralized network, there is no central directory server. Every peer directly connects to one or more peer(s). All requests are broadcasted to all of connecting neighbors, by flooding-based broadcasting mechanism. Gnutella is an example.

Centralized

Peers first approach the central directory server to obtain meta-information, such as the identity and location of the peer on which some information is stored, or to verify security credentials. Direct communication with a peer is conducted afterwards. A famous example is Napster.

Hybrid (Centralized + Decentralized)

A dynamic supernode acts as a centralized directory server, to which a certain number of peers connect, reflecting only a partial view of the whole network. While supernodes maintain Gnutella-like connections among themselves, a supernode is responsible for handling the requests by its connecting peers. KaZaA is an example.

To better illustrate each of the above topologies, we select their representative applications and analyze them in terms of their architectures and characteristics as shown in Table 1.

ISSUES, PROBLEMS, AND SOLUTIONS

In this section we discuss some important P2P related issues, the identified problems and the corresponding solutions that have been widely researched in the academic field. According to their studies, there are three major issues or problems relating to P2P computing, namely (1) non-compliance; (2) searching capability and efficiency; and (3) free riding and reputation management systems.

Non-Compliance

In a P2P network, compliance is the ability for a peer to effectively communicate with the directory server (if any) and other peers, without any problem. The most important consideration is whether two peers with different or even same P2P protocols are able to communicate with each another effectively. Non-compliance implementations are problematic both from the aspects of P2P users and the whole P2P community.

Possible results of non-compliance between communicating peers include (1) distorting the messages, making them different from the original; (2) corrupting the messages as they are being forwarded; (3) routing the messages improperly as they are being forwarded (Truelove, 2001).

One of the possible ways to route and handle the messages properly is by creating and maintaining a routing table, at a substantial cost of increased traffic and lost responses (Truelove, 2001). Another proposed solution is a framework to integrate various P2P file sharing protocols by using a P2P gateway for the sake of interoperability among all P2P users (Lui & Kwok, 2002). In this model, the P2P gateway converts the passing messages to suitable formats that are understandable by the recipients of messages.

Searching Efficiency and Flooding-Based Broadcasting

To efficiently locate the desired files (if any) is an increasingly important issue of P2P systems as the P2P community has grown significantly in size that it probably consists of users from all around the world. A recent study reported that, as of 21 April 2004, there were currently over 3.2 millions users in the FastTrack (KaZaA) network and 0.23 millions users in the Gnutella network (Slyck, 2004).

P2P users primarily locate their desired files by keyword searching, .i.e., users enter keywords consisted of regular expressions in an attempt to locate desired files that may or may not exist in the P2P networks. Advanced searching which XML (Extensible Markup Language) (XML, 2003) was introduced to improve the searching (Limewire, 2004).

From our review, it is obvious that prior researchers have mainly paid their attention to the searching efficiency and excessive network traffic posed by searching in the Gnutella-like decentralized network which is believed to best illustrate the P2P philosophy: a network that was only formed by equal and autonomous peers. One of the hot discussion topics is concerning the excessive network traffic, posed by the flooding-based broadcasting mechanism that is implemented to facilitate searching. The findings of public Gnutella network and network traffic were reported (Matei, Iamnitchi, & Foster, 2002; Portmann & Seneviratne, 2002). While Matei et al. (2002) suggested replacing the existing query flooding mechanism with smarter routing mechanisms, other researchers proposed adopting probabilistic searching protocols to minimize the network traffic by avoiding unconditional broadcasting, at the cost of search coverage and end-user latency (Menasce, 2003; Portmann &