European Journal of Scientific Research
ISSN 1450-216X Vol.34 No.4 (2009), pp.535-541

http://www.eurojournals.com/ejsr.htm

An Optimized Method for Analyzing the Peer to Peer Traffic

M Sadish Sendil
Research Scholar, Assistant Professor
Department of Computer Science and Engineering
SNS College of Technology, Sathy Main Road
Coimbatore-641035, Tamil Nadu, India
E-mail: sadishsendil@yahoo.com
Tel: +91 – 422 – 2666264; Mobile: +91 – 9994688771

N. Nagarajan
Principal and Research Supervisor
Coimbatore Institute of Engineering and Information
Technology, Coimbatore – 641 109

Abstract

The global Internet has emerged to become an integral part of everyday life. In the recent years, peer to peer networks have rapidly developed in the distributed and decentralized world of internet. Peer-to- Peer (P2P) is the logical antithesis of the Client-Server (CS) paradigm that has been the ostensible predominant paradigm for IP-based networks since their inception. Current research indicates that P2P applications are responsible for a substantial part of the Internet traffic. The number of users embracing new P2P technology is also increasing fast. It is therefore important to understand the impact of the new P2P services on the existing Internet infrastructure and on legacy applications. The majority of the unidentified traffic originates from peer-to-peer (P2P) applications like Napster, Gnuttella, etc. Identification of P2P traffic seem to fail because their existence by using arbitrary ports. The proposed scheme concentrates on the factors and characteristics of P2P communications with payload issues on P2P application based on network traffic collection. The method used here is based on a set of methods derived from the robust properties of P2P traffic. The system demonstrates the method with current traffic data obtained from Internet Service Providers. It has been found that the flow sizes Vs holding times, behavior of P2P users Vs total active users are also analyzed and results of a heavy-tail analysis are described. Finally, the system discusses the popularity distribution properties of P2P applications. The results shows that the unique properties of P2P application traffic seem to fade away during aggregation and characteristics of the traffic will be similar to that of other non-P2P traffic aggregation.

Keywords: P2P, CS, Napster, Gnuttella, ISP.

1. Introduction

The traffic generated by these P2P applications consumes the biggest portion of bandwidth in campus networks, overtaking the traffic share of the World Wide Web. A common feature in all of these P2P applications is that they are built on the P2P system design where instead of using the server and client

concept of the web each peer can function both as a server and a client to the other nodes of the network. This principle involves the adapting nature of P2P systems as individual peers join or leave the network. Another common feature of these P2P systems is that they are mainly used for multimedia file sharing (movies, music files, etc.), which frequently contain very large files (megabytes, gigabytes) in contrast to the typical small size of web pages (kilobytes).

In P2P environments, systems are no longer distinguished by thin clients and thick servers. i. e., every node (peer in P2P terminology) has, a priori, an equal status. This means that a peer offers services or resources to the community, but at the same time, it can consume services/resources from others in the system. An important property of P2P systems is the lack of a central administration. These properties make P2P file sharing systems most popular. But the P2P development is moving from simple file sharing to large scale decentralized and reliable systems.

A lot of new applications, e. g., reliable sensor networks (distributed data), ambient intelligence (distributed knowledge), and Ubiquitous Computing (distributed and highly interacting mobile devices) will benefit from this development. It is likely that evolving P2P systems will technologically be based on service oriented computing. A number of studies have been published in the field of P2P networking. Different P2P systems like Napster, Gnutella, KaZaA, and the traffic characterization and analysis of P2P traffic providing some interesting results of resource characteristics, user behavior, and network performance.

Further approaches propose structured P2P systems using Distributed Hash Table (DHT) with several implementations like Pastry, Tapestry, CAN, Chord. The P2P traffic characteristics are not fully explored today and there is a tendency that they will be even more difficult to analyze. In contrast to the first generation P2P systems the recent popular P2P applications disguise their generated traffic resulting in the problematic issue of traffic identification.

2. Optimized method for Identifying the P2P Traffic

The accurate P2P traffic identification is indispensable in traffic blocking, controlling, measurement and analysis. The problem is that P2P communications are continuously changing, from TCP layers using well known ports in some first versions to both TCP/UDP with arbitrary and/or jumping ports nowadays. A robust and accurate P2P traffic identification is vital for network operators and researchers but today there is a lack of published results on this field and this is the main motivation for the work presented in this paper.

Based on the literature survey, the following key factors are considered for identifying the peer to peer traffic which should be an optimized one. The first method is based on the fact that many P2P protocols, e.g. eDonkey, Gnutella, Fasttrack, etc., use both TCP and UDP transport layers for communication. Reasonably the unreliable UDP is often used for control messaging, queries, and responses while data transmission relies on TCP. However, the large volume of UDP traffic observed in the measurement data indicates that UDP could also be used for data transfer. Thus by identifying those IP pairs which participate in concurrent TCP and UDP connections the system can state that the traffic between these IP pairs is almost surely P2P.

The second method tries to separate web and P2P traffic from flows using HTTP/SHTTP ports, i.e. 80, 8080, 443 ... The typical difference between P2P and web communication of two hosts can be observed. In general, web servers use multiple parallel connections to hosts in order to transfer web pages text and images (also music, video contents in some cases). In contrast, data transmission between peers consists of one or more consecutive connections, i.e. only a single connection can be active at a time.

In the third method, P2P traffic is selected using default ports of P2P applications. P2P software often defines default ports for communication. It is true that in most cases peer users can change it to any arbitrary port (but it is not frequent since peer-to-peering is usually not prohibited for home users) or port can be dynamically chosen automatically or when firewall or port-blocking is observed.

This step cannot detect all P2P connections, but once the traffic is collected the system can be almost sure that it is from those concerned P2P systems. A table of well-known ports used by some popular P2P applications is collected for this step (see Table 3 for details). Flows containing these values in source_port or dest_port are all marked P2P.

Considering the fourth method, in normal TCP/UDP operation, at least one of the two ports is selected arbitrarily. It is not likely that flows with similar flow identities (source_IP, dest_IP, source_port, dest_port, prot_byte, TOS) exist in relatively short measurements. This happens, however, in the case of P2P connections; if both source and destination peers dedicate a fixed port for data transfer. File download of a file is often executed in several smaller chunks. Therefore multiple flows with the same flow identities can be generated by P2P software. This is the basis of this method: those identical flows are from P2P applications if at least two of each is found.

The last method is based on the fact that objects of P2P download often having large sizes from several MB in case of music files or smaller applications to hundreds of MB in case of video files and larger software packages. In addition, peer users are patient. P2P downloads can last some ten minutes or hours. By this method those flows are considered P2P flows which have flow size larger than 2 MB or flow length is longer than 12 minutes.

3. Implementation

3.1. ISP Network Backbone

From the Internet Service Providers, the following data can be collected for analyzing the traffic through the optimized methods.

- Number of connections available
- User's current status of operation
- Load capacity of the ISP backbone
- Connecting nodes to the network backbone

3.2. Traffic Measurements

In the chosen network segment, traffic of Asymmetric Digital Subscriber Line (ADSL) subscribers is multiplexed in some Digital Subscriber Line Access Multiplexers (DSLAM) before entering the ATM access network. Placed at the border of the access Network and the core network are some routers. Net flow measurements are carried out at two of these routers. Net flow collects all inbound and outbound flow information and exports the logs periodically. Some packet-level information was also recorded, including packet arrival times and packet sizes. The obtained data traces are the aggregate incoming traffic of more than 100 ADSL subscribers.

Table1: Summary of collected data sets

Data sets	Time of measurement	Number of flows	Total traffic (GB)
Call records 1 Inbound	15th July 2007	11 423 510	457.84
Call records 1 Outbound	15th July 2007	12 373 446	93.95
Call records 2 Inbound	9th Jan 2009	10 234 100	125.54

3.3. Traffic Characteristic Results

The analysis framework focuses on the fundamental differences between the P2P traffic and other Internet traffic. The comparison is done regarding several aspects of the traffic characterization i.e.

- Traffic intensities
- Traffic volume

The volume of P2P traffic, which is about 60-80% of the total traffic, exceeds by far the traffic volume of the non-P2P applications. This observation is especially true for outbound aggregate traffic.

The reason is that home users do not generate too much upload traffic, except for those users who use P2P applications. As a consequence the ratio of P2P traffic in the outbound direction is higher than in the inbound direction.

Figure1: The total and non Peer-to-Peer traffic intensity of the inbound Data set

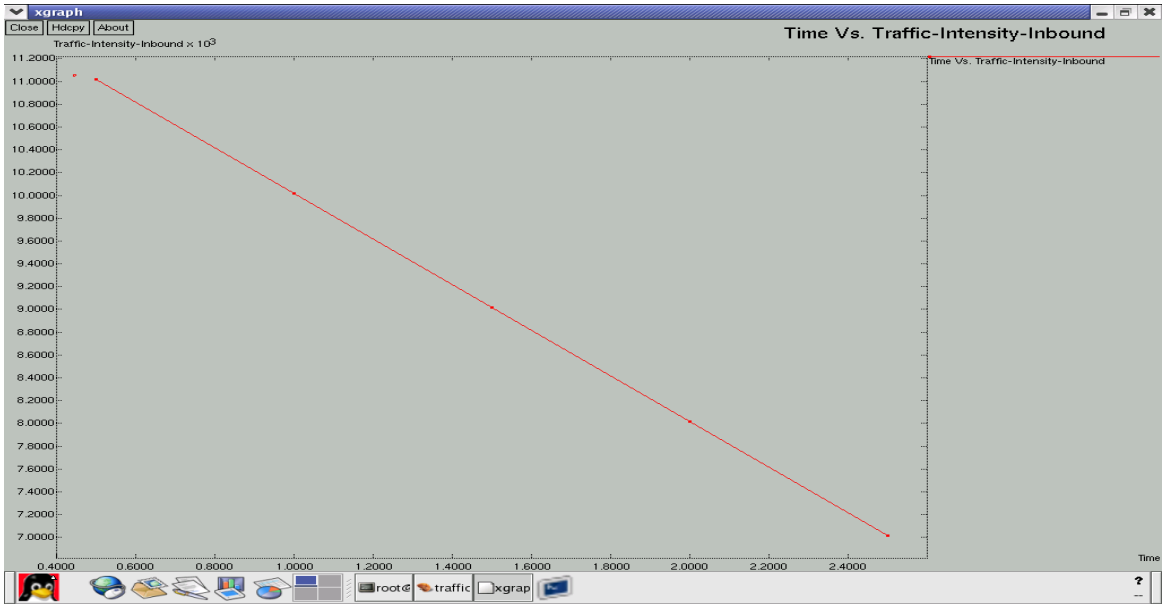
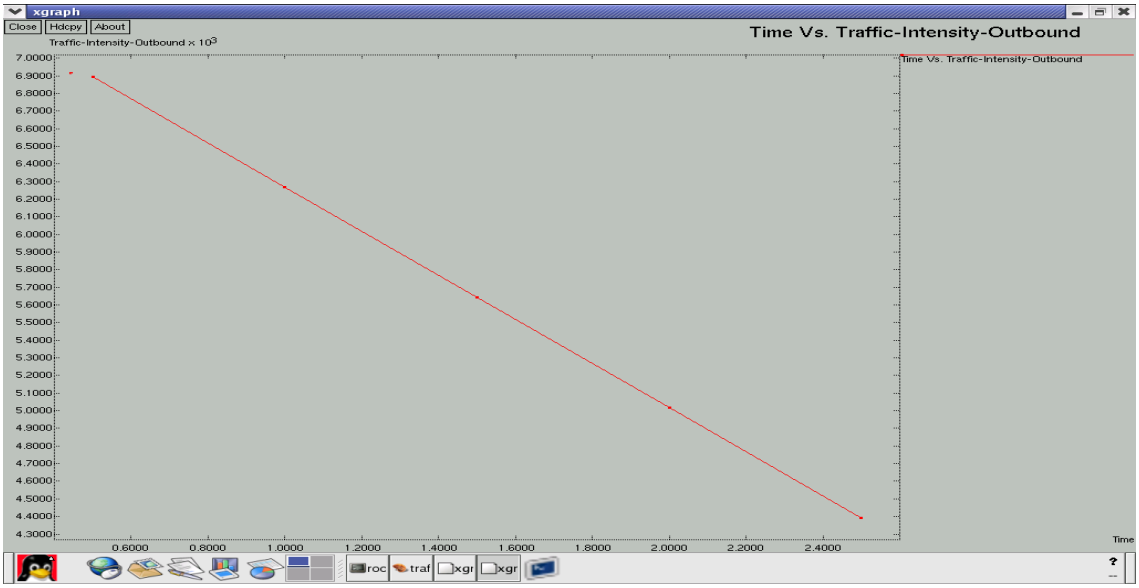


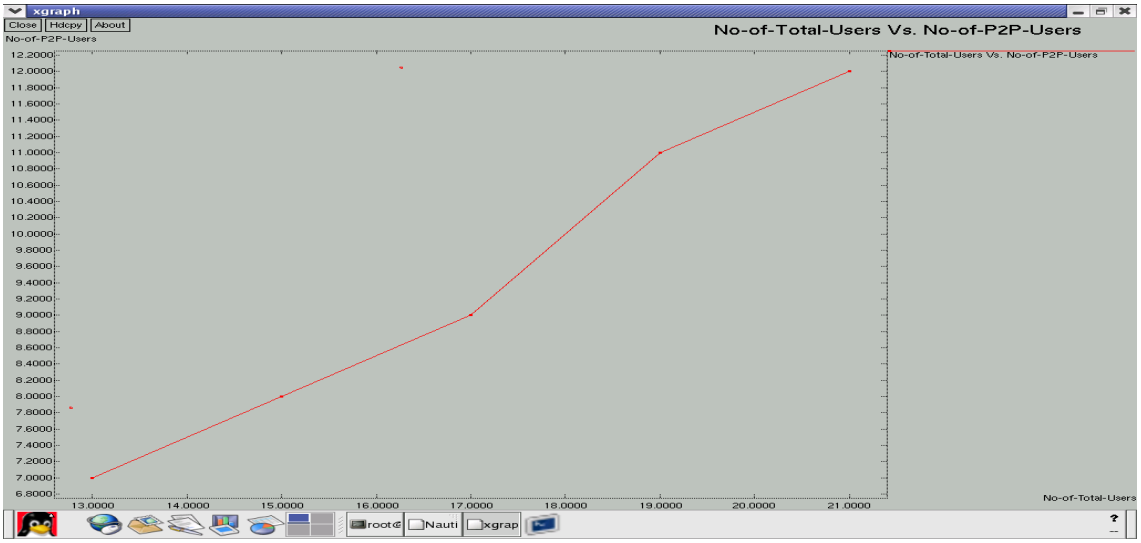
Figure 2: The total and non Peer-to-Peer traffic intensity of the outbound Data set



3.3.1. Number of P2P and total active users

In the measurement environment, Internet subscribes do not have fixed IP addresses. Each time a user connects to the Internet, a dynamic address is given to the user. Therefore it is impossible to determine exactly which data flow belongs to which user. However, less error is expected when the system choose to associate an individual IP address to a user. Since the ADSL contracts at the present Internet provider do not limit the time of connections, the average connection time is relatively long. The system assumes that during the measurements, which lasted at most 24h, only a minimum number of IP address wanderings occurred.

Figure 3: The relation between the number of P2P users and total active users are counted.



3.3.2. Number of active P2P user and bandwidth consumed

The relation between the number of active (P2P) users and the occupied bandwidth is also investigated. It is shown that a linear connection can be observed in both cases (P2P and non-P2P traffic). However, the variance of data around the assumed linear function is much higher than in the previous case (which is presented in Fig. 4). In addition, variation is higher and the slope of the line is much lower for non-P2P traffic. This means, P2P users (e.g. users, who use P2P applications as well) generate much more traffic in average than those users, who use only non-P2P applications.

3.3.3. Flow size and Holding times

The next comparison is about the properties of data transferring: flow size and flow holding time. The proposed system finds no significant divergence in these characteristics. In both cases the plots, disregarding flow sizes smaller than 0.1B, nearly follow a straight line in the log-log scale. This indicates a possible heavy tailed (Pareto) model for the flow size for both P2P (with shape parameter a=-0.3) and non-P2P flows (a=-0.25) and also for the overall traffic. (The assumptions of Pareto distribution were verified by several heavy-tailed tests: De Haan's moment method, Hill estimator, and QQ-plot.) The number of P2P flows which are larger than about 100 kB is somewhat higher than the number of non-P2P ones, which is also reasonable, but the difference is not significant.

- Traffic flow and throughput
- Traffic flow and P2P user demand variance
- Load balance of P2P active users

Figure 4: The data transferring flows and holding times of data is calculated.

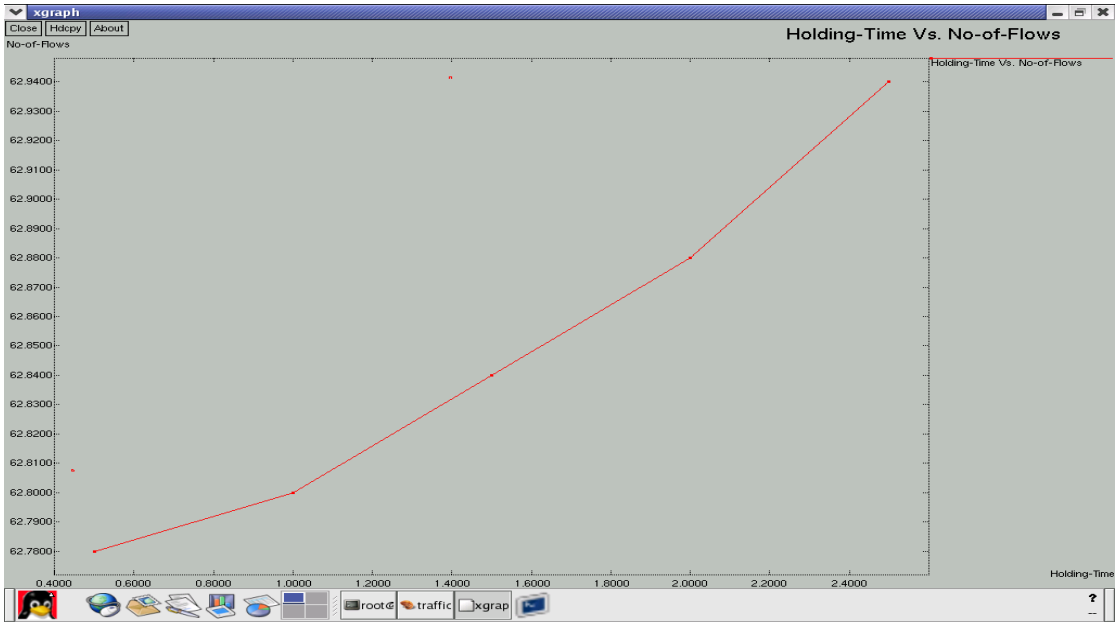
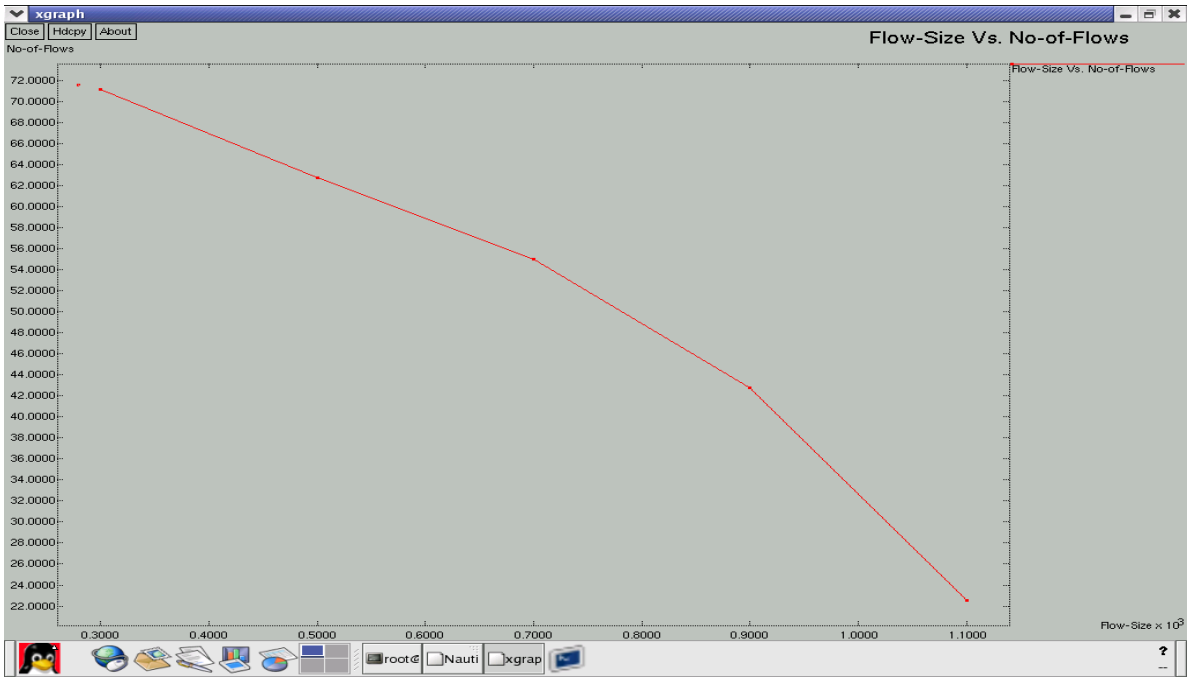


Figure 5: Comparison between the numbers of data flows from that flow size is calculated.



4. Conclusion and Future Work

In this paper the scheme first presented a novel P2P traffic identification method. The method collects a set of rules derived from the general behavior of P2P traffic. The proposed method does not use any payload information so it is easy to implement and use when payload cannot be evaluated because of legal or privacy obstacles or cannot be measured due to technical or financial problems. The validation results show that the proposed algorithm is able to identify the P2P traffic very efficiently. The method was used to identify P2P traffic in current measurement data. Traffic analysis study focusing on the most important characteristics like the behavior of active users, the ratio between the P2P users and the

total number of users, flow size and holding time distributions and the popularity distribution. The system showed that packet-level statistics of P2P and non-P2P data flows are basically similar.

The experimentation investigates the relationship between packet sizes and applications resulting in a list of typical applications belonging to various packet sizes. The analysis of the number of active users and total users revealed an almost linear relation. It suggests a very interesting and important result from a traffic dimensioning point of view: the ratio of active users and total users is almost constant. The major conclusions is that in spite of the different characteristics of individual P2P traffic the main characteristics of P2P aggregation do not differ significantly from the characteristics of other Internet traffic aggregation.

References

[1] Gummadi K.P, et.al., Oct, 2003, "Measurement, Modeling, and Analysis of a Peer-to-Peer File-Sharing Workload", in Proc. 19th ACM Symposium on Operating Systems Principles (SOSP-19), Bolton Landing, NY.

[2] Karagiannis. T, et.al., 2003, "File-Sharing in The Internet: A Characterization of P2P Traffic in The Backbone", Technical Report, UC Riverside,.

[3] Karagiannis. T, et.al., Oct, 2004, "Transport Layer Identification of P2P Traffic", in Proc. 4th ACM SIGCOMM Conf. on Internet Measurement, Taormina, Sicily, Italy.

[4] Kim. M, et.al., 2003, "Towards Peer-to-Peer Traffic Analysis Using Flows", DSOM: 55-67.

[5] Sen. S, et.al., 2004, "Analyzing Peer-to-Peer Traffic Across Large Networks", IEEE/ACM Transactions on Networking, 12(2):219-232.

[6] Sen. S, et.al., 2004, "Accurate, Scalable In- Network Identification of P2P Traffic Using Application Signatures", in Proc. 13th Int. Conf. on World Wide Web, NY, USA.

[7] Ohzahata. S, et,al., 2005, "A Traffic Identification Method and Evaluations for a Pure P2P Application", Lecture Notes in Computer Science, p55 Vol. 3431.

[8] Marcell Perényi, et.al., 2006, “Identification and Analysis of Peer-to-Peer Traffic”, Journal of Communications, Vol. 1, No. 7, November/December.

[9] Yunfei ZHANG, et.al., 2006, “Recent Advances in Research on P2P Networks”, Proceedings of Parallel and Distributed Computing, Applications and Technologies (PDCAT'06), 0-7695-2736-1/06.

[10] Thomas Karagiannis, et.al., 2004, “Is P2P dying or just hiding?”, Proceedings of IEEE Communication Society (Globecom), 0-7803-8794-5.

[11] Naimul Basher, et.al., 2008, “A Comparative Analysis of Web and Peer-to-Peer Traffic”, Proceedings of International World Wide Web Conference Committee (IW3C2), ACM 978-1-60558-085-2.

[12] Li Jun, et.al., 2007, “Active P2P Traffic Identification Technique”, Proceedings of IEEE Computational Intelligence and Security, 0-7695-3072-9/07.