Guest Editors' Introduction: Special Section on Peer-to-Peer-Based Data Management

Beng Chin Ooi, *Member*, *IEEE*, and Kian-Lee Tan, *Member*, *IEEE Computer Society*

✦

P EER-TO-PEER (P2P) computing has attracted much attention from both the academic community and industry. This is fueled by the successful deployment and adoption of many domain specific P2P systems. For example, Freenet and Gnutella enable users to share any digital files (e.g., music files, video, images), Napster allows sharing of (mp3) music files, ICQ facilitates exchanges of personal messages, SETI@home makes computing cycles of participants available, and LOCKSS pools storage resources to archive document collections.

In P2P systems, autonomous peers (computers) are treated as equals, i.e., perform the same functions. They can join and leave the system at any time. These peers pool together their resources (data, storage, computing cycles) to enable new capabilities greater than the sum of the parts. Data can be exchanged between peers directly and underutilized resources can be tapped. The potential of such a highly distributed and decentralized system is tremendous.

Interestingly, existing P2P systems lack data management capabilities that are typically found in DBMS. Although research in distributed (and heterogenous) databases has been pursued for many years, the database community has not been as aggressive in enhancing P2P systems with data management capabilities. We would add that the current P2P paradigm offers challenges beyond what has been previously done in the distributed database context. To list a few, the system may scale to over thousands or tens of thousands of peers which existing techniques cannot adequately handle, the dynamism of the system raises issues of information quality (e.g., completeness, consistency) that have not been previously considered, and the trustworthiness of the participating peers poses security threats not seen before. This special section aims to bring together current research activities that address some of these problems. The section contains six papers covering topics on data integration, search, consistency, trust, and identity. We hope this section will whet the appetite of our community to pursue this exciting field further.

In a peer-based data management system, it is practically impossible to construct a global schema that mediates semantic differences of shared data across a large number of autonomous peers. The first paper, "The Piazza Peer Data Management System" by Alon Y. Halevy, Zazhary G. Ives, Jayant Madhavan, Peter Mork, Dan Suciu, and Igor Tatarinov, proposes a solution to facilitate ad hoc, decentralized sharing and administration of data, and defining of semantic relationships. Every peer can contribute new data and relate the data to existing concepts and schemas and define new schemas for other peers to use as frame of reference for their queries. The paper also discusses query answering and optimization algorithms.

Replication and caching are very effective mechanisms that can bring the data/results closer to the users to improve performance. However, these mechanisms also introduce new challenges: Data that are replicated or cached have to be coherent with the source, updates to the data must be carefully disseminated from sources to their cached/replicated copies in other peers to minimize communication and computation overhead, and, in a P2P environment, the network should be resilient to failures so that data coherency is not completely lost even in the midst of failures. The second paper, "Resilient and Coherence Preserving Dissemination of Dynamic Data Using Cooperating Peers" by Shetal Shah, Krithi Ramamritham, and Prashant Shenoy, presents a dynamic push-based data dissemination architecture that ensures data coherency, resiliency, and efficiency.

The third paper, "Efficient Semantic-Based Content Search in P2P" by Heng Tao Shen, Yanfeng Shu, and Bei Yu, presents a three-tier framework to support semantic-based retrieval of documents. The framework summarizes the information content at different granularities: individual document level, peer level, where all documents within a peer are summarized, and superpeer level, where all summaries of peers managed by a superpeer are further combined and summarized. With the framework, queries can be routed to peers with similar content quickly. Corresponding to each tier of the framework is an index structure to facilitate speedy retrieval.

A critical issue in P2P systems is trust management—without a good solution to this problem, P2P systems are not likely to be deployed for serious applications. Essentially, peers need to manage the risk of communicating or cooperating with each other without prior experience and knowledge about each other. The next two papers address this important issue. The fourth paper, "Trust-*X*: A Peer-to-Peer Framework for Trust Establishment" by Elisa Bertino, Elena Ferrari, and Anna Cinzia Squicciarini, employs the notion of trust negotiation to establish trust between peers. In trust negotiation, the parties exchange digital credentials which may need to be protected. This

● *The authors are with the Department of Computer Science, National University of Singapore, 3 Science Drive 2, Singapore 117543.*
 *E-mail: {ooibc, tankl}@comp.nus.edu.sg.*

paper proposes Trust-X which embraces all aspects of negotiation, from the specification of the profile and policies of the peers involved to the selection of the best strategy by each peer to succeed in the negotiation. The fifth paper, "PeerTrust: Supporting Reputation-Based Trust for Peer-to-Peer Electronic Communities" by Li Xiong and Ling Liu, adopts a reputation-based model to manage trust. The proposed approach, PeerTrust, determines the trustworthiness of a peer based on feedback the peer receives through its transactions with other peers, the total number of transactions the peer performs, and the credibility of the feedback sources.

Finally, the sixth paper, "Efficient, Self-Contained Handling of Identity in Peer-to-Peer Systems" by Karl Aberer, Anwitaman Datta, and Manfred Hauswirth, addresses the issue of identification in P2P systems. It presents a decentralized, self-maintaining, lightweight, and secure directory service based on secure identification. The scheme separates identity from the network properties to realize the concept of logical independence in overlay networks. It also provides a general approach to identify entities and bind arbitrary information to them. The scheme has been applied to the P-Grid P2P system to remedy the problem of dynamic IP addresses.

In closing, we would like to thank the authors for their high-quality contributions to this special section as well as the referees for their careful assessment of the papers and for providing very helpful comments and suggestions. Special thanks are due to Philip Yu for offering the field the opportunity to have this special section published. We hope the readers will appreciate and enjoy it.

<div style="text-align:right">

Beng Chin Ooi
Kian-Lee Tan
Guest Editors

</div>
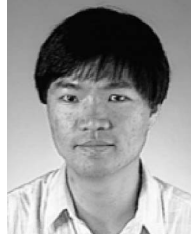
Beng Chin Ooi received the BSc (First Class Honors) and PhD degrees from Monash University, Australia, in 1985 and 1989, respectively. He is currently a professor of computer science at the School of Computing, National University of Singapore. His current research interests include database performance issues, indexing techniques, XML, spatial databases, and P2P/grid computing. He has published more than 80 conference/journal papers and served as a PC member for a number of international conferences (including SIGMOD, VLDB, ICDE, EDBT, and DASFAA). He is an editor of *GeoInformatica*, the *Journal of GIS*, *ACM SIGMOD Disc*, *VLDB Journal*, and the *IEEE Transactions on Knowledge and Data Engineering*. He is a member of the ACM and the IEEE.

Kian-Lee Tan received the BSc (Hons) and PhD degrees in computer science from the National University of Singapore in 1989 and 1994, respectively. He is currently an associate professor in the Department of Computer Science, National University of Singapore. His major research interests include query processing and optimization, database security, and database performance. He has published more than 100 conference/journal papers in international conferences and journals. He has also coauthored three books. Dr. Tan is a member of the ACM and the IEEE Computer Society.