SURVEY REPORT: A SHIFT FROM TRADITIONAL TOWARD PEER-TO-PEER (P2P) INFORMATION INTEGRATION

I Wayan Simri Wicaksana
Pusat StudiTeknologi Sistem Informasi, Universitas Gunadarma
Jl. Margonda Raya No. 100 Depok
iwayan@staff.gunadarma.ac.id

ABSTRACT

Peer-to-Peer (P2P) networking is not a new technology, the P2P has been introduced in the end of 70's, however the real implementation can be done in last decade synchronize with Internet progress. This literature study will evaluate from some research papers about the progress of P2P in many aspects. The main emphasize is in information searching and interoperability.
Basic idea of P2P arrived from social life which brings to the model of searching information in computer network model. From the basic P2P model that called pure P2P until the combination of some models to find better solution for special condition is discussed in this paper. Characteristic and function of P2P in many environments bring as well as the background knowledge in this area. This paper can bring contribution in selecting appropriate model of P2P.

Keywords: *topology, networking*

INTRODUCTION

History of P2P has been started since around 1970. USENET (1979) and FidoNet (1984) are two examples of completely decentralized networks of peers. Sun added object to Java language to speed up the development of peer-to-peer application in the late 1990, this effort is now being continued with the JXTA project. Microsoft also introduces dot-NET as one of P2P platform. Napster (2001) is music file sharing, this implementation is a trigger popularity of P2P. The P2P systems and application have attracted a great deal of attention from computer science research.

What is new and what is not new in P2P? Refer to (Milojicik, etc, 2002), give the comparison such as Table 1.

Computer terminology is often making confuse in academic, industry and users environment. P2P terminology drives misperception as well. To reduce the misperception, Milojicic et al. (2002) has collected some definitions from some experts as follow:

- The Intel P2P working group gives definition: "the sharing of computer resources and services by direct exchange between systems"
- Ross Lee Graham state P2P definition trough three key requirements: a). they have an operational computer of server quality; b) they have an addressing system independent of DNS; and c) the are able to cope with variable connectivity
- Clay Shirky of O'Reilly and Associate say "P2P is a class of applications that takes advantage of resources – storage, cycles, content, human presence – available at the edges of the Internet. Because accessing these decentralized resources means operating in an environment of unstable connectivity and unpredictable IP addresses, P2P nodes must operate outside the DNS system and have significant or total autonomy from central servers".
- Kindberg's definition is "P2P systems as those with independent lifetimes".

Table 1
What is New and what is Not New in P2P

Perspective	What is New		What is Not New	
Historical/Evolutionary (Computing)	Computing on the edge of the Internet	Scalability, availability, security, connectivity	Distributed	Concept, applications, distribution
Cultural/Sociological (Content	Direct sharing (privacy,		Decentralization	
Communciation / Collaboration	Dealing with disconnection		Ad-hoc NW, disconnected	
Architectural	Cost of ownership		P2P concept and application	
Algorithms/ Programming model	Particular algorithms		Distributed state algorithm in general	

(Milojicik, et al, 2002)

Our view from above definition P2P has special characters sharing, direct exchange, self-organized and independent, node can be server or client, independent addressing and connecting system. Therefore, to decide it is P2P or not, we have to look at the characters of system.

THEORETICAL BACKGROUND

After looking at the P2P systems, there are some specific characteristics that make the P2P system different form other systems. Some papers said the characteristics are goal as well for the P2P system. The main characteristics of P2P system are self-organizing, node has freedom to organize himself into network; symmetric communication, node are equal both request or offer services. So, nodes need have an operational computer of server quality, since each node can be act as a client and/or a server; and decentralized, no global directory or central control to every node.

From above main characteristic, there are other derivative characteristics, such as autonomy, cost of ownership, anonymity/privacy, scalability, ad-hoc connectivity, addressing system is independent of the DNS, and joint leave P2P nodes are at any time and unpredictable. Autonomy character derives from self-organizing. A peer can consider himself to decide P2P model, joint to which P2P network and so on. Cost of ownership characteristic come from understanding of ownership, shared ownership reduce cost of owning, and cost of maintaining. P2P systems implement this understanding. Anonymity/privacy, some forms from the authors of the free heaven, i.e. the location of a file is not known by its retriever, files move freely among systems, author/creator/publisher/ reader cannot be identified, servers do not know what documents they are storing, and servers cannot tell what document it is using to respond to a user's query. Decentralization gives better scalability, because limitation of scalability depend on power of centralized operation. Ideally connectivity can be variety, so peer can join and leave based on physical location or interest. For anonymity should be considered with digital right and intellectual property issues.

Refer to above characteristics of P2P, general pro and contra can be summarized as seen Table 2.

Some requirements are needed to run P2P systems based on the characteristics. The requirements are standard communication protocols are required, information exchange should be secure, information network should support policy-based authorization, information network should facilitate more effective search, information network should be easy to use and set up, information network should scale, and information network should be ubiquitous.

P2P system is proposed to answer some problem at many areas. P2P systems can be implemented for Community Web Network, can share data, information, sources among community group which has specific interest, such as e-business, p2p has possibility to add new capability at distributing and sharing

information, gaming, refer to messenger application, game has the same model, but need of bandwidth is much more, and collaboration development, such as rendering graphics.

Milojicic et al. (2002) give illustration application in cross matrix between P2P market and P2P activity, as can be seen in Table 3.

Table 2
Pros and Cons of P2P

	Pros	Contras
Collaboration	No agreements are needed before deploying server	Each collaboration is a new overhead
Standards	Plug-in capability is flexible and can track new standards	Need to support all variants
Costs	Running costs are low	Bu in costs are high
Performance	Limited only by extranet bandwidth and server CPU	May need high capacity server
Security	You have total control	You are responsible for server access control, etc

(Stephenson, 2002)

Table 3
P2P Application in matrix between P2P Market vs. P2P Activity

Type of Activity	Scope		
	Consumer	Enterprise	Public
@work	Collaboration, Communication	Distributed Computing, Storage, Communication, Collaboration	Communication, Digital rights management
@play	Games	HR-sponsored events	Digital media, Digital experience
@rest	Music sharing	Content consumption	Instant Messaging

(Milojicic, 2002)

The above implementations of P2P systems have give better illustration about ability of P2P. However in real world, some consideration should be taken to decide using P2P or not-P2P. The consideration is based on Roussopoulos et al. (2003). Tree decision for suitability of a P2P solution can be seen at Figure 1.

The consideration based on some factors. First, budget, it is main consideration to choose p2p, if the budget is limited, a key motivator in the choice of p2p system. Second, resource relevance to participants, is a peer interested to other peers? If the interesting is high, p2p system is needed. Third, trust consideration in mutual distrust between peers. Fourth, rate of system change; high rapid change in p2p system can make difficulty to provide consistency guarantees and defenses against flooding and other attack critically, if the system needs to solve critical system. We need carefully consideration about centralized and decentralized control for security and availability issues.

RESEARCH METHODOLOGY

The methodology to conduct the research as based on literature review. The main literature as from paper research, journal, proceeding and some white paper of real product. The important of white paper to bring an illustration to the classification as result of study.

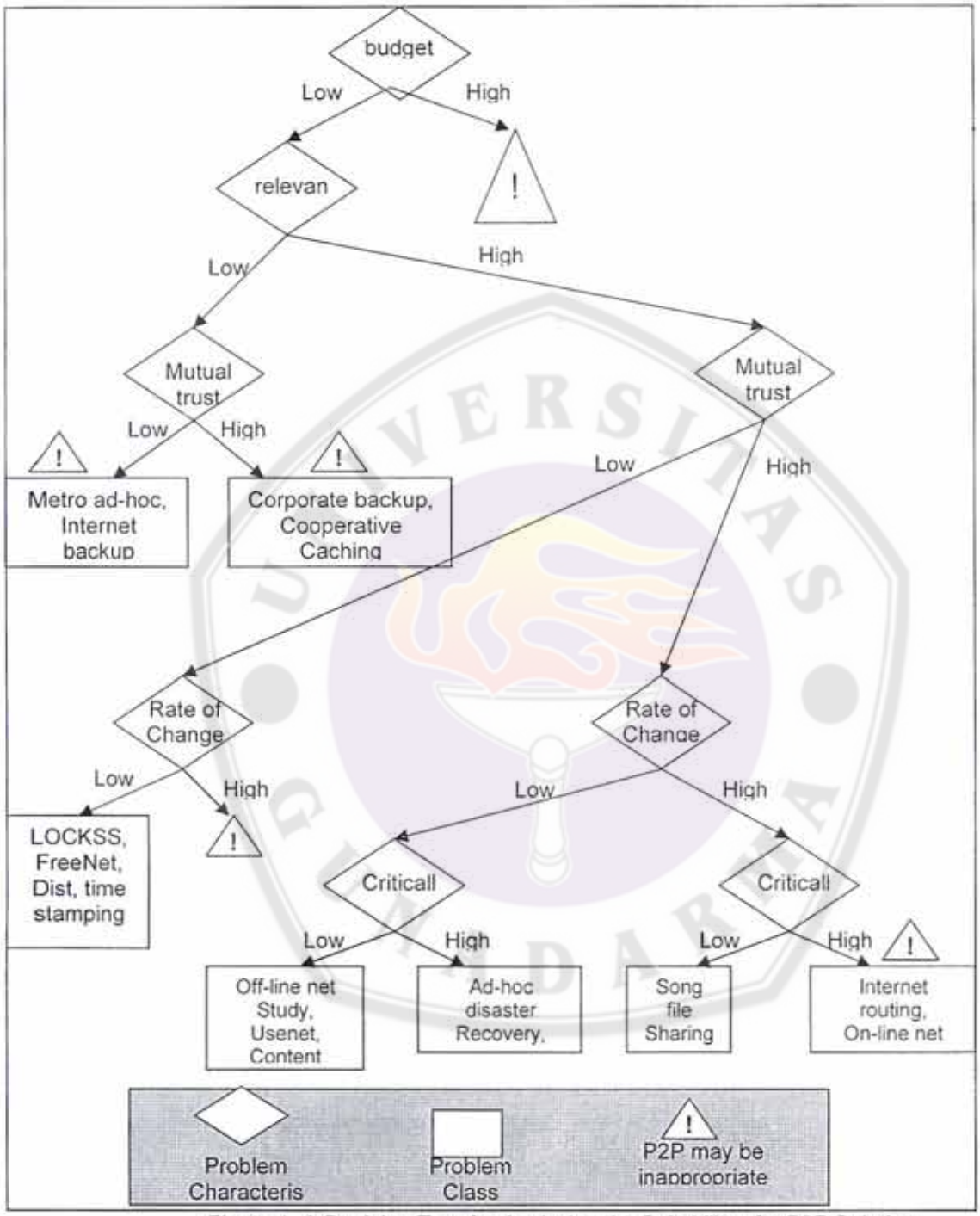

Figure 1. A Decision Tree for Analyzing the Suitability of a P2P Solution
(Roussopoulos, etc, 2003)

DISCUSSION

Towards of P2P Model: Architecture of a Peer

LsS Julee Mathew (Roussopoulos, 2003) gives modeling architecture of a Peer content three main layers, i.e. Web Service Foundation, Core Component, and Extension Modules. The model bases on a Peer use the HTTP protocol. Figure 2 illustrates internal structure of a Peer. Web server and Servelet Engine are foundation of the system. System plays with simple HTTP.

Request Manager is to handle HTTP request from node/peer (s) to five respond to be returned, or route the request to another module. Event Service has ability to carry communication among peers. General information at event service are the address of the creator, a local timestamp, a count indicating the local ordering, and globally unique identifier string. Buddy Manager is to organize the users and level of access. Module Container give new functionality for specific application or purpose of peer.

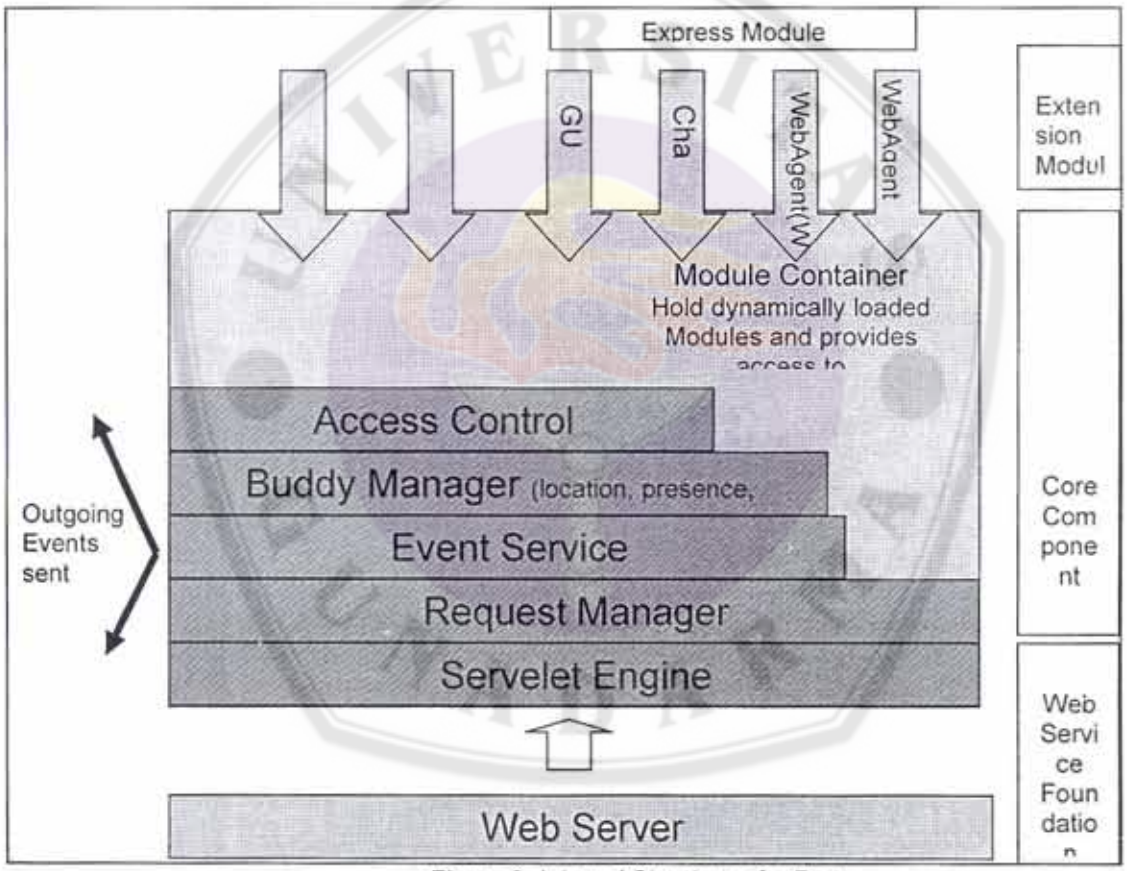

Figure 2. Internal Structure of a Peer (Mathew, 2002)

P2P Models

The terminology of P2P models, P2P network, P2P systems is changeable. In this paper, we tried to classification based on some point of view to make clear different P2P model. P2P model is more in logical level than physical level. First point of view is refer to degree of centralization, and then refer to network structure, organize network and the last one refer to searching method, figure 3 give summary of P2P Model.

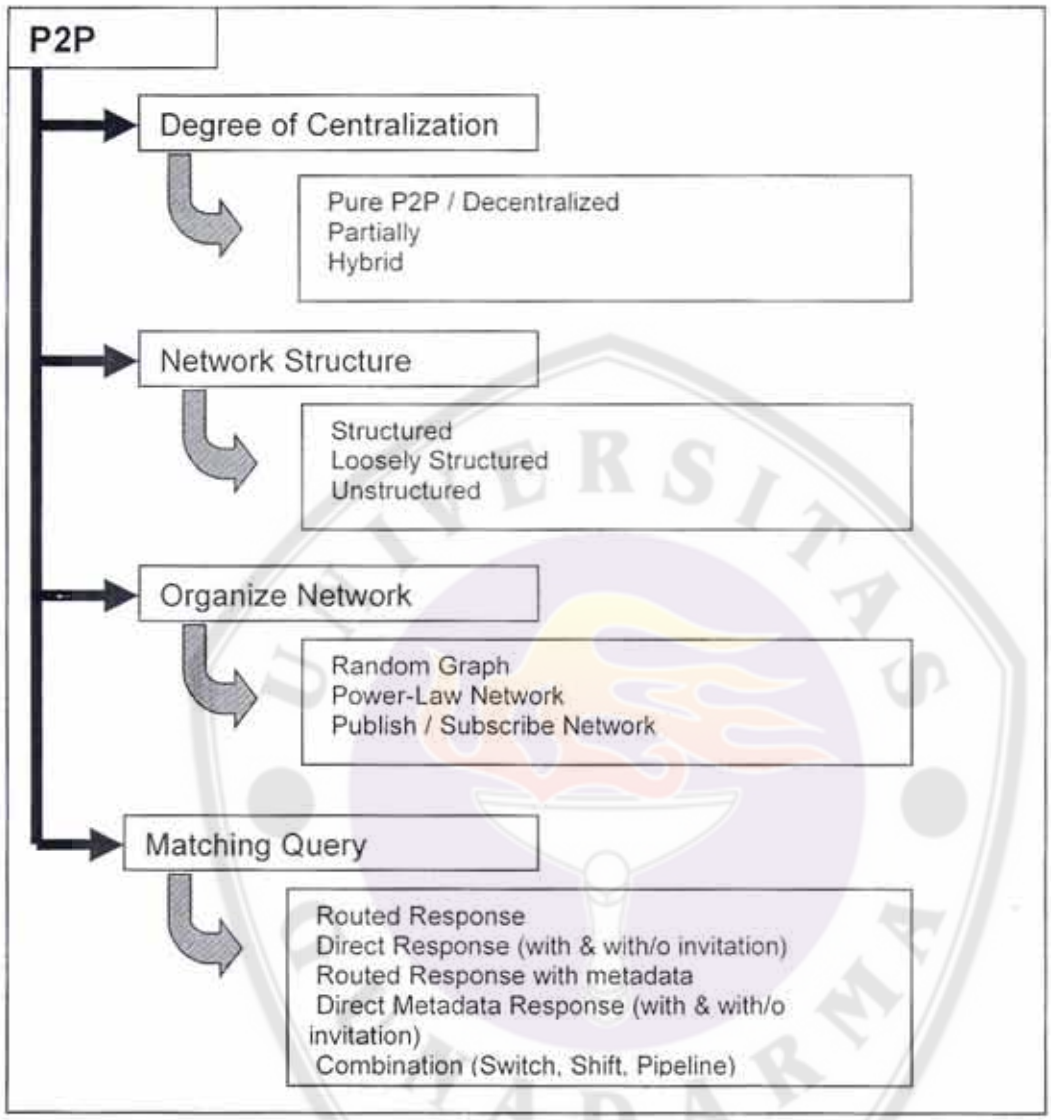

Figure 3. P2P Model

P2P Models Based on Degree of Centralization

Pure P2P/Decentralized

A pure P2P is no dedicated server to handle request. Each peer has same position and can be server or client (called servant). Method to find peer or information is flooding technique. For illustration can be seen Figure 4. Request from A will be broadcasted to all nodes (B, C, D, and E). For example node B can not find the request at its node, so the request will be re-broadcasted to all nodes. Example this model is Gnutella (www.gnutella.com).

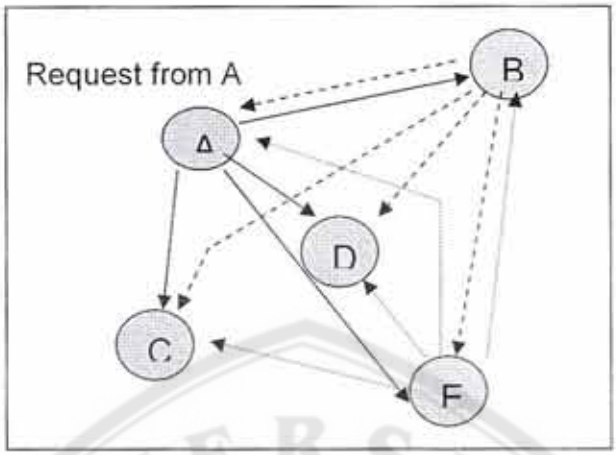

Figure 4. Pure P2P

Advantages of the model are no need server, independent to other peer, and limited per-node state. Disadvantages of the model are high traffic or bandwidth intensive, and slow search. Some improvements can be made are:

- To reduce the traffic by implementation TTL and/or maximum hops (Nejdl, 2003)
- Grouping or clustering peers at same interest, i.e. FUtella (Zahn, 2002). The similar approach is mapping for virtual topology (Ripeanu, 2000), i.e. Chord, CAN, SDS and OceanStore.
- Make history of success connecting at each peer, so searching will look at the history list (cache) from some networks. In other words, the request should not been sent to all nodes. The approach based on social network (friend-list) (Upadrashta, 2002)
- Edutella by implemented annotating resources (Nejdl, 2001)
- Range Addressable Network can be processed efficiently, research based on CAN system (Kothari, 2003)
- Combination of P2P and hierarchical tree structure architecture are to service discovery and resource allocation (Dowlatshahi, 2002)
- Butterflies approach is to improve CAN (Content Addressable Network) which DHT (distributed hash tables) for location of resources based on unique keys.(Datar, 2002)
- P2Prep is to choose reputable servents in P2P network by using Enhanced Pooling mechanism (Cornelli, 2001)
- Bidirectional direction for Chord is to optimal routing (Ganesan, 2003)
- Symphony is a novel protocol for maintaining distributed hash table in a wide area network. This protocol can be implemented for CAN to consider distance links per node. (Manku, 2003)

Model is based on pure P2P, but some nodes will be signed as super-peer or super-node than rest of the nodes. Some nodes act as Super-Peers because their capacity, connectivity or reliability. A Super-Peer can keep a list of connection nodes and speed up the join process. To select super node can be automatically or manual. In the implementation if a super-node down, nodes can select other super-node. If there is no more super-nodes, node can act as super-node for himself. Example is Kazaa, and recent Gnutella shift to this model.

Figure 5 as example of network, P0 sent query to super-peer SP1, SP1 will check at his member (P0 and P1), is there the answer? If no the query will route to other super-peer (SP2, SP3, and SP4) not directly to peer (P2, P3, and P4). This approach can reduce traffic and speed of searching. Improvement is Super-Peer-Based routing and clustering strategy with RDF-based (Nejdl, 2003).

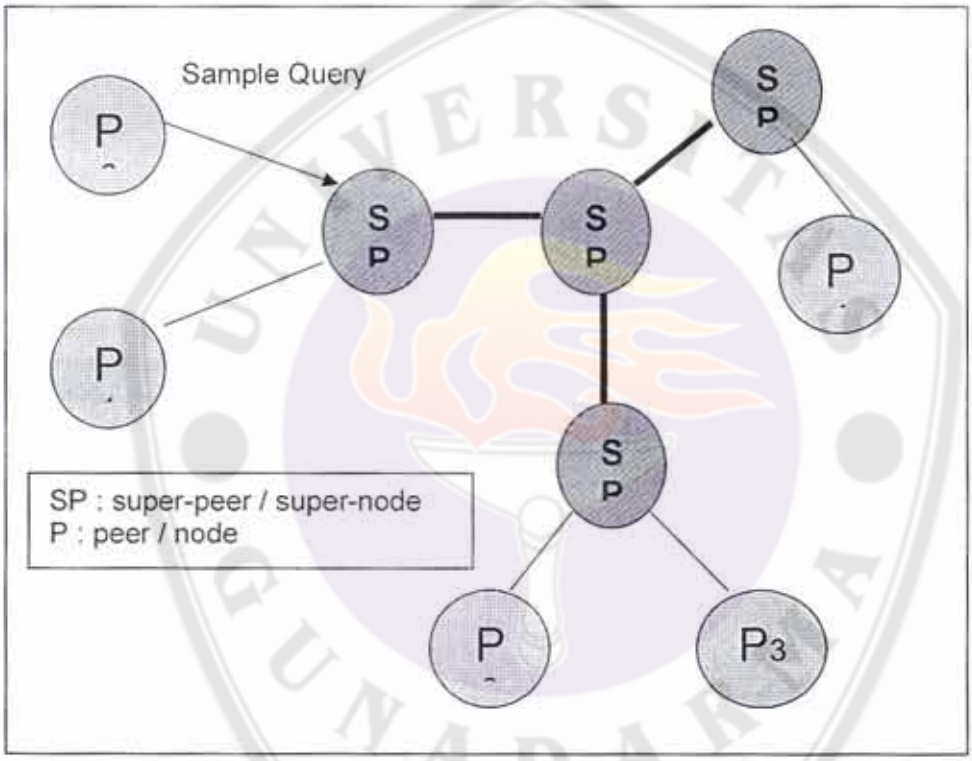

Figure 5. Partially P2P or Super-Node / Super-Peer

Hybrid P2P/Mediation

Basic mechanism hybrid P2P is same with pure P2P, but hybrid P2P need server(s) to improve the quality answering request. Currently, hybrid P2P is better than pure P2P, and pure P2P is appropriate for hundreds of peer. For illustration see figure 6, node 7 requests a Madonna song, the request sent to server to look at the index. For example the request of song is available at node 5. Node 7 will directly download the file of song from node 5 without server routing. Example of hybrid model is Napster (www.napster.com).

ISSN 0216-4728

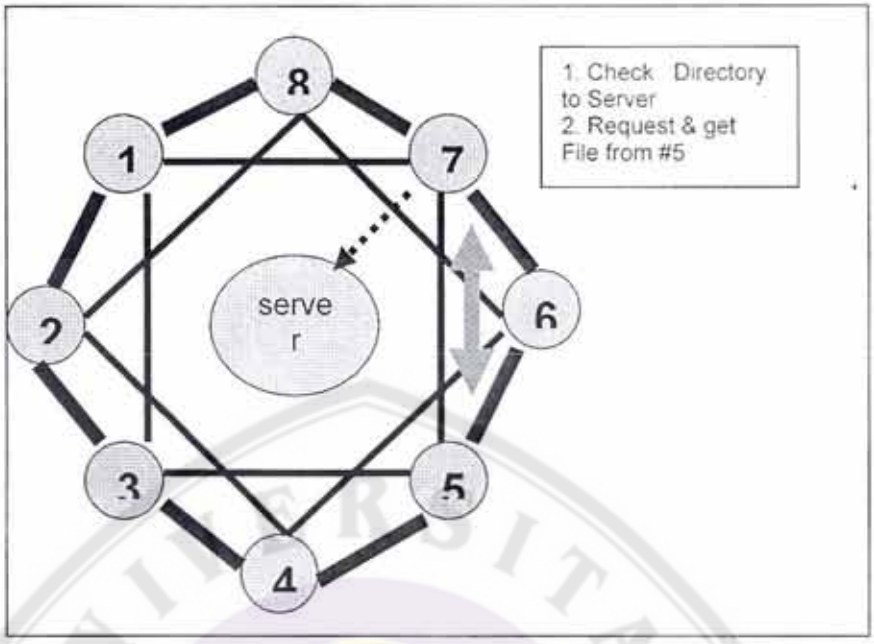

Figure 6. Hybrid P2P

Advantages of hybrid model are currently faster search than pure P2P, limited bandwidth usage, and no per-node state. Disadvantages of hybrid model are scalability depend on server, and system depend on server if server not active or network problem to server the system not run. Some improvements can be made are:

- Server replication where network connection will be more stable and memory is cheaper (Yang, 2001)
- Chained architecture is the best strategy for music-sharing systems (Yang, 2001).
- Advance P2P Architecture using Autonomous Agent to consider changeable system between pure to/from hybrid P2P .

P2P Model Based on Network Structure

Structured network has precisely specified location, or recognize form of network. The system provides a mapping between the file identifier and location, so queries can be efficiently routed to the node with desired file. Problem with structured network is to maintain the structure form, because joining and leaving node is high. Example is Chord, CAN, Past, Tapestry.

Loosely Structured network is in between of structured and unstructured network. The purpose to adopt the location in routing but has flexibility, if there are joining and leaving nodes. Disadvantage is not completely specified so not all searches succeed. Example is FreeNet.

The placement of data (files) is completely unrelated to the overlay topology. Main advantage for this network is high accommodating for big node population. Disadvantage is difficult to get file in efficient way. Example is Gnutella.

Improvements are AOTO (Adaptive Overlay Topology Optimization) (Liu, 2002) and Multiple random walks and uniform random graphs (Lv, 2001). Example of P2P model based on degree of centralization and network structure can be seen at Table 3.

Table 3
Example of P2P model based on degree of centralization and network structure

	Structured Networks	Loosely Structured Networks	Unstructured Networks
Pure Decentralized	Chord, CAN, Tapestry, Pastry	Freenet	Gnutella
Partially Decentralized			KaZaA
Hybrid Decentralized			Napster

(Androutsellis-Theotokis, 2002)

P2P Model Based on Organize Network

In the beginning, the organize network is just for pure P2P, but some papers also make simulation for hybrid by Power-Law Networks. Therefore, combination can be done as well for other than pure P2P.

Random Graph model is Gnutella Approach for flooding. Query may have TTL and maximum number of hops to reduce the traffic. Power-Law Networks approach look like super-node model, which a few nodes have high connectivity or power and many nodes low connectivity and power. This idea implemented for KaZaA (Kazaa). Publish/Subscribe Networks is by implementing P/S server for information consumers subscribe their need, propagate and aggregate subscription, publish their advertisements.

P2P Model Based on Matching Query

Request will be sent via one of neighborhood which called agent node. And from agent node sent to other nodes until get the answer, and this result will back to the originator through agent node using the path as before. Illustration can be seen at Figure 7. From originator (1) sends to agent node, and than sent to other nodes to get the answer (2,3, and 5). The result can be meet the request or not, if not find the request will sent to another node again (4 and 5). However if already find will return to agent node and originator via same path (6,7, and 8).

This routing will spend a lot of time to respond request. And the request can be sent more than one time to the same node.

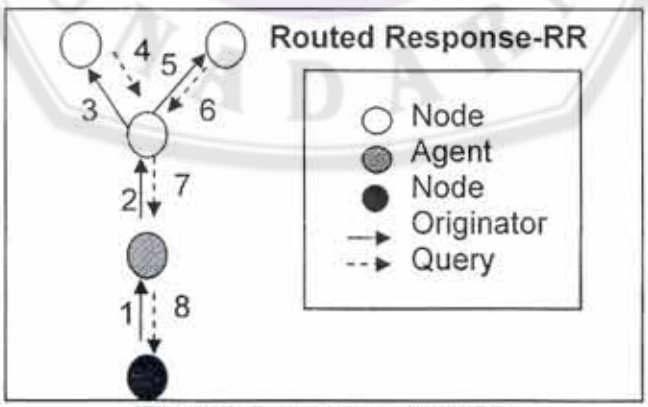

Figure 7. Routed Response (RR)
(Hoschek, 2002)

ISSN 0216-4728

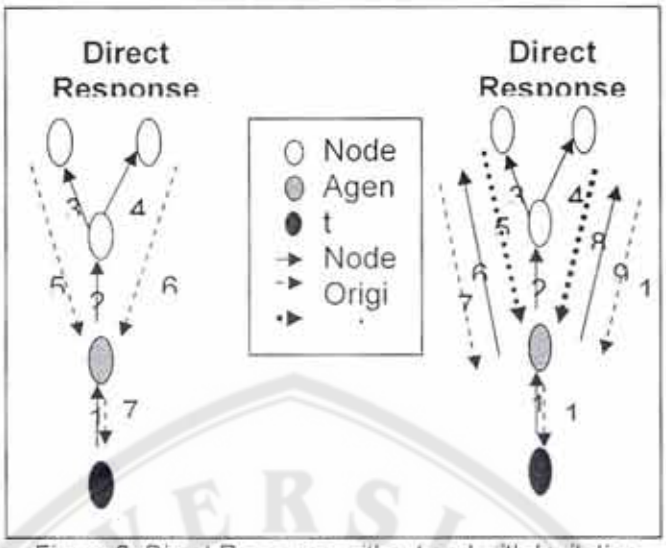

Figure 8. Direct Response without and with Invitation (Hoschek, 2002)

Direct response are two models: with and without invitation. Figure 8 at left site is illustrated Direct Response without Invitation. Request from originator will be sent to agent mode, and then agent mode sent to other nodes. If response of answer is available, the response will be sent directly to agent node (5 and 6), not through previous path. Problem for without invitation, agent node can be overloaded by incoming of results from many nodes. To overcome this problem, invitation procedure is implemented to ask nodes to send the answer by agent node, look at Figure 8 at right side.

Basic mechanism of routed metadata response is same with routed response (4.2.4.1), however there are two phase to give result. First phase is send metadata result. Second phase, based on metadata result, agent node can filter to choose appropriate answer. So, the agent node will get the full data after select from metadata. The path of routing can be seen at Figure 9.

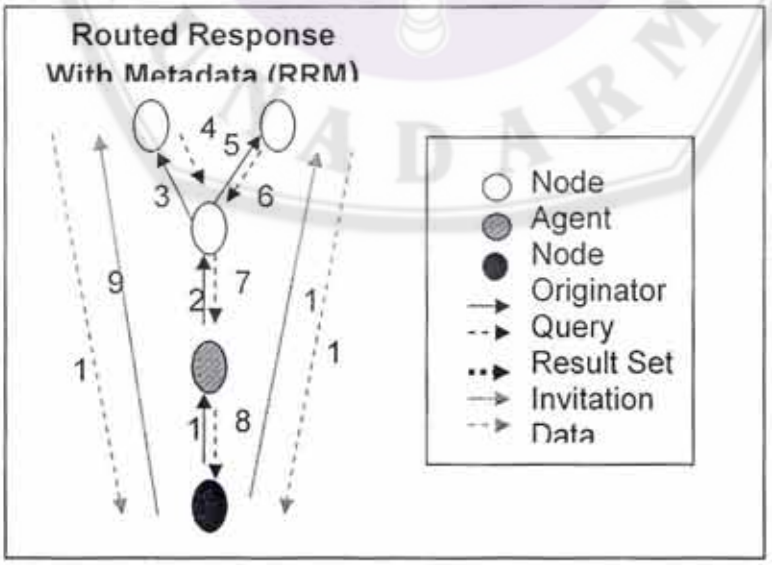

Figure 9. Routed Response with Metadata (RRM) (Hoschek, 2002)

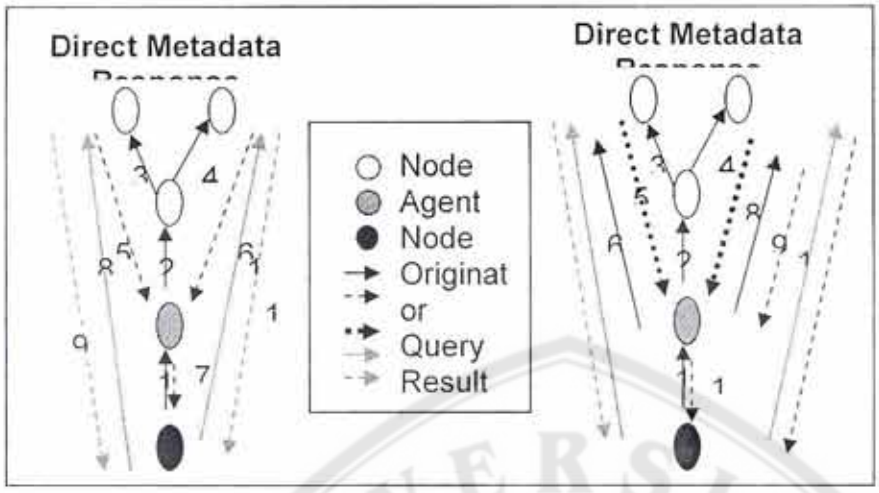

Figure 10. Direct Metadata Response with and without Invitation (DRM) (Hoschek, 2002)

Basic mechanism of direct metadata response (with and without invitation) is same with direct response (4,2.4.2). However there are two phase to give result. First phase is send metadata result. Second phase, based on metadata result, agent node can filter to choose appropriate answer. So, the agent node will get the full data after select from metadata. The path of routing can be seen at Figure 10.

Matching query can be improvement by using combination some four above models by implementing shift and switch method or adopt pipeline principle. By shift and switch can get the optimal method in request the response, however the problem what type matching will be used and which node to be the point of changing. Pipelining can improve time of searching, because to search next step without waiting previous request until completed.

CONCLUSION

P2P as one of distributed system solutions, has been reincarnation to meet with current technologies and problems. This paper has been discussed some aspects of P2P from characteristics of P2P, how to decide P2P or not P2P, some example of implementation.

Main content of the paper is summarizing current available P2P model, and some approach to optimize current model. In this paper P2P model categorized based on: degree of centralization, network structure, organize network and matching query. Many papers using mixed terminology, so the meaning network, system, model will be confusing. This paper tries to figure out the more complete picture of P2P model. Every model has disadvantages or problem, in this paper also write down some efforts have been done to eliminate the problems.

Other issues related with P2P are still there. The paper also outlines the summary main problem in purpose to bring P2P for information integration and sharing for the real world. Distribution of this paper can give foundation for implemented P2P in special case by looking at the complete picture of P2P model.

REFERENCES

Androutsellis-Theotokis, S. "A Survey of Peer-to_Peer File Sharing Technology", in Proceeding ACMII2002, 2002

Barkai, D. "An Introduction to Peer-to-Peer Computing,", http://www.intel.com/research/introp2p.html, access Jul 2005, 2000.

Cornelli, F., E. Damiani, S.D.C.d. Vimercati, S. Paraboschi and P. Samarati "Choosing Reputable Servents in a P2P Network (Slide)," pp. 1-19, 2001.

Datar, M. "Butterflies and Peer-to-Peer Networks,", in Proceeding ACMII2002, 2002.

Dowlatshasi, M., G. MacLarty and M. Fry "A Scalable and Efficient Architecture for Service Discovery," 2002.

Hoschek, W. "A Unified Peer-To-Peer database Framework For XQueries Over Dynamic Distributed Content And Its Application For Scalable Service Discovery," PhD Thesis, 2002.

Kothari, A., D. Agrawal, A. Gupta and S. Suri "Range Addressable Network: A P2P Cache Archtecture for Data Ranges,", http://www.opencontent.org/p2p/ran.html, accessed July 2005, 2003.

Manku, G.S., M. Bawa and P. Raghavan "Symphony: Distributed Hashing In A Small World," pp. 1-14, 2003.

Milojicik, D., etc "Peer-to-Peer Computing", http://citeseer.ist.psu.edu/milojicic02peertopeer.html, accessed Okt 2006, 2002

Nejdl, W., B. Wolf, S. Staab and J. Tane "EDUTELLA: Searching and Annotating Resources within an RDF-based P2P Network," http://citeseer.ist.psu.edu/staab01edutella.html, accessed Okt 2006, 2001.

Nejdl, W., M. Wolpers, W. Siberski, C. Schmitz, M. Schlosser, B. Ingo and A. Loser "Super-Peer-Based Routing and Clustering Strategies for RDF-Based Peer-To-Peer Networks," http://www.kbs.uni-hannover.de/Arbeiten/Publikationen/2002/www2003_superpeer.pdf, 2003.

Ripeanu, M "Peer-to-Peer Architecture Case Study: Gnutella Network", http://www.cs.uchicago.edu/~matei/PAPERS/gnutella-rc.pdf, accessed Jul 2005, 2000

Roussopoulos, M., M. Baker, D.S.H. Rosenthal, T. Giuli, P. Maniatis and J. Mogul "2 P2P or Not 2 P2P,", http://www.eecs.harvard.edu/~mema/courses/cs264/papers/iptps2004.pdf, accessed August 2006, 2003.

Stephenson, J. "Managed P2P - A New Architecture for Service Management," CBDI Jounal ,2002.

Upadrashta, Y. "Emerging Social Networks in Peer-to-Peer Systems,", http://citeseer.ist.psu.edu/706391.html, accessed March 2005, 2002.

Zahn, T., H. Ritter, J. Schiller and H. Schweppe "Futella - Analysis and Implementation of a Content Based Peer-to-Peer Network," 8th Netties Conference, Technicsche Universitat Ilmenau 2002.