社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－ＤＳＭ－46／2007－ＱＡＩ－24（10）
2007／7／19

自律的に平衡２分木を保とうとする構造型 P2P システム

山之上 卓† 中森 武‡

† 鹿児島大学学術情報基盤センター 〒890-0065 鹿児島市郡元 1-21-35
‡ 奈良情報システム 〒630-8247 奈良市油阪町 446-16
E-mail: † yamanoue@cc.kagoshima-u.ac.jp

あらまし 自律的に平衡２分木状の結合を保とうとする構造型 P2P システムについて述べる。新たにノードが加わったり離脱したりするとき、およそノード数の対数に比例した時間でこれらの処理が終了する。このシステムを応用した電話と放送を融合したシステムの開発を行っている。

キーワード P2P, TCP/IP, 自律システム, 平衡, 通信システム, 放送システム

A Structured P2P System Which Tries to Keep Its Balanced Binary Tree Shape by it-self

Takashi YAMANOUE† Takeshi NAKAMORI‡

† Computing and Commuications Center,
Kagoshima University 1-21-35 Korimoto, Kagoshima 890-0065, Japan
‡ Nara Information System, Ltd. 446-18 Abura-zaka, Nara 630-8247, Japan
E-mail: † yamanoue@cc.kagoshima-u.ac.jp

Abstract A structured P2P system, which tries to keep its balanced binary tree shape by it-self, is presented. When a new node take part in the group or when a node leaves from the group, these procedures are executed in O(log N) time. We are developing a system which is a fusion of telephone system and broadcasting using the P2P system.

Keyword P2P, TCP/IP, Autonomous system, balance, communication system, broadcast system.

1. はじめに

インターネットを使った VoIP やテレビ電話が普及しつつある。インターネットを使うことにより、安く手軽に、いつでも、どこでも、通話することが可能になりつつある。しかしながら、従来の SIP サーバを用いた IP 電話や、MCU を用いたテレビ会議システムの場合、サーバや MCU などの一箇所に機能が集中し、そこに障害が発生した場合の影響が大きい。また、その一箇所に負荷が集中しやすく、一斉に使われたときに大幅な機能低下が起ったり障害が発生したりする場合がある。また、簡単に稼動を止めることができない巨大な通信システムの場合、障害対策を厳重に行わなければならず、サーバの保守の時間調整を行うだけでも大変であり、管理運営の担当者の負荷も大きい。

このような問題を解決する一つの手段として、P2P 技術を使った実時間通信システムが注目を集めており、各所で研究が行われている。また、Skype[4] など、すでに広く利用されている P2P を使った通信システムも存在する。

我々は電子黒板[1][2]などの教育支援システムに関する研究を行っているが、これも P2P 技術を使った実時通信システムの一つである。しかしながら、従来の電子黒板には「グループマネージャ」が必要であった。グループマネージャは、ノードがグループに参加したりグループから離脱したりしたときにグループ内のノード間の結合を制御するプログラムである。このことは、グループ管理に関する機能がグループマネージャ一箇所に集中していることを表し、接続/離脱が集中したときの機能低下の原因になっていた。

我々は、従来の電子黒板が持つこのような欠点を克服し、電話と放送を融合した通信システム「NARA」の試作を行っている。NARA はノード間を平衡２分木状に接続することにより、ノードの検索や会話の遅延をノード数の対数に比例した時間に収める機能を持った構造型 P2P システムの一つである。この報告では、NARA が利用しているノード同士が自律的に平衡２分木を保つよう努力する機構を中心に述べる。

2. NARA の機能

NARA は P2P 技術とインターネット/イントラネットを用いてリアルタイムで音声・動画の交換、放送を可能にするシステムである。一種のテレビ電話システ

ムであり、一種の動画放送システムである。

NARA は、

- 通話先の識別は IP アドレスではなく、名前で行う。これにより、通話先のコンピュータがネットワークを移動したときでも名前を指定して接続できる。通話先のホストがどこにあるか検索を行うとき、全てのノード間通信の容量および遅延が同一で、ノードの処理時間も同じであるなら、ノード数が N のとき、検索時間は O(log N) 以内である。以下、同じ条件を仮定する。
- 新規利用者の参加や離脱も柔軟にできる。参加や離脱に必要な処理時間は O(log N)以内である。
- 複数の利用者で構成される複数のグループ内で同時に TV 通話が可能である。現在、1 つの通話グループ内で、受信者数は無制限で同時送信可能者数は 5 人としている。O(log N')(N' は受信者の数)の時間以内の遅延で全ての受信者に画像が行きわたる。
- 電話の転送が可能である。
- グループ内でお絵かきソフトを共有できる。このお絵かきソフトでパソコンの画面も共有できる。
- ノード間にファイヤーウォール(NAT)があっても通話できる。
- 高性能なサーバを必要としない。
- 一部の機器の障害が影響を及ぼす範囲が狭い。

などの機能を持つ予定であり、現在、LAN や JGNII を使った実験を通じてデバッグや改良を行っている。

3. NARA の構造

「NARA システム」(以下、システム) は「端末システム」同士が接続されることにより構成されている。端末システムは、利用者が使用するホストコンピュータで動作する通話端末プログラムである。端末システムは、このシステム上で利用者を識別する名前(ユーザ ID)、「通話先キャッシュ」、「待機グループノード」、「通話グループノード」、「代理オブジェクトの集合」、「代理オブジェクトのバックアップ集合」、「代理オブジェクトへの TCP 接続」、「代理オブジェクトのバックアップへの TCP 接続」「応用プログラムの集合」、「制御プログラム」、などを持つ。図 1 に端末システムの構成を示す。

「通話先キャッシュ」は以前通話した通話先の<名前,アドレス,ポート>の組を記憶しておくキャッシュで、利用者側から見ると通話先の検索時間を削減することができる。システム側から見ると通話先検索を削減するによる負荷軽減を図れる。

「待機グループノード」は待機グループのメンバーである。待機グループは通話先を検索するために構成されるグループである。待機グループノードが待機グループに参加することは、従来の電話システムにおける電話機が電話交換機に接続されることに対応する。待機ノードが待機グループに参加することを、その待機ノードを持つ端末システムが「システムに参加する」と呼ぶことにする。待機グループは、待機グループノードが平衡 2 分木状に結合されることにより構成される。この木の根より並列に相手先を検索することにより、2 章で述べた条件が満たされれば、ノード数の対数に比例した時間内で検索が終了する。このとき交換されるメッセージの総数はノード数に比例する。待機グループノードは上位ノードへの TCP 接続、2 つ上位のノードの IP アドレスとポート、左右のノードから接続される TCP ソケット(左右のノードからの接続)、左右に接続される部分木のノード数を表す変数「左の重さ」と「右の重さ」、その待機ノードを持つ端末システムへのリンクなどから構成されている。平衡 2 分木を構成するために左右の重さを使用する。左右の重さはノード間で定期的に情報交換することにより計算される。ある端末システムにおいて、その待機ノードの左に接続した待機ノードを持つ端末システムを「左下の端末システム(または単に左下)」、右に接続した待機ノードを持つ端末システムを「右下の端末システム(または単に右下)」と呼ぶことにする。

「通話グループノード」は通話グループのメンバーである。通話グループはそのグループに参加する利用者の間の会話や画像を共有するグループである。一対一の会話のとき、グループの参加者は 2 となる。通話グループはシステム全体で複数同時に存在することができる。通話グループも、通話グループノードが平衡 2 分木状に結合されることによって構成される。待機グループのときと同じ条件であれば、1 つの通話グループ内の通信はノード数の対数に比例した時間内の遅延ですべてのノードに伝わる。通話グループノードも待機ノードと同様に構成される。

「代理オブジェクト」は、インターネットやイントラネットのバックボーン側から見て NAT の背後にあるローカルアドレスを持った端末システム(「ローカル端末システム」)がシステムに参加するとき、そのノードの代理として、グローバルアドレス (イントラネットの場合は、イントラネットのバックボーンアドレス) にある端末システム(「グローバル端末システム」)が持つプログラムである。ローカル端末システムは、代理オブジェクトへの TCP 接続によって、システムに参加する。このとき、待機ノードは使われない。一般的には、1 つのグローバル端末システムが複数の代理オブジェクト(「代理オブジェクトの集合」)を持つ。そ

の集合の大きさがグローバルアドレス上の端末システムによって偏りが少なくなるように配置される。代理オブジェクトは、これが代理する端末システムが通話するための通話グループのノードを持つ。これにより、ローカル端末システムも、あたかもグローバル端末システムと同様に振舞うことができる。

「**代理オブジェクトのバックアップ集合**」は、代理オブジェクトのバックアップの集合である。代理オブジェクトを持つホストに障害が発生した場合にも、それぞれの代理オブジェクトに対応するローカル端末システムに対する障害の影響をできるだけ小さくしたい。そのため、代理オブジェクトを持つホストとは別のグローバルアドレスのホストにバックアップを持たせている。グローバル端末システムは、そのホストの左下の端末システムの代理ノードのバックアップの集合と、右下の端末システムの代理ノードのバックアップの集合の2つを持つ。

「**応用プログラムの集合**」は、画像・音声入出力を行う TV 電話端末プログラム、チャットプログラム、お絵かきプログラムなど、利用者と直接対応するプログラムの集合である。応用プログラムの入出力データは、通話グループノードを通じて他のホストノード間と交換される。ローカル端末システムの場合、その応用プログラムは代理オブジェクトへの TCP 接続を経由して、その代理オブジェクトが持つ通話グループノードと接続される。

「**制御プログラム**」はホストノードの構成要素を管理するプログラムである。

4. 平衡2分木の維持

本システムの待機グループと通信グループの各ノードは、定期的に左右の重さなどのノード情報を更新し、左右の重さができるだけ同じになるようにノードの追加を行うことによって、それぞれのグループが平衡2分木に近づくようにしている。以下に、定期的なノード情報の更新、新規ノードの追加処理、ノードの障害発生時と離脱時の処理について述べる。

4. 1 定期的なノード情報の更新

待機グループまたは通話グループに参加しているノード(以下ノード)は、一定期間ごとに以下を行い、左右の重さを計算する。

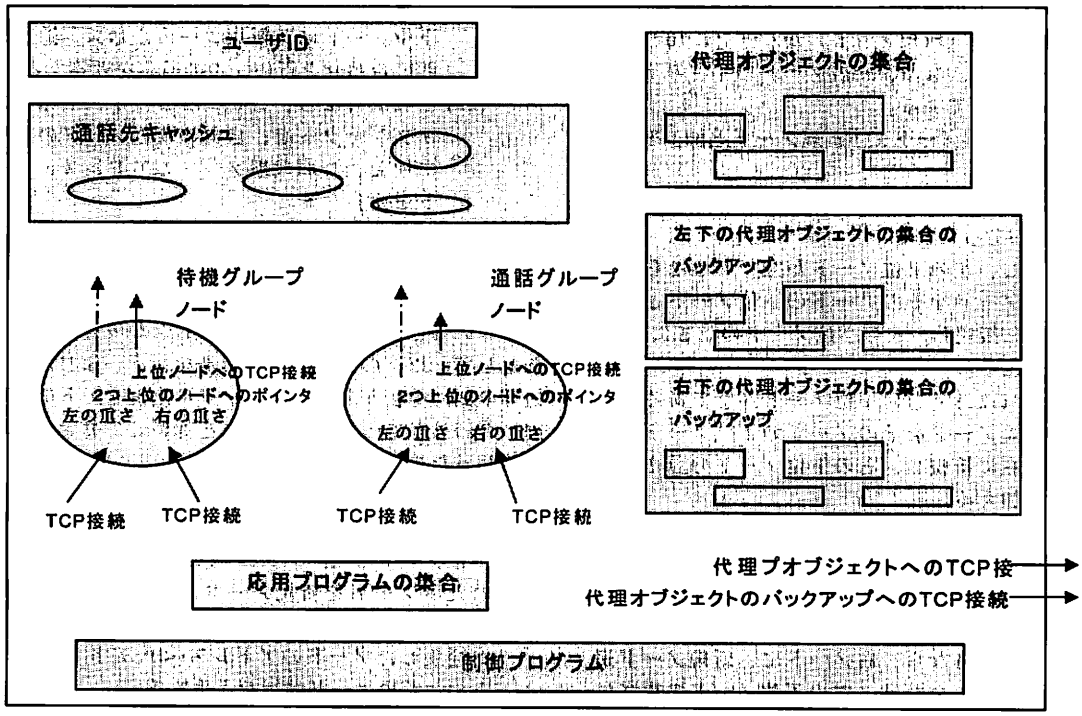

図 1. 端末システム

1. 左の重さ+右の重さ+1 を上位ノードの左(または右）の重さに代入する。もし自分が根である場合、上位ノードに重さを伝えない。
2. 左右それぞれについて下位ノードを持たない場合は、その下位ノードの重さを 0 とする。
3. 2つ上位のノードの IP アドレスとポートの検索し記憶する。この情報は 1 つ上位のノードに障害が発生した場合に使用する。根とそれに直接接続されるノードの場合は、「空」を記憶する。
4. 待機ノードの場合、自分の端末システムの代理オブジェクトの数を上位の待機ノードを経由して上位の端末システムに伝える。また、自分のノードの代理ノードの集合を上位ノードに伝え、バックアップする。このとき、対応するローカルノードはバックアップに対する TCP 結合も持つ。
5. 自分の上位ノードから反応がない場合、上位ノードに障害が発生したかグループから離脱したものとみなし、上位ノードの交換処理を行うことにより修復する。この「ノード交換処理」については後で述べる。

4.2 新規ノードの参加

新たなノード（新規ノード）がループに加わるとき、以下を行う。なお、新規ノードの左右の下位ノードは空となっている。

1. グループの根ノードに「新規ノード追加」を問い合わせる。
2. 任意のグループノードに新規ノード追加の問い合わせがあったとき、
   (ア) もし、左の下位ノードが空であるなら、新規ノードをそのノードの下位ノードとして左に接続し、左の下位ノードの重さに 1 を加える。
   (イ) それ以外で右の下位ノードが空であるなら、新規ノードをそのノードの下位ノードとして右に接続し、右の下位ノードの重さに 1 を加える。
   (ウ) それ以外で左ノードの重さが右ノードの重さ以下であるなら、左ノードに、新規ノード追加を問い合わせ、左の下位ノードの重さに 1 を加える。
   (エ) それ以外なら、右ノードに、新規ノード追加を問い合わせ、右の下位ノードの重さに 1 を加える。

以上のアルゴリズムにより、新規にノードが加わったとき、完全 2 分木状にノードが結合される。グループのノード数を N としたとき、新規ノードの参加に必要な時間は（全てのノード間通信容量や遅延時間が同じで、ノードの処理時間も同じであれば)、O(log N)であり、木の高さも O(log N)である。

ローカル端末システムがシステムに参加する場合は以下を行う。

1 ローカル端末システムが待機グループの根ノードに「新規ローカルホスト追加」を問い合わせる。
2 代理ノードのアドレスとポートが返り、それに対してローカル端末システムが TCP 接続を行う。
3 任意の待機グループノードに新規ローカルホスト追加の問い合わせがあったとき、以下を行う。
   (ア) もし、左右の下位ノードが空であるなら、新規ローカルホストの代理オブジェクトを、作成し、自分の代理オブジェクトの集合に加え、その代理オブジェクトのアドレスとポートを返答する。
   (イ) それ以外で左下の代理オブジェクト数が右下の代理オブジェクト数以下で、なおかつ、自分の代理オブジェクト数以下であれば、左下のノードに、新規ローカルホスト追加の問い合わせを行い、この返事を返答する。
   (ウ) それ以外で右の下位ノードの代理ノード数が自分の代理ノード数以下であれば、右の下位ノードに、新規ローカルホスト追加の問い合わせを行い、この返事を返答する。
   (エ) それ以外なら、新規ローカルホストの代理オブジェクトを、自分の代理ノードの集合に加え、この代理ノードのアドレスとポートを返答する。

4.3 ノードの障害発生時や離脱時の処理

定期的なノード間情報交換時に上位ノードからの反応がない場合、上位ノードが離脱したか障害が発生したとして、以下のノード交換処理を行う。なお、この処理の間、障害が発生したノードの上位ノードに各種問い合わせがあった場合、これを保留しておき、処理が終わった後、実行する。

1. 上位ノードの反応がない場合、「葉ノード取り出し処理」を行い、自分の<IP アドレス,ポート>と、上位ノードの<IP アドレス,ポート>と共に、得られた葉ノードを新規ノード候補として 2 つ上の上位ノードに伝え、2 つ上の上位ノードを、障害が発生したノードのアドレス(<IP アドレス,ポート>)と新規ノード候補を入手する。
2. 障害が発生したノードには左右 2 つのノードが接続している場合がある。障害が発生したノードの上位ノードは、先に受け取った新規ノード候補

を新規ノードとする。これに加えて結果と共に送られてくる左右の下位ノードの情報を入手して、「接続変更処理」を行う。

「ノード交換処理」により、処理を行った部分木の重さは一つ減り、木のバランスが崩れる場合がある。しかしながら、重さの変化が全ノードに反映されるO(logN)時間後より、新たにノードが追加されるとき、重さが軽い部分木に追加されるため、常に、ノードが完全2分木状になるよう、自律的に調整が行われる。

「葉ノード取り出し処理」

障害発生時でノード置き換えのために行う葉ノード取り出し処理は、以下のように行う。

1. 葉ノード取り出しの要求を受けたノードが葉ノードの場合(左右の下位ノードが空の場合)、結果として自分の<IP アドレス,ポート>を返す。
2. そうでない場合、下位ノードの左右の重さを比較し、左が右より重ければ、左のノードに対して葉ノード取り出し処理を行い、結果を返す。
3. そうでない場合、右ノードに対して葉ノード取り出し処理を行い、結果を返す。

以上の処理で、O(logN)時間以内で葉ノードを取り出すことができる。

「接続変更処理」

1. 補われるノードをA, Aの新規上位ノードをB, Aの新規の左の下位ノードをC, Aの新規の右の下位ノードをDとする。
2. BはA,C,D の<IP アドレス,ポート>を入手する。(C,D は ノード欠落時に、それぞれ自身からBに通知されている。Aは、C,D それぞれが葉ノード取り出し処理により候補を探し、どちらか早くBに通知されたものである。)
3. Bの左下位ノードが障害発生ノードの場合、Bの左下位ノードにAを入れる。Bの右下位ノードが障害発生ノードの場合、Bの右下位ノードにAを入れる。
4. BはAに C,Dを引き渡す。
5. AはAの左下にCを入れ、右下にDを入れる。C,D は空の場合もある。
6. AとCが同じ場合は、Aの左下ノードを空とする。同様に A と D が同じ場合は、A の右下ノードを空とする。

定期的なノード間情報交換時に、左右のどちらかの下位ノードからの反応がない場合、下位ノードに障害が発生したとして、以下のノード交換処理を行う。

1. 左右のノードのうち、反応のないノードの方の重さが2以上であるなら、2つ下のノードからノード交換処理の要求が来るのを待ち、要求が来たら対応する接続変更処理を行う。
2. そうでない場合、反応のないノードは葉ノードであるので、以下を行う。
   (ア) 反応のなくなった側の、代理ノードの集合のバックアップを、代理ノードの集合に加える。
3. 上下どのノードからも反応がない場合、自分を新規ノードとして待機グループに加える。

5. 通話の開始

通話を開始するとき、端末システムは相手の「ノード検索」を行い、結果が空であるなら、終了（相手が見つからない旨を表示）し、そうでなければ、以下を行う。

1. もし、自分がグローバル端末システムであれば、相手の通話ノードを通話グループの根ノードとし、根ノードに対して、自分の通話ノードの接続を試みる。
2. 自分がローカル端末システムであれば、相手の通話ノードを通話グループの根ノードとし、根ノードに対して、自分の代理オブジェクトの通話ノードの接続を試みる。

「ノード検索」

相手を呼び出すとき、最初に相手のユーザIDを持ったノードがどこにあるか検索する「ノード検索」を行う。ノード検索は以下のように行われる。

1. 自分の通話先キャッシュの中に相手ユーザIDを持ったノード情報が格納されているか調べる。もし、格納されていれば、そのノードに接続を試みて、成功すれば、これを呼び出しノードとする。失敗したら以下を行う(通信先ノードを保持することにより、待機グループにおける「ノード検索」の負荷を削減している)。
2. それ以外の場合、待機ノードグループの根ノードに対して、ノード並列検索を要求する。結果が空でない場合、これを呼び出しノードとすると同時に、これを通話先キャッシュに加える。結果が空である場合、相手がネットワークに接続していない旨を利用者に伝える。

「ノード並列検索」

任意の待機グループノードに対してノード並列検索が要求されたとき、以下を行い、結果を要求したノードに返す。

1. 検索アドレスが、自分(待機グループのノード)を含むノードのユーザIDならば、自分の通話グループノードのIPアドレスとポート番号を結果

とする。

2. 検索アドレスが、自分の代理オブジェクト集合のノードの中にあれば、その（代理集合の）ノードのIPアドレスとポート番号を結果とする。
3. それ以外で左右どちらのノードも空であるなら、空を返す。
4. それ以外の場合
   (ア) もし、左の下位ノードが空でないなら、左の下位ノードにたいしてノード並列検索を要求する。
   (イ) これと並行して右の下位ノードが空でないなら、右の下位ノードに対してノード並列検索を要求する。
   (ウ) 上の2項の結果が返るのを待つ。左右のどちらかから、空でない結果が返った場合、すぐにこれを結果とする。どちらとも空であるなら、空を結果とする。

ノード並列検索に必要なメッセージの総数は木を構成するノード数に比例する。しかしながら、今まで仮定していた条件を満たせば、検索に必要な時間はノード数の対数に比例する。

6. 関連研究

SKYPE[4]は5人までの同時会話が可能である。しかしながら、多人数でTV会議を行うことはできない。

木村は順序関係のあるノードIDを使うことにより、自律的に木を構成するP2Pシステムに関する研究を行っている[3]。このシステムでは、ノードIDの分布に偏りがある場合、木にも偏りが生じる。また、新規ノードが追加されたり離脱したりした場合、複数の場所で枝の付け替えが生じる場合がある。

BATON[5]は順序関係があるノードを、その順序関係を守ったまま自律的に平衡2分木を構成するものである。BATONの各ノードは、最大で全ノード数の対数に比例した大きさの表を持ち、その表を管理する必要がある。また、ノードの参加や離脱により複数のノードが移動する場合がある。

上田らはノード間の距離を使った階層的クラスタリングについて述べている[6]。これは一種の自律的な階層構造の構成方法であり、距離の情報が使われているため、ノード間通信の性能が良くなる。しかしながらこの論文は木を構成する方法について述べたものではない。ノード間結合にループがあるときグループ内で放送を行うときには無限ループが生じないように工夫が必要となる。

堀内らは多人数が参加するP2Pテレビ会議システムに関する自律的な木の構成方法について述べている[7]。このシステムでは複数のリーダノードのネットワークがあり、それぞれのリーダノードに木がぶら下がっているような構造を持っている。木の根より循環的（ラウンドロビン）に新規ノードを追加する枝が選ばれることにより、木のバランスが取られる仕組みを持っている。この機構の場合、ノードの離脱により木のバランスが崩れたとき、それが修復されない場合がある。

7. おわりに

このシステムでは端末システムがシステムに参加する時に根ノードを知っている必要がある。この欠点の克服については、次回以降に報告する予定である。また、相手先を検索したり新規ノードを追加したりするとき根ノードに、他と比べて大きな負荷がかかる欠点や、検索時にノード数に比例したメッセージ交換が行われる欠点を持っている。これらの欠点については今後検討していく予定である。

8. 謝辞

本システムの研究開発に際し、JGNIIを利用した。

参考文献

[1] Takayuki Hirahara, Takashi Yamanoue, Hiroyuki Anzai and Itsujirou Arita, "SENDING AN IMAGE TO A LARGE NUMBER OF NODES IN SHORT TIME USING TCP", Proceedings of the ICME2000, IEEE International Conference on Multimedia and Expo, pp.987-990, New York City, USA, July 30-Aug.2 ,2000.

[2] 平原貴行，山之上卓，安在弘幸，有田五次郎：TCPを利用した分散ネットワーク環境のための電子黒板システム，情報処理学会論文誌，vol.43, No.1, pp.176-184, 2002.

[3] 木村綾太郎，" P2P 論理ネットワークの自律構成手法とその実装"，NRI 技術創発 2004, vol. 4. pp.57-76, 2004.

[4] 池鴫俊「SKYPEの仕組み」日経BP社, 2005.

[5] H. V. Jagadish, Beng Chin Ooi, Quang Hieu Vu, "BATON: A Balanced Tree Structure for Peer-to-Peer Networks". Proceedings of the 31st VLDB Conference, pp.661-672 Trondheim, Norway, 2005

[6] 上田達也，安倍広多，石橋勇人，松浦敏雄 "P2P手法によるインターネットノードの階層的クラスタリング"，情報処理学会論文誌，Vol. 47, No. 4, pp. 1063-1076, 2006.

[7] 堀内英斗，若宮直紀，村田正幸，"多人数参加型P2Pテレビ会議システムにおける論理網構築手法の提案と評価，" 電子情報通信学会技術研究報告（IN2006-35), pp. 1-6, July 2006.