Vol.2011-DPS-146 No.26
Vol.2011-CSEC-52 No.26
2011/3/11

パソコン 150 台を使用した 低遅延 P2P 配信木モデルのシミュレーション実験

高原　誠[†]　田上敦士[††]　阿野茂浩[††]　鈴木健二[†]

近年，為替の更新情報や RSS など，小容量の情報を多数のユーザに向けて迅速に配信するサービスの需要が高まっている．これらの情報の配信において，P2P 方式が負荷集中による遅延の影響を受けないため，クライアント・サーバ方式に対して有利な場合がある．そのため， P2P 方式の適切なトポロジが検討されているが，ネットワークを模擬したシミュレーションによる評価しかなされておらず，実ネットワーク上での評価がない．

本稿では，筆者らが検討してきた配信木モデルから理論的に導出した最適トポロジと 150 台の PC が接続された実ネットワーク上の実測より導出した最適トポロジとを比較する．本比較により，本モデルを用いた理論的な導出手法の有効性を示すと共に，実装化に向けて考察したので報告する．

Experimentation of P2P Low Latency Multicast Tree with 150 Receiver Nodes

Makoto Takahara[†]　Atsushi Tagami[††]　Shigehiro Ano[††]
and　Kenji Suzuki[†]

Recently, the request for the realization of multicast services over the Internet becomes very strong. And much of the discussion has been organized by comparing the Client-Server and P2P model. We have found the merits of P2P method for the time sensitive short message distribution with the comparison to the Client-Server method in our early paper, where it can be possible to complete the data distribution to all recipients within the short period in addition to avoid the unnecessary overhead and link congestion at the server side. In order to assure this result, it is better to organize experimentation at the actual situation using many terminals and networks.

This paper compared the theoretical optimum topology obtained by the theoretical P2P model and the practical optimum topology obtained by the actual experimentation over the network with 150 receiving terminals. Through this evaluation, we have shown the validity of our P2P model and pragmatic way of parameters setting to the real implementation.

1. はじめに

近年，Web アプリケーションに代表されるように，インターネット上で提供されるサービスの多くはクライアント・サーバモデルで実現している．しかしながら，ユーザの増加・サービスの多様化に伴い，サーバの処理能力や回線の輻輳が課題となっている．また，通信のユビキタス化に伴い，光ファイバーや CATV 網を使用した広帯域通信，高速モバイル通信も発展してきた．また，パソコンや携帯端末機器自体も性能向上しており，端末が単に情報を受信する端末ではなく，その情報を新たに配信する能動的な端末として動作できる状況にある．このため，情報配信はクライアント・サーバ方式だけではなく，単に情報を受信する機能をもっていたクライアントが，受信情報を他の端末に配信するような P2P 通信も，極めて現実味を帯びた通信であることが分かる．とりわけ，災害情報や RSS などタイムクリティカルな通信では，P2P 通信の有効性が徐々に明らかになってきているのが現状である．

筆者等はこれまで，災害情報や RSS などの小容量データを一定数の受信者にできるだけ早く配信できる手法を模索してきた．このため，クライアント・サーバ方式や P2P 方式を比較検討したが，小容量のデータをすべてのユーザに配信完了する時間（総情報配信時間）を比較すると P2P 方式がクライアント・サーバ方式より有利であるというシミュレーション結果を得た．この結果は，従来の P2P モデルを用いた配信方式である ALM（Application Level Multicast）[1][2][3][4][5]が大容量データを扱うサーバやサーバに接続する回線の負荷分散を狙ったものであったのに対して，P2P の負荷分散特性を生かした上で，小容量データ配信の場合には，更に総情報配信時間の短縮にもつながるという重要な結果を示している[6][7][8]．これまで，情報配信木のモデル化と数値シミュレーションを継続し，最適トポロジを調査してきたが，ここで，シミュレーションを一歩進め，実際のネットワークで評価することが重要になってきている．

本稿では，P2P 配信方式のモデル化の妥当性を示すため，モデルによって導出される理論的な最適トポロジと，150 台のノードを用いた実ネットワークを用いた実測により導出した最適トポロジとの比較，およびその時の総情報配信時間を評価した．次いで，理論値を導出するためのパラメタ設定手法について検討したので，以下にその概要を報告する．

[†] 電気通信大学大学院
The University of Electro-Communications
[††]（株)KDDI 研究所
KDDI R&D Laboratories Inc.

Vol.2011-DPS-146 No.26
Vol.2011-CSEC-52 No.26
2011/3/11

2. 低遅延 P2P 配信木

本章では，まず想定するデータ配信について述べ，すべてのユーザに配信完了する時間（総情報配信時間）を定義する．次いで，総情報配信時間を最小とする配信木トポロジの理論的な導出手順ならびに実ネットワークを用いた実測手順を示す．

2.1 P2P 配信木を用いた小容量データ配信

1 つの配信サーバ（配信ノード）から多数の受信ユーザ（受信ノード）へ，10KB から 100KB の小容量データを配信するアプリケーションを想定する．配信ノードは予め，IP アドレスなどのすべての受信ユーザの情報を保持する．また，配信ノードを根とし，受信ノードを節点もしくは葉とする木構造ネットワーク（配信木）を構築する．配信木の根である配信ノードから葉に向けて情報伝達し，すべてのユーザへの情報配信を実現する．この際，節点となる受信ノードを中継ノード，中継ノードに接続されている子ノードの数を分木数と呼ぶ．配信木を用いた配信手法は，効率良くすべてのノードに対して配信することができ，P2P 配信によく利用される．また，災害情報や RSS など，本稿が想定するデータの配信頻度は低いため，ノード間は常時コネクションを維持せず，データ転送時にコネクションを生成し，転送完了後切断するものとする．

本アプリケーションにおいて，配信ノードが配信を開始してから，すべての受信ノードに配信完了するまでの時間を総情報配信時間と定義する．この時，配信木の最適トポロジとは，総情報配信時間を最小とする配信木トポロジである．システム実装において，各配信木トポロジにおける総情報配信時間の測定を行うことは，負荷が高く現実的ではない．このため，配信前の限られた情報により，事前に配信木の最適トポロジを理論的に導出することが重要となる．次節では配信木モデルを用いて配信木の最適トポロジを理論的に導出する手順と，各種配信木トポロジの総情報配信時間を実測することにより，配信木の最適トポロジを実測する手順について述べる．

2.2 最適トポロジの理論的な導出手順

本節では，配信木のモデル化により総情報配信時間を数式化し，総情報配信時間を最小とする配信木の最適トポロジの理論的な導出手順について述べる．本配信木モデルは，リンク遅延 R と処理遅延 ΔT という 2 種類の遅延を用いる．リンク遅延 R はノード間のデータ送信時間，処理遅延 ΔT はあるノードに情報送信開始してから次ノードに情報送信開始するまでに要する時間と定義する．

まず，完全 m 分木モデルの例を図 1 に示す．完全 m 分木とはすべての節点が m 個の子ノードを持ち，かつ，すべての葉の深さが等しい配信木である．この場合，受信ノード数を n とすると，理論総情報配信時間$\hat{D}_c(m)$は次式で表される．

$$\hat{D}_c(m) = \big((m-1)\times\Delta T + R\big)\times h \tag{1}$$

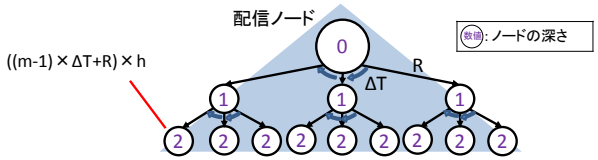

図 1　完全 m 分木モデル (m=3,h=2)

ただし，木の深さ h は以下の式を満たす．

$$n+1 = \sum_{i=0}^{h} m^i = \frac{m^{h+1}-1}{m-1}$$

完全 m 分木を構築するには，受信ノード数 n に制限がある．つまり，式(1)において木の深さ h は自然数の解を持つ必要がある．しかしながら実際には，受信ノード数 n は任意の値をとるため，前述の配信木モデルを一般化し，深さ$h-1$までは完全 m 分木，深さ$h-1$のノード（最終中継ノード）の分木数が$k\ (k \le m)$または$k-1$である配信木を考える必要がある．一般化した配信木モデルの例を図 2 に示す．この場合，理論総情報配信時間$\hat{D}(m)$は次式で表される．

$$\hat{D}(m) = \big((m-1)\times\Delta T + R\big)\times(h-1) + \big((k-1)\times\Delta T + R\big) + C \tag{2}$$

ここで，C は，深さ$h-1$までで最後に受信完了する最終中継ノードの分木数で決定され，$C \in \{0, -\Delta T\}$である．ただし，処理遅延ΔTはリンク遅延 R と比較して小さく，最悪ケースの総情報配信時間を短縮するため，以降$C = 0$と考える．

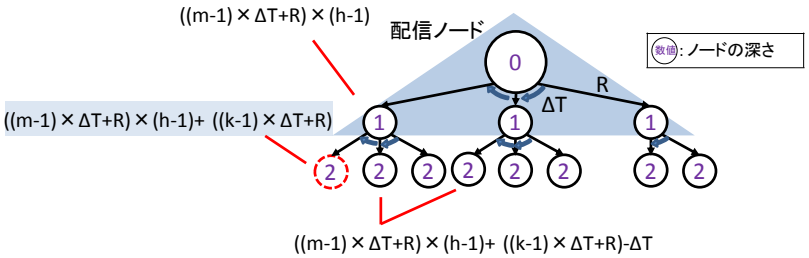

図 2　一般化した配信木モデル(m=3,h=2,k=3)

Vol.2011-DPS-146 No.26
Vol.2011-CSEC-52 No.26
2011/3/11

配信木の深さ h，最終中継ノードの分木数 k は，それぞれ以下で表される．

$$h = \lceil \log_m\{(m-1)(n+1)+1\} - 1 \rceil$$

$$k = \left\lfloor \frac{n - \sum_{i=1}^{h-1} m^i}{m^{h-1}} \right\rfloor$$

なお，⌈x⌉とはx以上の最小の自然数を表す．これにより，分木数 m が決定すれば配信木トポロジは決定することが分かる．すなわち，配信木の最適トポロジの導出は最適な分木数 m を導出することに等しい．この時，理論総情報配信時間 $\hat{D}(m)$ を最小とする理論最適分木数$\hat{m}$は，次式で表される．

$$\hat{m} = \arg\min_{m\in\{2,3,4\ldots\}} \hat{D}(m) \quad (3)$$

ただし，理論総情報配信時間$\hat{D}(m)$は複数のmにおいて同じ値をとることがあり．このため，理論最適分木数$\hat{m}$は複数存在する可能性がある．

2.3 最適トポロジの実ネットワークを用いた実測手順

理論値との比較のために，前節で述べた配信木モデルを実際に構築し，実測総情報配信時間 $\dot{D}(m)$を実測し，$\dot{D}(m)$を最小とする実測最適分木数$\dot{m}$を導出した．本計測に用いたネットワーク構成図を図 3, 使用したノードのスペックを表 1 に示す. 実測は，5 台のレイヤ 2 スイッチで構成される LAN 上に Gigabit Ethernet で接続した 1 台の配信ノードと 150 台の受信ノードを用いて行った．分木数 m は 2 から 12 ならびに，クライアント・サーバと等しくなることを意味する 150 とした．この時，実測総情報配信時間 $\dot{D}(m)$を最小とする分木数 m を実測最適分木数 $\dot{m}$ とした．配信データサイズは {10,20,30,40,50,60,80,100}KB とし，それぞれ総情報配信時間を 50 回実測した．

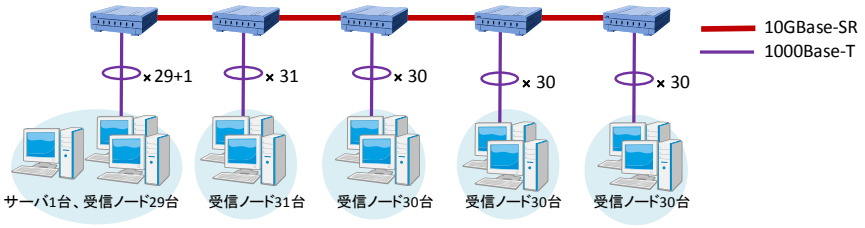

図 3　実測値の計測に用いたネットワーク構成図

表 1　ノードのスペック

CPU	Intel Core 2 Duo 2.83GHz
Memory	1GB
OS	Vine Linux 4.2 (kernel 2.6.16)

予め，配信ノードは配信先となるすべての受信ノードのアドレスをリストとして保持し，配信データと一緒に子ノードが担当するアドレスリストも送信する．アドレス数 L 個のリストと配信データを受信したノードは，次の配信アルゴリズムでデータを中継する．

[アルゴリズム]

1) L ≤ mの時，
   a) ノードは，L 個の子ノードにアドレスリストなしでデータ送信する．
2) L > mの時，
   a) L 個のアドレスから子ノードとなる m 個のアドレスを選択する．
   b) 2a)で選択されなかったアドレスを m グループに分割する．その m グループのうち p 個のグループは，(p − m)個のグループよりアドレス数が 1 つ多い．この p は次式となる．

$$p = (L - m) \bmod m$$

   ただし，x mod y とは x を y で割った時の剰余とする．
   c) 2b)で分割したアドレスリストをデータと共に 2a)で選出したノードに送信する．

図 4 に実測値の測定方法を示す．データ受信したノードは，UDP パケットを受信完了メッセージとして配信サーバに向けて送信し，一方配信ノードは，配信開始時刻と各受信完了メッセージを受信した時刻を計測する．計測結果より，150 個目の受信完了メッセージから配信開始時刻を引いた値の最頻値を実測総情報配信時間 $\dot{D}(m)$とした．本計測を 50 回行い，0.5ms 毎の分布を求め最頻値を求めた．これは，実測結果が小さく，平均値では異常値の影響を大きく受けるためである．この時，$\dot{D}(m)$が最小となる分木数を実測最適分木数$\dot{m}$とする．

Vol.2011-DPS-146 No.26
Vol.2011-CSEC-52 No.26
2011/3/11

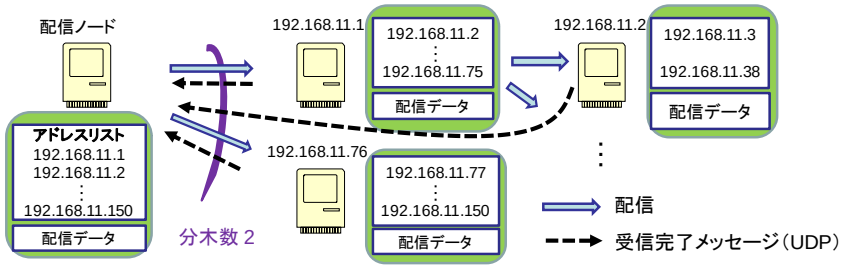

図 4　実測値の測定方法

3. 実験結果

本章では前章で示した理論最適分木数$\hat{m}$と，実測最適分木数$\dot{m}$を比較し，理論的な最適トポロジ導出手法の有用性を評価する．まず，理論最適分木数$\hat{m}$に必要なパラメタの設定手法を述べる．次いで実測結果から．理論最適分木数$\hat{m}$と，実測最適分木数$\dot{m}$の比較を行う．

3.1 理論的な最適トポロジ導出のためのパラメタ設定と総情報配信時間の補正

理論的な最適トポロジを得るためのパラメタであるリンク遅延 R と処理遅延 ΔT は，送信データサイズ毎の実測値を用いた．R は分木数 m=5 の測定開始から 1 番目の受信ノードからの受信完了メッセージを受信するまでの時間の 1/2，ΔT は m=150 の時の((最終受信完了メッセージの計測時刻)－(最初の受信完了メッセージの計測時間)) / 149 とした．それぞれ 50 回の計測値の平均値を用いた．

また，実測値 $\hat{D}(m)$ は計測手法の関係上，最後の受信ノードがデータを受信した時間に加えて，受信完了メッセージの伝送時間が含まれている．総情報配信時間の理論値$\hat{D}(m)$と実測値の$\hat{D}(m)$意味を等しくするため，理論総情報配信時間 $\hat{D}(m)$を以下のように補正する．

$$\hat{D}'(m) = \hat{D}(m) + R \qquad (4)$$

3.2 受信ノード数 150 の理論値と実測値の比較結果

受信ノード数 150 の時の，理論最適分木数$\hat{m}$，実測最適分木数 $\dot{m}$とデータサイズの関係を表 2 に示す．理論値$\hat{m}$と実測値$\dot{m}$は共に複数解を持つが，理論値$\hat{m}$が実測値$\dot{m}$にほぼ含まれている事が分かる．これより，2.2 節で述べた理論的な最適トポロジ導出手法が妥当であり，リンク遅延 R と処理遅延 ΔT により，総情報配信時間を最小とする最適トポロジが導出可能であると言える．

さらに，理論最適総情報配信時間 $\hat{D}'(\hat{m})$と実測最適総情報配信時間 $\dot{D}(\dot{m})$の関係を図 5 に示す．図より，誤差は最大 0.6ms 程度と小さく，総情報配信時間も高い精度で理論的に算出できていることが分かる．

表 2　各送信データサイズの最適分木数（受信ノード数 150）

データサイズ[KB]	理論$\hat{m}$	実測$\dot{m}$
10	5,6	5~11
20	5,6	5~8
30	5,6	5,6
40	5,6	5,6
50	5,6	5,6
60	5,6	5,6
80	5,6	5,6
100	5,6	3,5

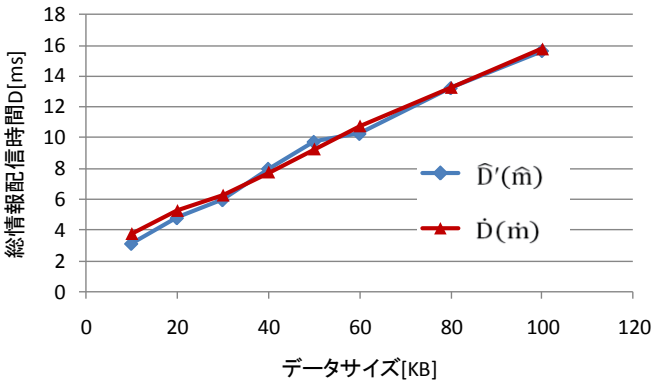

図 5　データサイズと総情報配信時間の理論値$\hat{D}'(\hat{m})$，実測値$\dot{D}(\dot{m})$の関係
（受信ノード数 150 の場合）

3.3 受信ノード数を変化させた際の理論値と実測値の比較結果

前節では，受信ノード数 150 として実験を行い，理論値と実測値の比較を行った．本節では，受信ノード数 n={60, 90, 120}の場合について追加で実測し評価を行う．送信データサイズは 10KB, 50KB, 100KB とし，2.3 節と同様に 50 回の計測から実測総情報配信時間 $\dot{D}(m)$を実測し，実測最適分木数$\dot{m}$を導出した．実測最適分木数$\dot{m}$と理論最

Vol.2011-DPS-146 No.26
Vol.2011-CSEC-52 No.26
2011/3/11

適分木数$\hat{m}$の比較を表 3 に示す．表より，ノード数が 150 以外の場合においても理論最適分木数$\hat{m}$が実測最適分木数$\dot{m}$にほぼ含まれており，受信ノード数 150 以外の場合においても最適トポロジの導出手法が有効であると言える．

さらに，理論最適総情報配信時間 $\hat{D}'(\hat{m})$と実測最適総情報配信時間 $\dot{D}(\dot{m})$の関係を図 6 に示す. 図より送信データサイズ 100KB において，最適総情報配信時間の差が大きくなっている事が分かる．しかしながら，その傾向はほぼ等しく，配信木モデルにより総情報配信時間 $\hat{D}'(m)$が十分な精度で表されていると言える．

表 3　各受信ノード数による最適分木数

ノード数	10KB		50KB		100KB	
	理論 $\hat{m}$	実測 $\dot{m}$	理論 $\hat{m}$	実測 $\dot{m}$	理論 $\hat{m}$	実測 $\dot{m}$
60	4	4~12	4	4	4	4
90	5	5	3	5	3	3,5
120	5	3,5~9,11	3	5	3	3,5
150	5,6	5~11	5,6	5,6	5,6	3,5

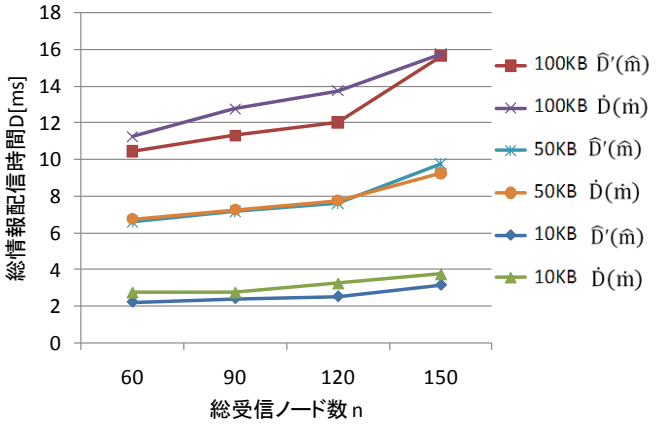

図 6　受信ノード数と総情報配信時間の理論値$\hat{D}'(\hat{m})$，実測値$\dot{D}(\dot{m})$の関係

4. 考察

4.1 実装に向けたパラメタの決定

3.2 節において，リンク遅延 R および処理遅延 ΔT が分かれば，配信木の最適トポロジと総情報配信時間を理論的に導出できることを示した．しかしながら，3.1 節で示したパラメタ（リンク遅延 R および処理遅延 ΔT）の計測方法は，取扱うデータサイズ毎に配信ノードが全ユーザに対して該当するデータサイズを事前に 50 回直接配信し，計測した遅延の平均値をとる方法である．実際にサービス提供のために実装する時，受信ノード数が大きい場合や配信時期が分からない場合，本手法はサーバへの負荷となり現実的でない. このため実装に向け, 前回の情報配信時の測定結果を用いて，最適トポロジに収束させるような手順が必要であると考える．本節ではその手順について考察する．

パラメタのうち，処理遅延 ΔT は受信ノードの性能に依存するため，事前に計測した値を用いることが可能である. 一方, リンク遅延 R はネットワークに依存するため，実際に受信ノードと計測する必要がある．以下に情報配信と同時に R を計測し分木数 m を導出する手順を示す．R の計測は 3.1 節で述べたように受信完了メッセージを用いるものとする．

1. $m = 2$ として配信木を構築し R を計測する．
2. 計測した R と事前に求めていた ΔT を用いて，最適な分木数を導出する．
3. 2.で導出した分木数から配信木を構築し R を計測する．以後 2 から繰り返す．

表 4 に 2.3 節での送信データサイズ 50KB の実測値を用いて，上記手順により最適トポロジを導出した結果を示す．表より，5 回の配信結果を用いることにより理論最適分木数$\hat{m}$に収束可能であることが分かる．これより，前回の情報配信時の測定結果を用いて，最適トポロジに収束させるような手順が実装可能であると言える．

表 4　試行回数と最適トポロジの変化

試行回数	受信ノード数				
	30	60	90	120	150
1	3	3	3	3	3
2	3	3	3	3	3
3	3	4	3	3	5,6
4	3	3	3	3	3
5	3	3	3	3	3
6	3	4	3	3	5,6
7	3	4	3	3	5,6
8	3	4	3	3	5,6
9	3	4	3	3	5,6
10	3	4	3	3	5,6
理論最適分木数$\hat{m}$	3	4	3	3	5,6

Vol.2011-DPS-146 No.26
Vol.2011-CSEC-52 No.26
2011/3/11

4.2 広域網への適用

3 章では LAN に接続された 150 台の PC を用いて評価した．しかし，対象とする情報配信サービスは，インターネットなどの広域網でのサービスを想定している．このため，広域網での理論最適分木数$\hat{m}$について考察する．まず，総情報配信時間$\hat{D}(m)$の定義である式(2)の両辺を ΔT で割ると次式となる．

$$\hat{D}(m)/\Delta T = (m - 1 + R/\Delta T) \times (h - 1) + (k - 1 + R/\Delta T) \quad (5)$$

ここで, 木の深さ h および k は m, n によって決定される値である. すなわち, $\hat{D}(m)/\Delta T$は受信ノード数 n とリンク遅延と処理遅延の比$R/\Delta T$により決定される．ここで，処理遅延 ΔT は分木数 m に依存しないため，式(3)は次式と等価である．

$$\hat{m} = \arg\min_{m \in \{2,3,4\ldots\}} \hat{D}(m)/\Delta T$$

すなわち，理論最適分木数$\hat{m}$は受信ノード数 n，リンク遅延と処理遅延の比$R/\Delta T$により決定される．図 7 に本パラメタと理論最適分木数$\hat{m}$の関係を示す．ここで，処理遅延 ΔT は配信ノードの処理能力に依存し，リンク遅延 R はネットワークに依存する．このため，広域網に適用した場合，リンク遅延 R のみが大きくなり，リンク遅延と処理遅延の比$R/\Delta T$は大きくなる．図 7 よりリンク遅延と処理遅延の比$R/\Delta T$が大きくなるに従い，理論最適分木数$\hat{m}$は大きくなる．ここで，最適分木数 $\hat{m}$ が受信ノード数 n と一致する時，配信木の最適トポロジはすべての受信ノートが配信ノードに直接接続している．すなわちクライアント・サーバ方式が P2P 方式より迅速に配信できることを意味する．これは，P2P 方式は受信ノード数 n が大きい，またはリンク遅延と処理遅延の比$R/\Delta T$が小さい場合に有用であることを意味する．

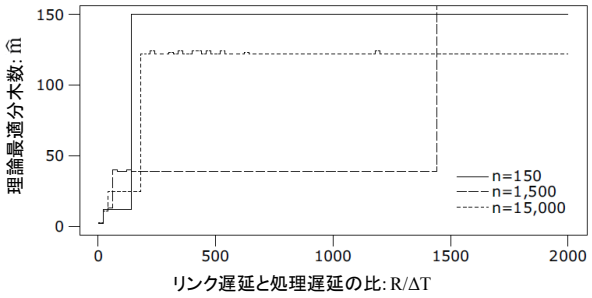

図 7　パラメタ値と理論最適分木数の関係

5. おわりに

本稿では，筆者らが提案した P2P 配信木モデルの実ネットワークでの妥当性を示すため，150 台の PC を LAN に接続した環境での実測を行い，理論値と実測値の比較を行った．この結果，提案する配信木モデルの有効性を示した．これにより，リンク遅延と処理遅延の 2 つのパラメタを測定すれば，配信木の最適トポロジが導出できることが示せた．さらに，実際にシステム構築するために，2 つのパラメタの計測手順と，その時の最適トポロジへの収束の考察を行った．さらに，広域網の適応を見据えた理論的な最適トポロジについても考察した．これらにより，提案手法の実装についての道筋をつけた．

今後は，ノードの性能，ネットワークからの離脱，ネットワークの近接性を考慮した総情報配信時間の評価を進めていきたい．また，計測による負荷と性能のトレードオフについて調査したい．

参考文献

1) Hosseini, M., Ahmed, D. T., Shirmohammadi, S. and Georganas, N. D.: A Survey of Application-Layer Multicast Protocols, *IEEE Communications Surveys & Tutorials*, Vol.9, No3, pp.58-74 (2007).

2) Pendarakis, D., Shi, S., Verma, D. and Waldvogel, M.: ALMI: An Application Level Multicast Infrastructure, *3rd Usenix Symp. Int'l. Tech. and Sys*., pp. 49–60 (2001).

3) Chu, Y., Rao, S. G. and Zhang, H.: A Case for End System Multicast, *IEEE Journal on Selected Areas in Communications*, Vol. 20, No. 8, pp. 1456–1471 (2002).

4) Ratnasamy, S., Handley, M., Karp, R. and Shenker, S.: Application-Level Multicast Using Content-Addressable Networks, *3rd Int'l. Workshop Net. Gr. Commun*., pp.14–29 (2001).

5) Castro, M., Druschel, P., Kermarrec, A.-M. and Rowstron, A.: SCRIBE: A large-scale and decentralized application-level multicast infrastructure, *IEEE Journal on Selected Areas in Communications*, Vol.20 No.8, pp.1489-1499 (2002).

6) 高原　誠，鈴木健二，田上敦士，阿野茂浩: P2P プラットフォームによる更新情報の低遅延配信方式の提案，マルチメディア・分散・協調とモバイル(DICOMO2009)シンポジウム，5E-3, 1047-1054 (2009).

7) 高原　誠，鈴木健二，田上敦士，阿野茂浩: 偏りのある配信木を用いた低遅延 P2P 情報通知方式の解析，情報処理学会研究報告，マルチメディア通信と分散処理研究会報告，Vol.2010-DPS-142 No.18, Vol.2010- CSEC-48 No.18 (2010).

8) 高原　誠，田上敦士，阿野茂浩，鈴木健二: 小容量データの同報配信のための効率的な P2P 配信方式の提案と解析，情報処理学会論文誌，Vol.52，No.2，(2011) (掲載予定).