C-008

P2P システムの耐故障性に関する検討
On the Fault Tolerance of P2P System

小澤 一平 † 寺田 夢人 † 新井 雅之 † 福本 聡 †
Ippei Ozawa Yumeto Terada Masayuki Arai Satoshi Fukumoto

1. はじめに

Chord や Symphony [1] に代表される構造化 P2P(Peer-to-Peer) ネットワークは，分散ハッシュテーブル（DHT, Distributed Hash Table）を用いることでルーティングの効率化を実現している．その高い性能と拡張性から，ノード間の接続形態，距離に関する定義，キー探索方法などに関する種々の研究が報告されている．

最近では，構造化 P2P ネットワークの持つ耐故障性に注目して，データの分散配置に木構造を導入して故障の影響を回避する方式 [2] などが提案されている．また，2 レベルのキャッシュを導入して性能と信頼性を向上させ，故障に対する弾力性を耐故障性の評価尺度として議論したものもある [3]．しかしながら，最も代表的な構造化 P2P ネットワークである Symphony そのものが持つ耐故障性についての定量的な特性評価については，これまでに必ずしも明確に議論されていない．

本論文では，Symphony のオブジェクトとハッシュテーブルの複製を可能な限り積極的に利用する場合と，深さ優先探索を許容する場合の，耐故障性について定量的に考察する．

2. Symphony の耐故障性

2.1 安定化プロトコル

P2P ネットワークではノードの参加・離脱に対する一貫性維持がシステム全体で定期的に実行される．故障したノードが管理する部分ハッシュ空間の受け渡し手続きは，通常時のノードの離脱に対する一貫性管理のための安定化プロトコルと区別することなく扱われる．

ノードの離脱と故障に対する回復処理を実現するためには，各ノードが $r(\geq 2)$ 個の連続する後任ノードに割り付けられたハッシュ値のリストを保持する必要がある．これを “後任ノードリスト” と呼ぶ．後任ノードリストは，故障ノードの管理するハッシュ空間を吸収するときの情報として使用される．

2.2 オブジェクトの複製に基づく耐故障性

Manku ら [1] は，Symphony において，各ノードが自身の持つオブジェクトの f 個の複製を，f 個の連続する後任ノードに配置する手法について考察している．

それぞれのノードは，f 個の後任ノードすべてに対する接続を維持する．この複製配置は，f 個のノードの同時故障を許容することができる．新たなキーの値がリング上に挿入されるとき，ノードは一貫性維持のために f 個の後任ノードにその情報を伝播させる．すべての探索は f 個の後任ノードが同時に故障するまで継続することができる．

2.3 ハッシュテーブルの複製配置に基づく耐故障性

本研究では，各ノードが後任ノードリストに含まれる r 個のノードのハッシュテーブルすべてを複製保持する場合の耐故障性について考察する．各ノードは，自身を後任ノードリストに含む r 個の先任ノードと r 個の後任ノードとで，互いのハッシュテーブルを共有することになる．このうち，r 個の先任ノードとの共有テーブル情報が時計回りのルーティングにおいて有効となる．このことは，そのノードがルーティングの中間ノードであるとき，後任ノードリストに含まれるノードが持つハッシュテーブルのリンクすべてを追加的なロングリンクとして使用できることを意味する．

2.4 深さ優先探索による耐故障性

本研究では，さらに，探索の中間ノードにおいて深さ優先探索を許容する場合の耐故障性を考察する．通常のルーティングの過程において，あるノードに対しすべてのリンクが切断されていると探索は失敗とするが，深さ優先探索ではバックトラックすることで，リング上の切断された空間の迂回が可能となる．始点ノードは木構造の根に対応し，目的ノードに到達するか全ノードにアクセスするまで，貪欲アルゴリズムによって再帰的に探索する．探索の過程において，リンク先のノードがすべて故障しているときは，そのノード ID を “禁止リスト” に追加して，バックトラックする．

しかしながら，P2P ネットワークおいては多数のノードが参加し，数本から数百本の探索パスが存在しうるため，探索に膨大な時間を要する可能性がある．また，ホップ数に比例して禁止リストの情報量が累積していく問題がある．そこで，探索が成功しそうにないときのバックトラックを無効にする為の “打切り故障率” を導入し，探索のための期待ホップ数の増加に制限を加える．すなわち，ネットワーク全体の推定故障率が一定の閾値を超えるとバックトラックを中止し，通常の探索に切り替える．推定故障率は，1 ホップ毎にリンク先ノードで得られる故障情報から累積的に計算する．また，各ノードにおいてバックトラックに使用できるリンクの本数に上限を与える．通常のロングリンク k 本中利用出来る本数を k' 本，

†首都大学東京 大学院 システムデザイン研究科, Graduate School of System Design, Tokyo Metropolitan University

ショートリンク r 本中利用出来る本数を r' 本とする．

3. シミュレーション評価

本研究では評価の前提として，ノード故障の発生から安定化プロトコルによる定期的な一貫性保守が完了するまでの過渡的な状態について考える．これは，この期間に最も信頼性が低下し，ネットワークのルーティング機能に障害が発生しやすいためである．故障モデルとしてノードのフェイルストップ故障を仮定する．故障は一定の確率でランダムに発生するものとする．評価尺度として探索成功確率を用いる．

図 1，図 2 は，ノードの故障率に対する探索成功確率を表している．オブジェクトの複製数とハッシュテーブルの複製数以外のパラメータは，ノード数 $N = 2^{12}$，ロングリンクの本数 $k = 4$ で共通である．

図 1 の $(r = 0, f = 0)$, $(r = 0, f = 4)$, $(r = 4, f = 0)$, $(r = 4, f = 4)$ は，それぞれ，複製を両方とも配置しない場合，複製配置がオブジェクトだけの場合，ハッシュテーブルだけの場合，オブジェクトとハッシュテーブルの両方を配置する場合に対応している．オブジェクトの複製だけを配置する場合に着目すると，故障率 0.2 程度までなら複製数 $f = 4$ で十分に信頼性を維持できることが解る．ハッシュテーブルの複製だけを配置する場合は，目的ノードの故障率で探索成功確率の上限が与えられる．オブジェクトとハッシュテーブルの両方の複製を配置する場合，すなわちトータルでは，より高い故障率に対する信頼性の維持が可能となっている．

図 2 では，バックトラックしない場合と，それぞれ打切り故障率が $c = 0.3$，$c = 0.4$，$c = 0.5$，$c = 1.0$ の場合の探索成功確率を示している．バックトラックする場合はリンクの使用本数の上限値を $k' = 2$，$r' = 2$ で共通としている．また，オブジェクトの複製数 $f = 4$ で目的ノードの冗長化を行っている．$c = 1.0$ は上限がなく，全探索するということを意味する．打切り故障率を引き上げることで探索成功確率は改善されるがホップ数も増大するため，ノード数やレプリカ数などに応じた適切な打切り故障率を選択する必要がある．深さ優先探索によって，平均探索時間の増加を調整しながら，探索成功確率を向上させることができることが解る．

4. まとめ

本研究では，代表的な構造化 P2P システムの一つである Symphny を取り上げ，目的ノードの持つオブジェクトの複製配置に加えて，ハッシュテーブルの複製を連続する複数ノードへ配置する場合の耐障害性について議論した．また，深さ優先探索によってネットワークの切断を迂回する探索方法を検討した．

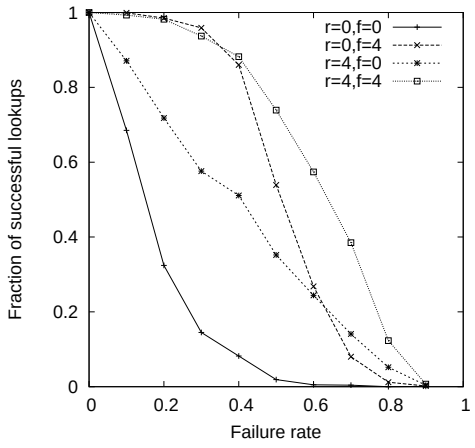

図 1: 複製を配置した場合の探索成功確率

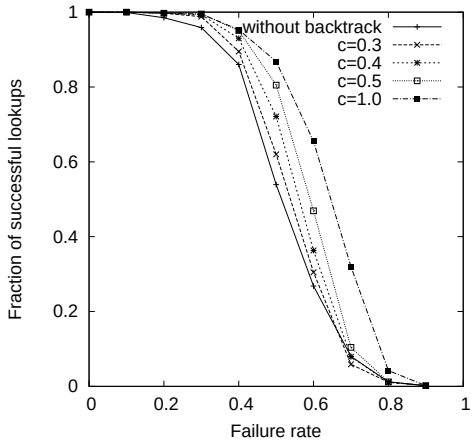

図 2: 深さ優先探索した場合の探索成功確率

参考文献

[1] G. S. Manku, M. Bawa, P. Raghavan, “Symphony: Distributed Hashing in a Small World,” *Proc. USENIX symp. Internet Technologies and Systems,* 2003.

[2] P. Garbacki, D.H.J. Epema, and M. v. Steen, “The Design and Evaluation of a Self-Organizing Superpeer Network,” *IEEE TRANSACTIONS ON COMPUTERS,* Vol. 59, No. 3, pp. 317-331, Mar. 2010.

[3] Z. Li, G. Xie, K. Hwang, and Z. Li, “Churn-Resilient Protocol for Massive Data Dissemination in P2P Networks,” *IEEE Transactions on Parallel and Distributed Systems,* Vol. 22, No. 8, Aug. 2011.