様式 C-19

# 科学研究費補助金研究成果報告書

平成２１年５月１９日現在

研究種目：若手研究（Ｂ）
研究期間：2006～2008
課題番号：１８７０００６９
研究課題名（和文）P2P ノード集合上で分散プロセスを実行する P2P 基盤ソフトウェアの開発
研究課題名（英文）Development of a P2P platform which enables executing distributed processes on P2P nodes

研究代表者
安倍　広多（ABE KOTA）
大阪市立大学・大学院創造都市研究科・准教授
研究者番号：４０２９１６０３

研究成果の概要：P2P システムでは，参加しているコンピュータ（ピア）が予告なく離脱（もしくは故障）するため，通常の P2P アプリケーションでは耐故障性を考えた設計・実装を行う必要があり，実装は容易ではなかった．本研究では複数のピアを仮想ピアとしてグループ化し，各ピアで一貫性を保ちながら同一のプログラムを動作させる方式を提案・実装した．仮想ピアは実質的に離脱しないピアと見なせるため，これを利用することで P2P アプリケーションを容易に実現できる．

交付額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2006 年度 | 800,000 | 0 | 800,000 |
| 2007 年度 | 1,100,000 | 0 | 1,100,000 |
| 2008 年度 | 600,000 | 180,000 | 780,000 |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 2,500,000 | 180,000 | 2,680,000 |

研究分野：分散システム，基盤ソフトウェア
科研費の分科・細目：情報学，計算機システム・ネットワーク
キーワード：分散システム，P2P，分散プロセス，基盤ソフトウェア，フォールトトレランス，Paxos 合意アルゴリズム

## １．研究開始当初の背景

ピアツーピア(P2P)方式によるネットワークサービスが注目されている．P2P 方式は従来のサーバクライアント方式と比較してスケーラビリティ，信頼性，管理コストなどの点で優れている．P2Pはファイル共有サービスなどで利用されてきたが，最近は並列計算のためのコンピューティングプラットフォームとして利用しようという動きがある．本研究ではその動きをさらに進め，不安定なP2Pノード群上で安定的にプロセスを実行する基盤ソフトウェア(P2POS:仮称)の設計，実装，評価を行う．P2POSが動作するノード集合全体が様々なソフトウェアを実行する基盤となる．

## ２．研究の目的

本研究では，P2Pネットワーク上で安定してソフトウェアを動作させるための方式を検討・実装する．具体的には以下の機能の実現を目指す．

- プロセスの実行状態保存，復元（チェックポインティング）（他のノードでの実行状態復元も可能）
- 実行中のプロセスをP2POSが動作する他のノードに転送し実行継続（プロセス移送）
- 同一プロセスの複数のノードでの実行（冗長プロセス）

P2POSで解決すべき主要な技術的課題を以下に挙げる．

- ノードでP2Pプロセスを安全に実行するためのセキュリティ機構
- 悪意あるノードからのシステム全体の保護
- 異種混合環境での動作：チェックポインティングやプロセス移送時に保存するデータには機種依存性が含まれていてはならない
- 全体が見渡せない大規模な自律分散環境でのプロセススケジューリング
  - 停止・離脱したノードで実行していたプロセスの再開方法
  - 公平性の実現（特定のプロセスが多数の資源を占有しないように）
- 冗長プロセス構成では，論理的に同一なプロセスが複数同時に実行される．このような環境でどのように一貫性を保ってサービスを提供するか．

３．研究の方法

(1) 開発する P2POS（仮称，後に musasabi と命名）は，大阪大学が中心となって開発している PIAX（ユビキタスコンピューティングのためのエージェントベースの P2P プラットフォーム）を拡張することで実装することにした．PIAX は全て Java 言語により実装され，構造化オーバレイネットワークの 1 つである Skip Graph を提供している．また，PIAX 上でエージェントを実行する機構を備えている．musasabi 上のユーザプロセスは PIAX のエージェントを拡張して実現した．

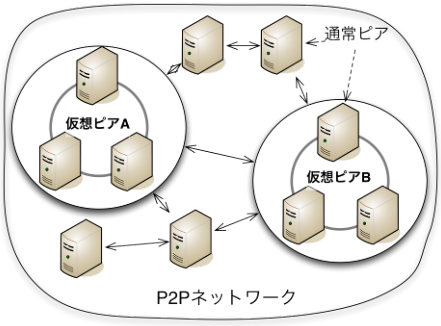


図 1 仮想ピアと通常ピア

(2) プロセス移送機能の実現
musasabi 上で，実行中のプロセスを異なる計算機上に転送し，実行を継続する機能（プロセス移送機能）を実現した．実装したプロセス移送機能では，実行中のプロセスを他のノードに実行コンテキストごと移動できる（強モビリティと呼ばれる）．通常の Java 言語では強モビリティはサポートされていないため，これを実現するために，Java 言語で Continuation(継続)を実現するライブラリである Apache Javaflow を用いている．Javaflow はバイトコード変換を用いて Continuation を実現している．実現したプロセス移送機構の実装は異種混合環境でも動作する．Windows 上で動作していた musasabi のプロセスを，MacOS X 上の musasabi にプロセス移送し，動作を継続させることが可能である．

(3) 同一プロセスを複数のノードで実行する方式（冗長プロセス）の実現：
同一プロセスを複数のノードで同時に実行することで，プロセスの耐故障性を向上する方式を実現した．提案方式では musasabi が動作する複数のノード（通常ピア）を**仮想ピア**としてグループ化し，その上で**仮想プロセス**を動作させる（図１，図２参照）．ノードの一部が離脱（もしくは故障）しても仮想プロセスの実行を継続できる．提案方式では各ピア上で同一のプロセスが同時に（冗長的に）一貫性を保ちながら動作する．この方式を実現するためには，各ノードが同一の入力を同一の順序で受信する（アトミックマルチキャストに相当）機構が必要である．musasabi では分散合意アルゴリズムの 1 つである Paxos を用いて実現した．また，仮想ピアを構成するノード数が減少しないように，仮想ピアを構成するノードが離脱すると，P2P ネットワーク上のピアを選び，離脱したピアと交替さ

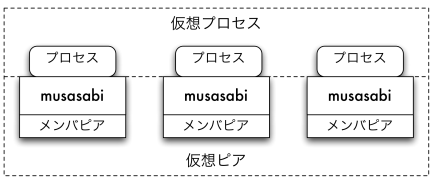


図 2 仮想ピアと仮想プロセス

せるようにした．新しく仮想ピアに参加したピアでは，他のピアで既に実行中のプロセスと同じ状態のプロセスを実行する必要があるが，このために(2)のプロセス移送機能を用いている．

(4) 仮想ピア（仮想プロセス)を単位としてP2P サービスを構築するためには，仮想ピアと通常のピア，あるいは仮想ピアと仮想ピア間で通信するための仕組みが必要である．仮想ピアは複数のノードから構成されるため，通信方法は単純ではない．仮想ピアが複数のノードで構成され，互いに一貫性を保ちながら動作していることを考慮した通信アルゴリズムを考案した．

(5) 各ノード上で musasabi 上のプロセスを安心して実行できるように，Java セキュリティモデルを利用するようにした．musasabi のプロセスはサンドボックス内で動作する．

## ４．研究成果

P2P システムに参加しているノード（ピア）は，予告なく離脱（あるいは故障）するが，そのようなピアを複数用いてプログラムを長期間継続して実行する方式（仮想ピア方式）を明らかにし，実際に実装した．仮想ピアを単位として P2P サービスを構築することで，従来の P2P アプリケーションでは不可欠だったリモートピアの障害対策（データの複製を複数のノードにさせる，バックアップポインタを維持するなど）を省力化できる．これにより，P2P アプリケーションを容易に実装できるようになる．また，ハイブリッド型 P2P システムにおける単一故障点である物理サーバを仮想ピアに置き換えることで，システムの耐故障性を向上させるといった応用も考えている．

仮想ピアに参加するピア（メンバピア）の数を増やすことで，仮想ピアが機能を失う可能性を下げることができる．図 3 はメンバピアの数を変化させながら，仮想ピアの稼働時間と仮想ピアの信頼度（故障せずに動いている確率）の関係をプロットしたグラフである（計算の前提：ピアの離脱間隔は指数分布で，１時間でピアの半分が離脱する．ピアが離脱してから交代ピアを参加させるまでの時間は 60 秒とする)．グラフから，仮想ピアはメンバピア数 7 程度で高い信頼性を得られることが分かる．

国内外での類似研究としては，Paxos 合意アルゴリズムを P2P システムで用いる例は存在する（P2P ネットワーク上で Key-Value Store を実現する Scalaris)．また，クラウドコンピューティングでの例としては Google の Chubby 分散ロックサービスが Paxos を使用している．しかし，本研究で提案している仮想プロセスのような，P2P システム上でプロセスの耐故障性を高めるものは筆者の知る限り見当たらない．

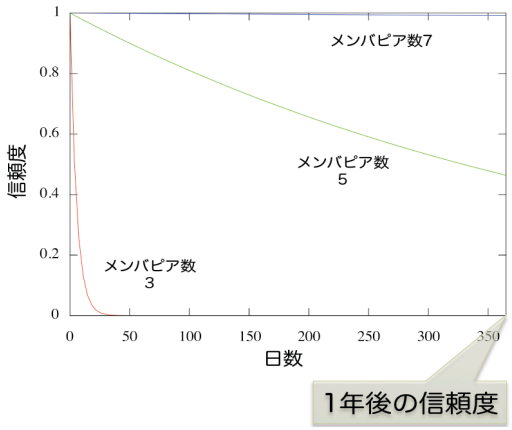


図 3 仮想ピアの信頼度

本研究の成果は，NICT(独立行政法人情報通信研究機構)の「高度通信・放送研究開発委託研究」として(株)アクタスソフトウェアと大阪市立大学が共同開発しているプロジェクト（平成 20 年度から 3 年間の予定）でも利用することになっているため，引き続き研究開発を行っていく．このプロジェクトでは，移動端末から位置情報が付与された動画ストリームを配信するシステムを P2P システムとして実現する．これを複数の仮想ピアを用いて実現する予定である．

musasabi のソースコードは安定次第オープンソースとして公開する予定である．

今後の課題としては，以下が挙げられる．

- 仮想ピアを構成するピア（メンバピア）の効率的な選択：メンバピア間では頻繁に通信する必要があるため，仮想ピアの性能を考慮するとメンバピアは互いにネットワーク的に近いことが望ましい．しかし，同一組織や同一 AS 内に複数のメンバピアが存在すると今度は耐故障性の面で問題がある．このため，丁度良いメンバピアを選択する方法を考案する必要がある．このためには，P4P 的なアプローチ，あるいは Vivaldi のようなネットワーク座標系を用いる方式が有望と考えている．
- 仮想ピア間通信の実装：現在，通常のピアと仮想ピア間の通信は実装しているが，仮想ピア間で通信する方式は検討は終わっているが実装が完了していない．これを実装する．

・仮想ピアから仮想ピアを動的に生成する機能の実装：１つの仮想ピアはスケーラビリティを得ることはできない．スケーラビリティを得るに複数の仮想ピアが必要であり，必要に応じて動的に仮想ピアから仮想ピアを生成する機能を検討している．これを実現する．
・実装の安定化：NICT のプロジェクトなどで musasabi の仮想ピアを用いた P2P システムを開発することで，実装の安定化を図る．
・仮想ピアを用いた P2P システムでのセキュリティモデルの検討：仮想ピアを用いた P2P システムは，従来の P2P システムとはモデルが異なるため，どのような攻撃が考えられ，どのような防衛手段があるのか考察する必要がある．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計３件）

① 鹿野将典，上田達也，安倍広多，石橋勇人，松浦敏雄，P2P基盤ソフトウェアmusasabiの仮想ピアにおける通信方式，情報処理学会研究報告 Vol.2009-DPS-139, No.2, pp.1–8，2009，査読無．
② 市村大資，安倍広多，松浦敏雄，石橋勇人，強モビリティを用いたChordアルゴリズム実装の試み，大阪市立大学大学院創造都市研究科都市情報学専攻電子紀要「情報学」：<http://ojs.info.gscc.osaka-cu.ac.jp/JI/>, (ページ番号なし), 2009，査読無．
③ 安倍広多，P2Pシステム上での安定したサービス提供基盤musasabi, 情報処理学会研究報告，Vol. 2009-IOT-4，pp.131–136，2009，査読無．


〔学会発表〕（計２件）

① 安倍広多，Paxos合意アルゴリズムを用いて仮想的に"落ちない"ピアを提供するP2P基盤ソフトウェアmusasabi，第3回広域センサネットワークとオーバレイネットワークに関するワークショップ，2009年5月9日，大阪大学．
② 安倍広多，P2P-based OS: musasabi，第2回広域センサネットワークとオーバレイネットワークに関するワークショップ，2008年11月1日，慶応義塾大学．


〔その他〕
ホームページ等
http://rabbit.media.osaka-cu.ac.jp/research/index.php/Musasabi

６．研究組織
(1)研究代表者

安倍　広多（ABE KOTA）
大阪市立大学・大学院創造都市研究科・准教授
研究者番号：４０２９１６０３


(2)研究分担者
なし

(3)連携研究者
なし