様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２１　日現在

機関番号：　１４３０３
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００１６９

研究課題名（和文）分散処理環境における動的最適化制御のための知的ネットワークシステムの研究

研究課題名（英文）Study on Intelligent Network System for Dynamic Optimization Control in Distributed Computing Environment

研究代表者

布目　淳（Nunome, Atsushi）

京都工芸繊維大学・情報工学・人間科学系・助教

研究者番号：６０３３５３２０

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：大規模な分散処理環境において性能を最大限に引き出すためには、ソフトウェア及びハードウェア両面からのチューニングが重要である。本研究では、様々な性能のノードが混在する非均質環境に加え、実行時に生じる実効性能差も考慮して、それらの分散処理環境で動的に最適化制御を行う方式を提案する。この制御に必要な情報は実行時に収集する必要があるため、従来はオーバヘッドの大きさが問題であった。提案方式は、ネットワーク装置でシステムの管理情報を適切に付加することによって、最適化制御に伴うオーバヘッドを抑制する。今後、より複雑な動的最適化制御を開発し管理情報の量が増加する場合にも対応可能な基盤技術の一つを確立した。

研究成果の概要（英文）：In a massively distributed computing environment, it is important to tune the system from both aspects of software and hardware to bring out potential performance to the maximum. In this study, we have proposed a scheme to perform dynamic optimization control for the heterogeneous environments where various types of node are mixed. It also considers differences in the effective performance which is occurred at run-time. Formerly, a high overhead is regarded as a problem because the information needed for such control must be collected at run-time. In our scheme, the overhead accompanied with the optimization control will be suppressed by appending the system management information onto ordinary packets by the network device. We have established one of a fundamental technology which is capable of applying even if the management information enlarges for more complex dynamic optimization control in the near future.

研究分野：　情報工学

キーワード：　計算機システム　ハイパフォーマンスコンピューティング　分散処理　動的負荷分散制御

１．研究開始当初の背景

(1) 分散処理環境において、ノード間の通信制御は重要な研究課題の１つである。1990 年代後半以降、機器の高性能化と低価格化により、システムの大規模化が特に進んでいる。こうした多数のノードの状態を把握し、タスクやデータを適切に配置することは、システム全体の性能を最大限に引き出すために極めて重要である。中でも、様々な種類のソフトウェアを同時に実行する場合や、異なる性能のノードから成る非均質環境においては、動的な割り付け制御が不可欠となる。このように、実行時に管理情報（リソースの利用状況、ノードやネットワークの負荷情報など）を収集することは、動的な負荷分散制御のみならず、分散ストレージに適切にデータを配置したり、フォールトトレランスを実現したりするなど、実行環境の変化に適応するために、様々な分野で必要な処理となっている。一方で、こうした処理はトラヒックの増大を招き、オーバヘッドの増加にも繋がるため、管理情報を低負荷かつ迅速に交換する技術の開発が急務であった。

(2) こうした技術的背景から、本研究代表者は大規模な実行環境を対象とした動的負荷分散制御方式に関する研究を行ってきた。

(3) まず、負荷情報を通常の情報伝達に用いるネットワークパケットに併合することで、多数のノード間で負荷情報を効率良く交換する方式を開発した。これにより、個々のノードが個別に情報交換を行うことでネットワークが混雑する問題を抑制できることを明らかにした。

(4) また、IP ネットワーク上で非 IP ネットワーク用の通信プロトコルを使用する際に生じるネットワークフレームの空き領域を、管理情報を交換するために有効利用する方式を開発した。近年の高速ネットワーク規格においては、大きな最大転送単位（MTU）によって転送効率を向上させることが主流であるが、従来の転送路を前提とした SCSI プロトコルをそのまま IP ネットワークに持ち込む iSCSI では、小さなパケットサイズでの情報交換が頻繁に行われている。このギャップを管理情報の交換に用いることで、別途パケットを生成することを回避することで、ネットワークトラヒックを抑制した。

(5) 本研究は上記(3)と(4)の研究から着想を得た発展的な研究である。

２．研究の目的

(1) 本研究は、本研究代表者がこれまで行ってきた研究を発展させ、より大規模なノード構成を有する環境において、より多様な応用ソフトウェアに対応できる分散処理環境を実現するための基盤技術を確立することを目的とする。

(2) 実行環境の動的な変化に対してシステムが自律的に適応する「動的最適化制御」を行う機構を開発する。応用ソフトウェアだけがこうした制御を担当する場合、実行時のオーバヘッドの大きさから実行性能に与える影響が大きい。そこで本研究では、ネットワークなどのハードウェアと協調することにより、低オーバヘッドかつ低遅延な動的最適化制御を実現する。

３．研究の方法

(1) 本研究では、分散処理環境を構成する各ノードの機器構成が異なっているだけでなく、実行時にかかる負荷の偏りのために、実効性能についても差が生じるものと仮定する。このような「ノードの動的な非均質性」が存在することを前提とし、タスクおよびデータを適切なノードへ移送するための基盤技術を開発する。

(2) 実行ノードやストレージノードの負荷状況をシステム内で共有するために、ノード間で交換する管理情報を決定する。併せて、ノード間で交換するネットワークパケットの構造を決定する。この管理情報を通常のノード間通信で用いるユニキャストフレームに併合することで、独立したネットワークトラヒックの発生を抑制する。

(3) 次に、管理情報の交換頻度を適正レベルに維持する方式を設計する。管理情報の正確さを保つためには十分に短い間隔で情報交換する必要があるが、過度に交換を行うことは、ネットワークトラヒックの増大を招き、システムの実効性能を低下させる。そのため、ノード状況の変化が少ない場合には情報交換の頻度を下げ、動的最適化制御のオーバヘッドを削減する方式を開発する。

(4) 幅広い種類の応用ソフトウェアの挙動に対応するため、応用ソフトウェアの動作をモニタし、適切なノードに対してデータおよび負荷を移送する方式を提案する。

(5) 小規模な評価用環境を構築し、動的最適化制御にかかる実際の処理時間を計測する。この結果から、シミュレーション時のパラメータ値を決定する。このパラメータ値を用いて、大規模な分散処理環境での性能をシミュレーションにより評価し、提案方式の有効性を検証する。

４．研究成果

(1) ストレージ上のデータをアクセス状況に応じて適切なノードに移送する自律分散ストレージシステムにおいて、ストレージノードの状況をノード間で交換するための管理情報を決定した。ここでは SAN (Storage Area

Network）環境下で広く用いられる iSCSI プロトコルを対象とした。各ストレージの状態を 16 バイトの固定長データに集約することで、ハードウェアで容易に処理できるようにした。

(2) ZFS ストレージに格納したサイズ 100MB のファイルに対するシーケンシャルアクセスを観測し、Read アクセスと Write アクセスを行った場合の iSCSI パケット長を計測した。その結果、Read アクセスにおいては 4,166 バイトのフレーム長、Write アクセス時においては実行パケットの 76%が 3,018 バイトのフレーム長であることが判明した。ギガビットイーサネットに代表される現在の高速ネットワーク規格では、通常のイーサネットの最大フレーム長を越えるジャンボフレームが一般的に用いられており、9,000 バイト以上のフレーム長が広く利用可能になっている。

(3) ストレージ情報をシステム内に伝搬させるため、ローカルノードが生成したストレージ情報には最高の優先順位を与える。一方で、遠隔のノードから受け取ったストレージ情報は伝搬遅延の影響で情報の鮮度が劣るため、転送優先順位を低く設定する。送信側ノードにおいては、まずローカルストレージに関する情報を元の iSCSI パケットに併合する。次に、イーサネットジャンボフレームにまだ追加する余地がある場合、遠隔ノードから収集したストレージ情報を併合する。この時、古い情報が伝搬しないようにするため、新しいタイムスタンプを持つストレージ情報を選択する。

(4) 図 1 は iSCSI パケットに対するストレージ情報の併合によるフレーム数の削減率（Frame Reduction Ratio）を示している。iSCSI リード処理においては、iSCSI イニシエータが送信するフレームの 49.8%、iSCSI ターゲットが送信するフレームの 36.2%をそれぞれ削減可能であることが明らかになった。同様に、iSCSI ライト処理においては、iSCSI イニシエータが送信するフレームの 16.7%、iSCSI ターゲットが送信するフレームの 49.8%をそれぞれ削減できることが分かった。

(5) システム内で管理情報を交換する時間間隔を各ノードが自律的に調整する方式を設計した。これにより、低負荷のノードにおいては管理情報の送信パケットを削減し、負荷変動の大きいノードでは管理情報の送信頻度を高めることが可能になった。2 種類の応用ソフトウェアを用いて、240 台のストレージノードからなる環境でシミュレーションによる評価を行った結果、より大規模な環境に対してもネットワークトラヒックの増加を伴わずに提案方式が適用できることが明らかになった。

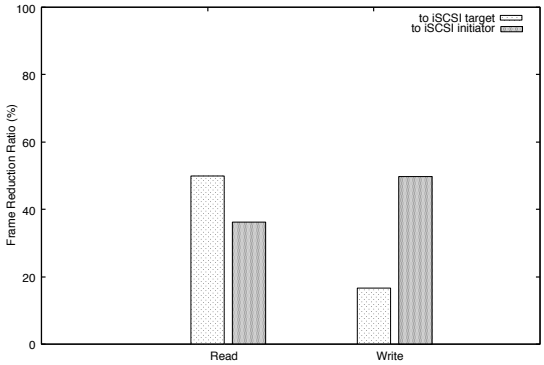

図 1 iSCSI パケットに対するストレージ情報の併合によるフレーム数削減率

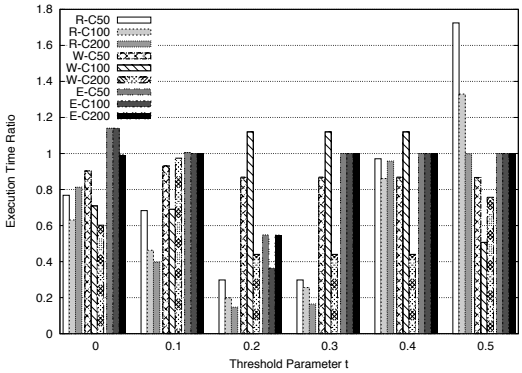

図 2 移送閾値 t を変化させた場合の実行時間比

(6) 上記(5)までの成果を基に、ストレージ上のデータブロックをその利用パターンに応じて適切なストレージノードに移送する方式を開発した。システムで実行される応用ソフトウェアによってストレージへの読み書き比率が異なるため、提案方式ではデータブロックに対する読み書き頻度に応じてストレージノードの読み書き性能に重み付けを行い、その重み付け後の性能を基にストレージノードを序列化する。これを「動的移送方式」とし、重み付けを行わない従来方式（静的移送方式）と比較した。動的移送方式では、例えば、読み込みアクセスが頻繁に行われるデータブロックは読み込み性能が優れたストレージノードへ移送することでアクセス性能を改善する。

(7) シミュレーションによる評価の結果、提案方式は特に読取り集約型アプリケーションにおいて、実行時間を短縮できることがわかった。図 2 に静的移送方式での実行時間を 1 とした場合の動的移送方式の実行時間比（Execution Time Ratio）を示す。読み取り集約型アプリケーションを R、書き込み集約型アプリケーションを W、読み書き均衡型アプリケーション E と表記している。クライアントノード数 C を 50 台、100 台、200 台と変化させて評価したところ、移送先ノードの候補とするための移送閾値（Threshold Parameter）t を適切に設定すれば、200 台構成の場

合に特に優れた性能を示し、従来方式よりも実行時間を約 80%削減できた。このことから、提案方式が大規模システムに十分に対応可能であることが示されたと言える。

(8) 近年、大規模な分散処理環境を安価に構築できるようになってきた一方で、その管理に要する作業量は上昇する傾向にある。作業量の削減には自律的な動的最適化制御が有効であるが、これまではオーバヘッドの大きさが問題であった。本研究によって、ネットワークにかかるオーバヘッドを抑制することが可能で、自律的な動的最適化制御が性能向上に有効であることが確認できた。これは、適切にオーバヘッドを抑制することによって、複雑な動的最適化制御を行ったとしても性能向上を引き出せることを示した点で、大きな意義があったと言える。本研究の成果により、動的最適化制御に必要な管理情報を低コストで交換できることが明らかになり、将来的にこれまで不可能であったような、より高度な手法を導入できる可能性を示すことができた。

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 15 件）

① Atsushi Nunome, Hiroaki Hirata, "A Data Migration Scheme Considering Node Reliability for an Autonomous Distributed Storage System", Proceedings of the 5th International Conference on Computational Science/Intelligence and Applied Informatics (CSII 2018), 査読有, 2018

② Hiroki Yamasaki, Atsushi Nunome, Hiroaki Hirata, "Parallelizing the Construction of a k-Dimensional Tree", Proceedings of the 3rd International Conference on Big Data, Cloud Computing, and Data Science Engineering (BCD 2018), 査読有, 2018

③ Kohei Fujisawa, Atsushi Nunome, Kiyoshi Shibayama, Hiroaki Hirata, "Design Space Exploration for Implementing a Software-based Speculative Memory System", International Journal of Software Innovation (IJSI), 査読有, Vol. 6, No. 2, pp. 37-49, 2018, DOI: 10.4018/IJSI.2018040104

④ Shingo Shimano, Atsushi Nunome, Yuta Yokoi, Kiyoshi Shibayama, Hiroaki Hirata: "A Dynamic Configuration Scheme of Storage Tiers for an Autonomous Distributed Storage System", Information Engineering Express, 査読有, Vol. 3, No. 4, pp. 91-104, 2017

⑤ Shingo Shimano, Atsushi Nunome, Yuta Yokoi, Kiyoshi Shibayama, Hiroaki Hirata, "An Autonomous Configuration Scheme of Storage Tiers for Distributed File System", Proceedings of the 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/ Distributed Computing (SNPD 2017), 査読有, pp. 453-458, 2017, DOI: 10.1109/SNPD.2017.8022761

⑥ Sekai Ichii, Shohei Hayashi, Atsushi Nunome, Hiroaki Hirata, Kiyoshi Shibayama, "Performance Evaluation of Delayed-Committing Transactional Memory", Proceedings of the 18th IEEE/ ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/ Distributed Computing (SNPD 2017), 査読有, pp. 445-451, 2017, DOI: 10.1109/SNPD.2017.8022760

⑦ Kohei Fujisawa, Atsushi Nunome, Kiyoshi Shibayama, Hiroaki Hirata, "A Software Implementation of Speculative Memory", Proceedings of the 18th IEEE/ ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/ Distributed Computing (SNPD 2017), 査読有, pp. 437-443, 2017, DOI: 10.1109/SNPD.2017.8022759

⑧ 出島貴史、布目淳、平田博章、"スレッドレベル並列投機実行のためのデータ依存解析機構"、情報処理学会第 79 回全国大会講演論文集、査読無、2G-09、Vol. 1、pp. 67-68、2017

⑨ Atsushi Nunome, Hiroaki Hirata, Kiyoshi Shibayama, "An Interval Control Method for Status Propagation in an Autonomous Distributed Storage System", Proceedings of the 15th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2016), 査読有, pp. 723-728, 2016, DOI: 10.1109/ICIS.2016.7550844

⑩ Hiroaki Hirata, Atsushi Nunome, Kiyoshi Shibayama, "Speculative Memory: An Architectural Support for Explicit Speculations in Multithreaded Programming", Proceedings of the 15th IEEE/ACIS International Conference on Computer and Information Science (ICIS

2016), 査読有, pp. 715-721, 2016, DOI: 10.1109/ICIS.2016.7550843

⑪ Yuki Shoji, Atsushi Nunome, Hiroaki Hirata, Kiyoshi Shibayama, "A Large-Scale Speculation for the Thread-Level Parallelization", International Journal of Computer and Information Science (IJCIS), 査読有, Vol.17, No.1, pp. 24-32, 2016

⑫ 小路勇気、布目淳、平田博章、柴山潔、"スレッドレベル並列投機実行のためのデータ値予測機構"、第14回情報科学技術フォーラム(FIT2015)講演論文集、査読無、C-005、Vol.1、pp.243-244、2015

⑬ Sekai Ichii, Saki Tashiro, Atsushi Nunome, Hiroaki Hirata, Kiyoshi Shibayama, "Hardware Transactional Memory with Delayed Committing", In Proceedings of the 3rd International Conference on Applied Computing and Information Technology (ACIT 2015), 査読有, pp. 161-168, 2015, DOI: 10.1109/ACIT-CSI.2015.38

⑭ Yuki Shoji, Atsushi Nunome, Hiroaki Hirata, Kiyoshi Shibayama, "A Large-Scale Speculation for the Thread-Level Parallelization", Proceedings of the 3rd International Conference on Applied Computing and Information Technology (ACIT 2015), 査読有, pp. 169-175, 2015, DOI: 10.1109/ACIT-CSI.2015.39

⑮ Shingo Shimano, Atsushi Nunome, Hiroaki Hirata, Kiyoshi Shibayama, "An Information Propagation Scheme for an Autonomous Distributed Storage System in iSCSI Environment", Proceedings of the 3rd International Conference on Applied Computing and Information Technology (ACIT 2015), 査読有, pp. 149-154, 2015, DOI: 10.1109/ACIT-CSI.2015.36

〔学会発表〕(計12件)

① Atsushi Nunome, A Data Migration Scheme Considering Node Reliability for an Autonomous Distributed Storage System, The 5th International Conference on Computational Science/ Intelligence and Applied Informatics (CSII 2018), 2018

② Hiroki Yamasaki, Parallelizing the Construction of a k-Dimensional Tree, The 3rd International Conference on Big Data, Cloud Computing, and Data Science Engineering (BCD 2018), 2018

③ Yuta Yokoi, An Autonomous Configuration Scheme of Storage Tiers for Distributed File System, The 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/ Distributed Computing (SNPD 2017), 2017

④ Shohei Hayashi, Performance Evaluation of Delayed-Committing Transactional Memory, The 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2017), 2017

⑤ Kohei Fujisawa, A Software Implementation of Speculative Memory, The 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/ Distributed Computing (SNPD 2017), 2017

⑥ 出島貴史、スレッドレベル並列投機実行のためのデータ依存解析機構、情報処理学会第79回全国大会、2017

⑦ Atsushi Nunome, An Interval Control Method for Status Propagation in an Autonomous Distributed Storage System, The 15th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2016), 2016

⑧ Hiroaki Hirata, Speculative Memory: An Architectural Support for Explicit Speculations in Multithreaded Programming, The 15th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2016), 2016

⑨ 小路勇気、スレッドレベル並列投機実行のためのデータ値予測機構、第14回情報科学技術フォーラム(FIT2015)、2015

⑩ Saki Tashiro, Hardware Transactional Memory with Delayed Committing, The 3rd International Conference on Applied Computing and Information Technology (ACIT 2015), 2015

⑪ Yuki Shoji, A Large-Scale Speculation for the Thread-Level Parallelization, The 3rd International Conference on Applied Computing and Information Technology (ACIT 2015), 2015

⑫ Shingo Shimano, An Information Propagation Scheme for an Autonomous Distributed Storage System in iSCSI Environment, The 3rd International Conference on Applied Computing and Information Technology (ACIT 2015), 2015

〔図書〕（計０件）

〔産業財産権〕

○出願状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
布目　淳　　(NUNOME, Atsushi)
京都工芸繊維大学・情報工学・人間科学系・助教
研究者番号：60335320

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
一井　世界 (ICHII, Sekai)
嶋野　真吾 (SHIMANO, Shingo)
小路　勇気 (SHOJI, Yuki)
田代　早紀 (TASHIRO, Saki)
出島　貴史 (DEJIMA, Takashi)
林　昇平 (HAYASHI, Shohei)
平田　博章 (HIRATA, Hiroaki)
藤澤　昂平 (FUJISAWA, Kohei)
山崎　寛季 (YAMASAKI, Hiroki)
横井　雄太 (YOKOI, Yuta)