様式 C-19

科学研究費補助金研究成果報告書

平成 23 年 6 月 13 日現在

機関番号：82636
研究種目：若手研究(B)
研究期間：2009～2010
課題番号：21700032
研究課題名（和文）大容量メモリを高度に活用しシステム性能のディペンダビリティを向上する技術の研究
研究課題名（英文）A study on advanced memory management techniques that improve performance dependability
研究代表者
河合栄治（KAWAI EIJI）
独立行政法人情報通信研究機構・連携研究部門テストベッド研究推進グループ・主任研究員
研究者番号：40362842

研究成果の概要（和文）：
本研究では、メモリデータが保全されないメモリ領域をアプリケーションから利用するための基本的なプログラミングモデルの開発および基本的な機能の実装を行った。また、それらを連携動作させ、実際にシステム上において人工的なワークロードを用いて、メモリ不足時にもページングを発生させず、安定した性能を維持するための各種パラメータの設定を調査した。さらには、本研究で得られた知見を元に、クラウドコンピューティング環境をターゲットとして、急激なメモリ利用状況の変化にも耐えうるメモリ管理のフレームワークを提案した。

研究成果の概要（英文）：
In this study, a basic programming model for unguaranteed memory areas where the data might be lost in physical memory shortage and the APIs for such memory accesses were developed. In the experiments, the parameter settings under a wide variety of synthetic workloads to improve the stability of performance were investigated. In addition, according to the results obtained from the experiments, a memory management framework for cloud computing environments was proposed, which improved the performance dependability in drastic changes in system-wide memory usage.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,300,000	390,000	1,690,000
2010 年度	1,900,000	570,000	2,470,000
年度			
年度			
年度			
総　計	3,200,000	960,000	4,160,000

研究分野：総合領域
科研費の分科・細目：情報学・ソフトウエア
キーワード：オペレーティングシステム、メモリ管理

１．研究開始当初の背景

従来の OS メモリ管理基盤であるページングは、二次記憶をデータ保全の担保として、仮想メモリ空間への永続的なアクセスを保障することで、プロセスからは簡素で一貫性のあるメモリアクセスモデルを実現している（図１）。しかし、メモリ欠乏時に二次記憶（ハードディスク）へのアクセスが頻発し、性能に対するペナルティが非常に大きいという問題がある。この問題は、現在のメモリの大容量化、二次記憶との性能ギャップの拡大、ネットワークの超広帯域化によるスルー

プット要求の高まりにより、ますます顕在化してきている。特に、近年普及が進みつつある仮想化されたホスト環境では、二重のページングにより物理メモリも仮想化され、問題を一層複雑化している。

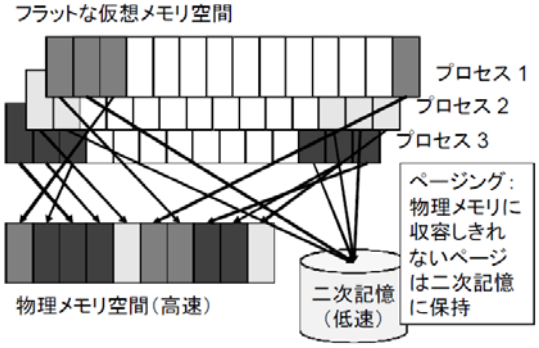

図 1：ページングのしくみ

２．研究の目的

本研究の目的は、OS のメモリ管理の基盤であるページングの性能に対するペナルティを軽減するメモリ管理手法を開発することである。具体的には、ページングによる二次記憶へのメモリデータの保全が不要となる新しいメモリ管理手法を提案する。要点は次の 3 点である。

(1) 二次記憶による担保がなくメモリアクセスに失敗すると、プロセスは異常終了する。これを回避するための一貫性のあるフレームワークをライブラリと OS の協調で実現する。

(2) 高い性能を得るために、メモリ管理の非同期化かつイベント駆動化を実現する。すなわち、メモリ需給状況が逼迫すると、OS は早急にメモリを解放する一方で、プロセスは非同期的に OS からメモリ解放イベントを受け取り、後処理を行う。

(3) ページングと連携してメモリ解放の管理を柔軟に行うことを可能にし、現実性の高いワークロードに対するサービススループット指向な Graceful Degradation を実現する。

本手法により、複数のゲスト OS 環境が動作したり、多様なサービスがホスト上に展開されたりするクラウド環境において、メモリ利用方法が高度化されることが期待される。

３．研究の方法

H21 年度は、メモリデータが保全されないメモリ領域をアプリケーションから利用するための基本的なプログラミングモデルの開発および基本的な機能の実装を行った。具体的には、次の三つの機能となる。

(1) 保全されないメモリ領域を安全に使うための基本的な API の設計

本 API は、メモリ領域の状態遷移を暗黙的に管理することで、非同期的に解放されてしまったメモリ領域へのアクセスに対して安全にエラーを返すことが出来る。図２に用いたメモリ領域の状態遷移図を示す。

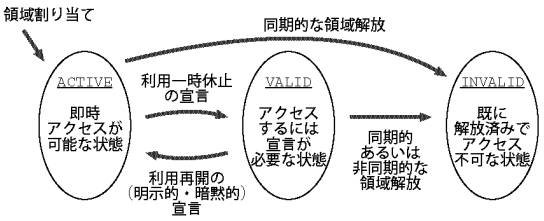

図 2：プログラミングモデルにおける状態遷移

(2) 保全されないメモリ領域を実現するための基本的なシステム機能

本機能は、割り当てたメモリ領域のページングを抑制し、かつメモリ割り当ておよび解放に対してそれぞれ確実に物理ページの割り当ておよび解放を実現する。

(3) アプリケーションプロセスと非同期的にメモリを解放するメモリマネージャスレッド

メモリマネージャスレッドはアプリケーションプロセス内で独立して動作し、システムのメモリ使用状況をモニタリングし、必要に応じて非同期的にメモリ領域を開放する。図３に実装したシステムの概要図を示す。

H22 年度は、H21 年度に開発した(1)～(3)

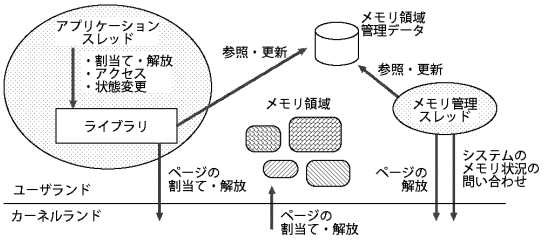

図 3：ライブラリとメモリマネージャースレッド

を連携動作させ、実際にシステム上において人工的なワークロードを用い、メモリ不足時にもページングを発生させず、安定した性能を維持するための各種パラメータの設定を調査した。

さらに、これらの実験を通じて得られた知見を元に、クラウドコンピューティング環境をターゲットとして、急激なメモリ利用状況の変化にも耐えうるメモリ管理のフレームワークを提案した。

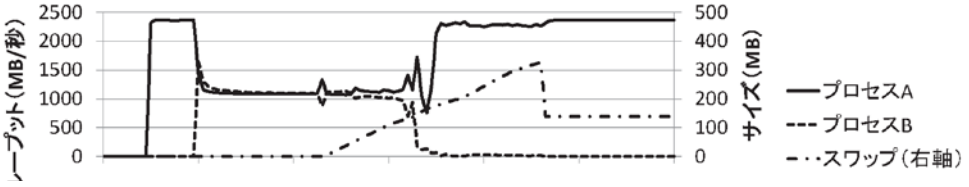

図 4：従来手法の場合の性能

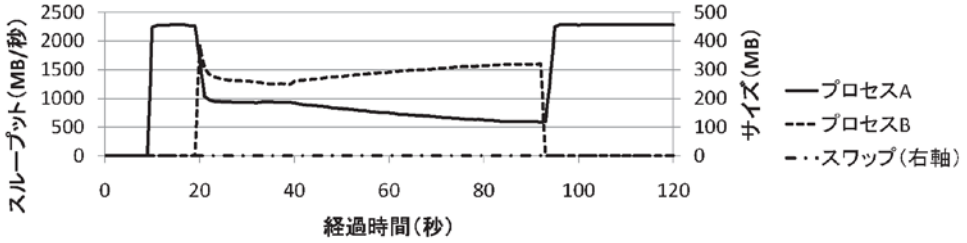

図 5：提案手法の場合の性能

４．研究成果

本研究で実施した性能評価について、図４に従来手法の場合の性能および図５に提案手法の場合の性能を示す。いずれも、プロセスＡが先行してメモリを確保し使用している状況で、プロセスＢがシステムにおいてメモリ不足が発生するだけのメモリを使用しようとした場合の、各プロセスの性能およびスワップ領域の使用状況を示している。

この結果から分かるように、従来手法ではページングが大量発生し、プロセスＢは全く性能が得られない不平等な状況が発生した。一方で、プロセスＡに提案手法を実装した場合、ページングを全く引き起こさず、状況に合わせて自身のメモリ領域を縮減している。

以上の性能評価結果により、メモリが余っている時にはキャッシュ的にメモリを活用してサービス性能を最大化する一方で、システムでメモリ不足が発生した際にはページングが与える性能へのペナルティを最少化する、基本的な要素技術が確立された。最終的には、非常に大量のメモリを急激に使用してもページングを引き起こさないための設定が可能となった。

これらの技術を用いると、特に大規模クラスタ上を用いたボランティアコンピューティングのような環境に応用が可能だと考えている。ボランティアコンピューティングとは，ネットワーク上の計算機の遊休資源を集め，巨大な計算やデータストレージのための分散計算環境を実現するものである．実装例としてはBOINCが良く知られる．

このボランティアコンピューティングでは，個々の計算機の本来のタスク実行への影響を最小化することが必要となる．これまで，CPUについてはバックグラウンドジョブのスケジュールなど広く議論されており，ディスクI/O，ネットワークI/Oについても多数研究開発されている．

一方で，メモリについては，使用メモリ量を監視することはあっても，その量を積極的に制御するということはなされてこなかった．メモリの場合，たとえプロセスの実行を一時的に停止したとしても，利用中の物理ページを回収して本来のタスクに割り当てるためにはページングが必要となることから，性能への影響は不可避である．さらには，ファイルキャッシュ以外のデータにおいて積極的に使用メモリ量を制御するには，プログラムの振る舞いを変化させることが必要であり，プログラミングモデルやフレームワークの確立が必要である．

提案手法は，こうしたボランティアコンピューティングのスケジューリングにおいて，メモリに着目した制御を実現する基礎として動作することができ，他のタスクへの影響を抑えつつ，積極的なメモリの活用が可能となる．

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）

〔学会発表〕（計６件）

1. Takeshi Okuda, A Mechanism of Flexible Memory Exchange in Cloud Computing Environments, In Proc. of IEEE 2nd International Conference on Cloud Computing Technology and Science, The University Place Conference Center, Indianapolis, USA, November 30, 2010.
2. 河合栄治、データ保全しない非同期的なメモリ管理機構、電子情報通信学会、コンピュータシステム研究会、岡山大学、岡山、2010年11月12日
3. Takeshi Okuda, A Remote Swap Management Framework in a Virtual Machine Cluster, In the Proc. of the 3rd International Conference on Cloud Computing (IEEE Cloud 2010), Hyatt Regency Miami, Miami, USA, July 9, 2010.
4. Eiji Kawai, Dynamic and Efficient Memory

Sharing for Cloud Computing Environments, In the Proc of the 18th Pacific Rim Applications and Grid Middleware Assembly (PRAGMA) Workshop, UCSD, San Diego, USA, March 3, 2010.
5. 永井洋太郎、ramfs，iSCSI，LVM を用いた仮想サーバクラスタにおけるメモリ共有手法、情報処理学会第 8 回インターネットと運用技術研究会、作並温泉一の坊、宮城県、2010 年 3 月 1 日
6. Eiji Kawai, Advanced Resource Sharing in the Cloud, In the Proc. of Asian Forum on Information and Communications Technology 2009 (AFICT 2009), Amari Watergate Hotel, Bangkok, Thailand, December 16, 2009.

６．研究組織

(1)研究代表者

河合栄治（KAWAI EIJI）
独立行政法人情報通信研究機構・連携研究部門テストベッド研究推進グループ・主任研究員
研究者番号：40362842

(2)研究分担者

（　　）

研究者番号：

(3)連携研究者

（　　）

研究者番号：