論 文 の 内 容 の 要 旨

論文題目　A Study on Application-Transparent Optimal File Placement in Hierarchical Storage
（階層ストレージにおけるアプリケーション透過なファイル最適配置手法に関する研究）

氏　　名　　松澤　敬一

ITシステムが扱うデータ量は年々増加傾向にある。そのため、ストレージにはより大容量・高性能が求められている。従来主要な記憶メディアとしてはHard Disk Drive (HDD)が用いられてきたが、2000年代後半にはNAND Flashを用いたSolid State Drive (SSD)が市場に流通し始め、性能の高さとHDDとの互換性の高さからその後急速に普及した。2010年代後半には、高速なインターフェースNVM Express (NVMe)が登場し、メディアのアクセス性能はさらに向上した。2019年には不揮発メモリが広く入手可能となり、今後の普及が見込まれる。このように現在は記憶メディアとして、性能・容量・コストが異なる複数の選択肢がある。そこで、データをアクセス特性に応じて適正な記憶メディアに配置し、ITシステム全体でデータ格納における性能・容量・コストのバランスの適正化を図る階層ストレージの必要性が高まっている。

一方、ソフトウェアにおいては、アプリケーションからストレージを利用する方法として、ファイルを基本とするインターフェースが長らく利用されており、今後も継続して利用されるとみられている。そのため、階層ストレージを多様なアプリケーションに適用するには、これらのインターフェースを維持することが必須である。

そこで本論文は、アプリケーションに対し透過的に利用できる階層ストレージのためのデータ最適配置技術について論じる。本論文における透過的とは、既存のアプリケーションに変更なく適用でき、かつエラーや大幅な性能低下などの不具合を起こさないことを指す。また、最適配置とは、データの応答性能に優れた配置であることを示す。

本研究では、既存のアプリケーションで広く利用されているファイル及びPOSIXのインターフェースに着目し、データの最適配置を行う。このインターフェースを維持することで、アプリケーションの変更を不要としたまま階層ストレージを適用可能とする。

また、アプリケーションによるファイル単位のアクセスパターンに着目し、そのパタ

ーンを用いてデータの配置や移動順を定め、応答性能を向上させる。

本論文では、三つの主題について議論する。一つ目の主題は、古い機器を新しい機器に移行するために、データを機器間で複製することである。このような機器間のデータ複製は、今後新たな記憶デバイスを備えた機器が登場する際にも、既存のデータやそのデータを利用するアプリケーションを継続的に利用するために必要な技術である。本項では、既存データをファイルとして格納する旧ファイルサーバから、新規に導入する新ファイルサーバに移行する手法について論ずる。その際、稼働中の旧ファイルサーバや、そのファイルを参照する外部のアプリケーションに対し、変更や長時間の停止時間を要さず、格納されたデータを複製する手法を提案する。提案手法では、ファイル単位のPost-Copy方式で複製を行うことで複製に伴う待ち時間を最小化する。これは、データの複製開始前にアプリケーションが接続先のファイルサーバを切り替えておき、アプリケーションが新ファイルサーバ上で未複製のファイルにアクセスしようとすると、その時点で旧ファイルサーバからアクセス要求への応答に必要なデータだけ複製するものである。一般的なデータセットにおいては、この複製による待ち時間は数秒程度で収まる。一般的なファイルシステムにおけるタイムアウト時間の30〜120秒未満であることから、広範囲のデータセットのファイルサーバに適用できることを確認した。また、Post-Copy方式を実現する機構は全て新ファイルサーバで備えることで、既存の旧ファイルサーバやアプリケーションに変更を要しない手法であることを実現した。

二つ目の主題は、計算機内にある複数の記憶メディア間で、データをアクセス状況に応じて再配置することでストレージのアクセス応答時間を短縮する手法である。階層ストレージにおいて平均入出力性能を高める一般的な手法として、高速な記憶メディアを低速な記憶メディアに対するキャッシュとみなし、今後のアクセスが予測されるデータをプリフェッチする手法がある。しかし、近年のクラウド環境で用いられるInfrastructure-as-a-Service (IaaS)のように、仮想マシンが多数動作し、それぞれが異なるパターンでデータアクセスを行う実行環境においては、ホストOSと仮想マシン間で多層に重なったストレージの処理によって、仮想マシン内のアプリケーションが生じさせるアクセスの局所性をホストOSにおいて観測できなくなる。その結果、アクセスの予測精度が低下し、プリフェッチの効果が減少する。提案手法では、これら仮想マシン内のアプリケーションは、仮想マシン内のファイルに対して空間的なアクセスの局所性を持つことに着目し、仮想マシン内のファイルと物理ディスク上の配置を対応付けるレイアウト情報を仮想マシンからホストOSに送信する。これにより、ホストOS内でプリフェッチ対象のデータ領域を選択する際、物理ディスク上で不連続であっても、仮想マシン内のファイル上で連続する領域を、同時にプリフェッチすることで、アクセスの局所性に基づくプリフェッチの効果改善を図る。提案手法では、TPCx-Vベンチマークにお

いて、レイアウトを認識しない場合にくらべ、17.1%のトランザクション処理性能の改善が見られた。

三つ目の主題は、近年登場した不揮発メモリにデータを格納する際の、性能向上のためのソフトウェアインターフェースの検討である。現在のストレージに対するソフトウェアのインターフェースでは、OSが介在することでI/Oスケジューリングなど記憶メディアへのI/Oを軽減するためのソフトウェア処理と、複数プロセス間でデータやディレクトリ構造の一貫性を保つための排他処理に伴うソフトウェアの性能オーバヘッドが生じる。前者は従来の低速な記憶メディアに対しては有効であるが、不揮発メモリのような高性能メディアに対しては、むしろCPUの実行時間を要しオーバヘッドとなる。また、後者は実用アプリケーションでは稼働中にデータを他アプリケーションと共有することは珍しく、やはり排他処理が実質的に不要なオーバヘッド要因となっている。提案手法では、標準Cライブラリと連携してユーザー空間のライブラリとして動作し、アプリケーションにプロセス固有のディレクトリを提供するユーザー空間ファイルシステム方式を提案する。本ファイルシステムは、プロセスのメモリ空間上にマップした不揮発メモリ上にファイルデータを格納し、CPUのLoad/Store命令で読み書きするため、ファイルアクセスにおけるOSの関与を不要とする。また、本手法では、標準Cライブラリの備えるファイルアクセスインターフェースを維持することで、アプリケーションの変更なく適用可能となる。ファイルアクセスを行うfilebenchベンチマークにおいて、提案手法はXFSファイルシステムに対し6.67倍、先行研究である不揮発メモリ向けのファイルシステムの1.33倍の性能を出せることを確認し、提案手法の性能優位性を確認した。

本論文における研究成果により、計算機間および記憶メディア間において、既存のアプリケーションの変更や再構成を要せずデータを移動し、データアクセスの応答時間を短くする階層ストレージが実現できるようになった。本成果の適用により、既存の多くのアプリケーションに対し、そのデータを適切な記憶メディアに配置・移動し、性能・容量・コストのバランスを適正化できる見込みを得た。