仮想マシンを考慮したIaaSクラウドファイルシステム

2012SE221 佐藤颯哉
指導教員：宮澤元

1 はじめに

近年，クライアントが必要とする計算資源を必要な時に必要な分だけ提供できるクラウドコンピューティングというIT基盤が注目を集めている．特に, IaaS(Infrastructure as a Service) クラウドは，他の様々なクラウドサービスの基盤となっているもので，広く普及している．

IaaSクラウドでは，ユーザに対して仮想マシン (Virtual Machine, VM) と呼ばれるリソースを提供している．仮想マシンとは，物理ホスト上に，仮想化ソフトウェアによって作り出される仮想的なコンピュータのことである．仮想マシンを利用することによって，多数のユーザやプログラムが1台のコンピュータを相互に干渉せずに独立して並行に使用できる，それぞれ別々のOSやソフトウェアが使用できるといった利点がある．

仮想マシンが実行している物理ホストがメンテナンスなどで停止する場合，仮想マシンを構成するメモリやディスクのイメージを別の物理ホストに丸ごと移動し，移動先で仮想マシンの実行を再開することができる．これを仮想マシンのマイグレーションと呼ぶ．仮想マシンのマイグレーションが発生すると，仮想マシンのメモリやディスクのイメージをマイグレーション先に転送する必要がある．このデータ量は非常に大きいので，これをネットワーク転送するオーバヘッドは無視できない．

本研究の目的は，仮想マシンのマイグレーションに伴うメモリとディスクのイメージを転送するオーバヘッドを低減することである．仮想マシンのメモリイメージは，仮想マシンが実行している物理ホストのみに存在するので，マイグレーション時に必然的に転送しなければならない．一方，仮想マシンのディスクイメージはファイルシステムが仮想マシンに対して提供しているものであり，マイグレーションを制御するクラウド基盤ソフトウェアとファイルシステムが連携することにより，転送オーバヘッドを低減できる可能性がある．

本稿では，仮想マシンのマイグレーション時のデータ転送のオーバヘッドを低減するIaaSクラウドファイルシステムを提案する．クラウド基盤ソフトウェアと連携し，ファイルを構成するチャンクを適切に配置することにより，マイグレーションのオーバヘッドを低減できる．

2 クラウドファイルシステム

クラウドコンピューティングにおけるファイルシステム (クラウドファイルシステム) では，非常に巨大なデータを多数の計算ノードで並列処理するために，ファイルを一定サイズのチャンクに分割して管理し，チャンクを複数のサーバに分散配置している．これにより，ファイルを構成する複数のチャンクに並列アクセスしてスループットを向上したり，チャンクごとに複数の複製を作って冗長性を確保したりすることができる [1][2][3].

IaaSクラウドでは，仮想マシンがアクセスするディスクイメージがファイルとしてクラウドファイルシステムに格納されるので，仮想マシンがマイグレーションする際のディスクイメージの転送オーバヘッドが非常に大きくなる可能性がある．仮想マシンがディスクイメージにアクセスする際には，ファイルサーバから仮想マシンが動作する計算ノードにディスクイメージファイルを構成するチャンクが多数ネットワーク転送される．キャッシュなどを用いればファイル転送のオーバヘッドは軽減できるが，仮想マシンのマイグレーションが発生すると，キャッシュも無効となりディスクイメージ全体をマイグレーション先ホストに転送し直さなければならない．

3 マイグレーションオーバヘッドを低減するファイルシステム

我々は仮想マシンのマイグレーション時のデータ転送のオーバヘッドを低減するIaaSクラウドファイルシステムを開発している．ファイルシステムが仮想マシンのマイグレーション管理を行うCloudStack[4] のようなクラウド基盤ソフトウェアと連携することにより，チャンクの配置情報を仮想マシンのマイグレーション先のホストを決定する際のヒントとして利用できる．

3.1 チャンク管理

チャンク管理はファイルシステムが行う．チャンク管理には，チャンク管理ファイル及びチャンクの送信履歴を用いる．チャンク管理ファイルの内容は，チャンクの送信先のうち仮想マシンのディスクイメージが存在するサーバを示す．チャンクの送信履歴の内容は，全てのチャンクの送信先を示し，既に同一のチャンクが存在するサーバをマイグレーション先に定める目的で使用する．本システムのチャンク管理は他のクラウドファイルシステムにあるメタデータ管理と同様であるが，本システムでは，クライアントもチャンクのコピーを保持する．

ファイルをサーバに書き込む際に，仮想マシンのディスクイメージ全体があるチャンクの送信先サーバをチャンク管理ファイルに書き込み，チャンクを送信した全ての送信先サーバを送信履歴ファイルに記録する．このシステムでは，ランダムに定めた複数のサーバに同一のチャンクを保存するが，チャンク管理ファイルには仮想マシンのディスクイメージ全体があるチャンクの送信先のみ記録する．ディスクイメージ全体が保存されていないサーバに保存されたチャンクはディスクイメージの一部でしかないので，

ディスクイメージ全体の保存されているサーバのみ記録する．送信履歴にはチャンクの送信先を全て記録する．

ファイルをサーバから読み込む際は，チャンク管理ファイルを読み込み，チャンク管理ファイルで指定されたサーバからチャンクを読み込む．チャンク管理ファイルにはディスクイメージ全体のあるサーバが記録されているので，そのサーバからデータを受信することになる．

マイグレーション時は，チャンク管理ファイルに記録されているチャンクの位置をマイグレーション元からマイグレーション先に書き換えることにより，仮想マシンのディスクイメージ全体のあるサーバの情報を変更する．

3.2 クラウド基盤ソフトウェアとの連携

クラウド基盤ソフトウェアとの連携はマイグレーション時に行う．例えば CloudStack[4] における管理サーバのように，クラウド基盤ソフトウェアでは仮想マシンの管理などを行うソフトウェアが存在するので，提案するファイルシステムは，このようなソフトウェアと連携する．マイグレーション時に，管理サーバはチャンクの送信履歴を参照し，マイグレーション元以外で同一のチャンクを送信したサーバを探し，それらからマイグレーション先を決定する．管理サーバはこれをもとにマイグレーションを実行する．

4 システムの実装

提案したファイルシステムのプロトタイプを FUSE(Filesystem in Userspace)[5] を用いて実装した．システムの全体図を**図 1** に示す．

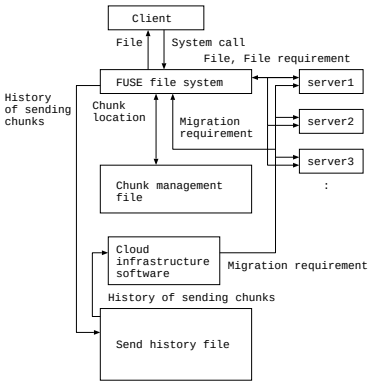

図 1 本システムの全体図

クライアントがファイル要求を出すと，FUSE ファイルシステムにシステムコールを送り，複数台のリモートのサーバと通信し，サーバプログラムにファイル要求を送る．サーバプログラムは，サーバ上のファイルを FUSE ファイルシステムへ送り，FUSE ファイルシステムはサーバから送信されたファイルを処理する．

ファイルはチャンクに分割して保存する．ファイルを読み込む際は，保存したチャンクから読み込み，ファイルをローカルに保存する場合はチャンクから元のファイルを復元する．

3.2 節のクラウド基盤ソフトウェアとの連携については現状では実装できていない．

5 関連研究

マイグレーションオーバヘッドの削減に関する研究には，文献 [6] や文献 [7] がある．文献 [6] の方法は，重複除去によりディスクイメージの転送量を抑えている点が本研究とは異なるが，本研究と同様にディスクイメージの転送量を減らすことによって転送時間が短縮されているので，本研究においても転送オーバヘッドを削減できると考える．文献 [7] のように，オンデマンドでディスクイメージを転送する方法もあるが，この方法を本システムへ適用することも検討する必要がある．

6 おわりに

本稿では，クラウド基盤ソフトウェアと連携してマイグレーションオーバヘッドを低減するファイルシステムを提案した．チャンク管理ファイル及びチャンクの送信履歴をクラウド基盤ソフトウェアに送信することにより適切にマイグレーション先を選択することができる．現在までに，複数のサーバ上のチャンクを読み込み，クライアントに保存することができるファイルシステムを実装した．

作成したファイルシステムをクラウドファイルシステムとして実装し，実験により効果があることを確認することが課題である．また，より効果を高めるために，できるだけ多くのファイルをマイグレーション先のサーバに集約し，実験により効果を確認する必要もある．

参考文献

[1] Ghemawat，et al.：“The Google file system”, ACM SIGOPS operating systems review, Vol.37 No.5, ACM(2003).

[2] Hortonworks：HDFS, `http://hortonworks.com/hadoop/hdfs`, 2016.1.5.

[3] Weil, Sage A., et al.：“Ceph: A scalable, high-performance distributed file system”, Proceedings of the 7th Symposium on Operating Systems Design and Implementation, USENIX Association(2006).

[4] The Apache Software Foundation：Apache CloudStack: Open Source Cloud Computing, `https://cloudstack.apache.org/`, 2016.1.5.

[5] SourceForge.net：FUSE: Filesystem in Userspace, `http://fuse.sourceforge.net`, 2016.1.5.

[6] 森田 大希 他：“仮想クラスタを構成する複数ディスクイメージの効率的移送手法” 情報処理学会研究報告 Vol.2012-OS-121 No.10(2012.5)

[7] 広渕崇宏，他：“仮想計算機遠隔ライブマイグレーションのための透過的なストレージ再配置機構” 情報処理学会論文誌 vol. ACS26 152-165(2009)