様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　６月　１日現在

機関番号：12601
研究種目：若手研究（Ｂ）
研究期間：2010～2011
課題番号：22700022
研究課題名（和文）　仮想化技術による分散ストレージ管理基盤

研究課題名（英文）　Distributed Storage Management with Virtualization Technology

研究代表者
品川　高廣（SHINAGAWA TAKAHIRO）
東京大学・情報基盤センター・准教授
研究者番号：40361745

研究成果の概要（和文）：本研究では，クライアント・マシンの HDD 等のストレージの内容をサーバで集中管理しつつ，クライアント側の計算資源は最大限活用することが出来るシステムの実現を目的として,仮想化技術を用いた分散ストレージ管理基盤に関する研究をおこなった．我々が研究開発している独自の仮想マシンモニタ「BitVisor」を活用して，ストレージ以外のデバイスへのアクセスはパススルーとしてゲスト OS がハードウェアの機能をオーバーヘッドゼロで全ての機能を最大限に活用できるようにしつつ，ストレージ・デバイスへのアクセスを部分的に仮想化することで，OS からは完全に透過的にサーバに転送することを可能にした．実験の結果，既存の Windows や Linux 等の OS の設定を一切変更することなくネットワークブートできることを確認した．また，サーバ OS のキャッシュが有効の場合には起動時間が 14 秒程度短縮されることを確認した．

研究成果の概要（英文）：The purpose of this work was to develop a distributed storage management system that allowed central management of contents of storage devices, such as HDDs, of client machines while the client machines could fully utilize their hardware resources. We exploited our original hypervisor called BitVisor to allow maximum utilization of hardware functionalities from guest OSs with no extra overhead by allowing pass-through access to the hardware resources, while storage access was transparently transferred to remote servers by partially virtualizing storage devices with the hypervisor. Experimental results confirmed that our hypervisor could network-boot existing OSs such as Windows and Linux without any modification to the configuration of the OSs. The results also confirmed that the boot time was reduced by 14 seconds compared to that of local boot when the server cache was enabled.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,700,000	510,000	2,210,000
2011 年度	1,400,000	420,000	1,820,000
年度			
年度			
年度			
総　計	3,100,000	930,000	4,030,000

研究分野：総合領域
科研費の分科・細目：情報学・ソフトウェア
キーワード：仮想マシン，分散ストレージ，ネットブート

様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

１．研究開始当初の背景

近年，仮想マシンモニタの研究が再び脚光を浴びている．PC の性能向上により仮想化技術が容易に利用可能になってきたほか，セキュリティ向上や管理コスト削減など従来のメインフレームとは異なる仮想化技術の応用例が増えている．

申請者らは，仮想マシンモニタによるセキュリティ向上を目指した「セキュア VM プロジェクト」で中心的役割を果たし，平成 18 年度科学技術振興調整費（研究代表者：加藤和彦）による支援のもと，純国産の仮想マシンモニタ「BitVisor」を研究開発した．BitVisor は，準パススルー型という新しいアーキテクチャにより，仮想マシンモニタ自身を小さくシンプルに保ちつつ，ゲスト OS からは透過的に暗号化などのセキュリティ機能を実施できるソフトウェアで，その新規性・有用性は国際的にも認められている．また，申請者らは BitVisor の機能を応用して Kernel Rootkit からのファイル保護を実現するなど，BitVisor の特徴を生かした更なる応用に関する研究をおこなっている．

本研究では，BitVisor の特徴である準パススルー型アーキテクチャの利点（高い信頼性・安全性，低オーバーヘッド）を分散ストレージ管理の分野に応用することにより，仮想化の利点を保ちつつ，高い透過性や低オーバーヘッドを実現できる新しい仕組みの構築を目指している．

２．研究の目的

企業や大学など多数のデスクトップ・コンピュータを用いる環境では，セキュリティの向上や管理コスト削減のためにコンピュータを集中管理したいという要望が強い．しかし従来方式では，サーバ側に高価なマシンが必要であったり，OS の機能に依存したりして，維持管理が難しくなってしまう．本研究では，仮想化技術による分散ストレージ管理基盤を実現する．申請者らが提案している準パススルー型という独自アーキテクチャの仮想マシンモニタ『BitVisor』を応用することで，ストレージをサーバで集中管理しつつ，クライアント・コンピュータの資源を活用して高価なサーバを不要にし，かつ任意の OS を無改変で稼働可能にして維持管理を容易に行えるようにする．

本研究では，多数のコンピュータがネットワーク接続された環境で，以下のような利点を持つシステムを実現する．

(1)ストレージの内容をサーバで集中管理し，面倒な多数の計算機の管理作業を軽減する

(2)仮想化のオーバーヘッドをほぼゼロにし，個々の計算機の資源を最大限活用可能にする

(3)クライアント側の OS は一切改変不要とし，Windows を含む任意の OS を稼働可能にする

これらの点は，申請者らが開発した仮想マシンモニタ「BitVisor」活用することで実現する．BitVisor は「準パススルー型」というユニークなアーキテクチャを提案・実装しており，原則としてゲスト OS にハードウェアを直接アクセスさせつつ，必要最小限の I/O のみを捕捉・修正してセキュリティ機能などを実現する．本研究では，準パススルー型アーキテクチャの特徴を最大限活用して，以下のようなユニークな機能を持つストレージ管理機能を研究開発する．

(a)ストレージ（ハードディスク等）へのアクセスを I/O レベルでサーバに転送する

ストレージアクセスをサーバに転送することにより，各計算機のストレージの内容を集中管理できる．仮想マシンモニタで I/O のレベルで転送することにより，ゲスト OS からは完全に透過的に実現可能になり，起動中を含めた全てのストレージアクセスを集中管理することが可能になる．

(b)仮想化の ON/OFF を稼働中にダイナミックに切り替え可能にする

ストレージアクセスをサーバに転送するモードと，クライアントのハードディスクに直接アクセスするモードを動作中に切り換えられるようにすることにより，ストレージの内容はサーバで集中管理しつつも，ローカルのハードディスクに既にデータがある場合は直接アクセスしてオーバーヘッドを低く抑えることを可能にする．さらに，ストレージ（及びネットワーク）以外のハードウェアは完全パススルーでゲスト OS が直接アクセス可能なため，ハードディスクの準備が整った時点で仮想化機能を完全に OFF にして，完全ゼロオーバーヘッドでの実行も可能になる．

これらの機能を実現することより，例えば以下のような運用が可能になる．

○任意の OS のネットワークブート

クラスタ環境などにおいて，多数の計算機を新しい OS や新しい設定の OS イメージで起動することが可能になる．従来のネットワークブートと比べた利点は，ホストコントローラのレベルで仮想化するため，Windows を含む任意の OS を無改変で起動可能であるという点である．従来の仮想マシンモニタと比べた利点は，ストレージ以外は仮想化せずゲスト OS がハードウェアを直接アクセス可

能なため，オーバーヘッドが極めて小さい点である．ローカルのディスクをキャッシュとして用いることにより，一度アクセスしたセクタデータはサーバにアクセスすることなく読み書き可能になるため，ストレージアクセスのオーバーヘッドも低く抑えることが可能である．

○インクリメンタルインストール＆ゼロオーバーヘッド実行

上記の応用として，OS を実行しながらインストールすることが可能になる．すなわち新しい OS を起動した場合，ネットワークブートでインストール作業を待つことなく利用可能にしつつ，実行しながらディスクへの書き込み作業を並行しておこなうことで，インストール作業をインクリメンタルにおこなうことが出来る．さらに，いったんインストール作業が終了した後は，ストレージを含むハードウェアの仮想化が不要になるため，通常のインストールをおこなった場合と同等，すなわちゼロオーバーヘッドでの実行が可能になる．

３．研究の方法

本研究で提案するシステムは，ベースとなる仮想マシンモニタ BitVisor のコア機能に加えて，(1)ATA マネージャ，(2)NIC マネージャ，(3)ストレージマネージャ，(4)ストレージサーバ，の４つのコンポーネントから構成される．本研究では，まず「ATA マネージャ」及び「NIC マネージャ」を設計・実装して，ゲスト OS から ATA デバイスへのアクセスを NIC 経由でネットワーク越しのサーバに転送できるようにする．この段階では，ストレージサーバとしては既存の技術（ATA over Ethernet）を活用して，クライアント側の動作を確認する．次に，「ストレージマネージャ」を開発して，ゲスト OS からのアクセスをローカルとリモートに振り分けられるようにするとともに，「ストレージサーバ」を拡張して多数のクライアントのストレージを適切に管理できるようにする．最後に，これらを組み合わせて実環境での動作確認・性能測定をおこなう．

平成２２年度には，分散ストレージ管理システムの各種コンポーネントの設計・実装をおこなう．提案システムは，大きく分けて，(1)ATA マネージャ，(2)NIC マネージャ，(3)ストレージマネージャ，(4)ストレージサーバ，の４つのコンポーネントから構成される（図１参照）．

ATA マネージャは，ATA 関係の３つのコンポーネント（物理 ATA コントローラ，ATA ドライバ，ストレージマネージャ）を仲介するコンポーネントである．ATA ドライバからの I/O アクセスに対して，(1)物理 ATA コントローラにそのまま転送する「パススルーモード」か，(2)ストレージサーバへ転送する「ネットワークモード」のいずれかで動作する．パススルーモードで動作時は，ATA ドライバが物理 ATA コントローラに対して直接アクセスできるため，オーバーヘッドが非常に低く抑えられる．一方，ネットワークモードで動作時は，ローカルのハードディスクにデータがなくてもストレージアクセスが可能になる．ローカルのハードディスクとストレージサーバを混在して利用可能にするために，ATA マネージャはセクタ単位でパススルーモードとネットワークモードを切り替え可能にする．また，ローカルのハードディスクをキャッシュとして利用するために，物理 ATA コントローラへのアクセスを多重化して，ATA ドライバとストレージマネージャが同時にアクセスできるようにする．

NIC マネージャは，ストレージマネージャが物理 NIC を経由してストレージサーバにアクセスできるようにするためのコンポーネントである．NIC ドライバとストレージマネージャが物理 NIC に同時にアクセスできるようにするために，NIC マネージャは物理 NIC へのアクセスを多重化する．現在の BitVisor でも NIC 多重化機能は搭載されているが，現在の実装ではデバイスの初期化などをゲスト OS の NIC ドライバに任せているため，起動時など OS が動作する前でも物理 NIC にアクセスするための機能を新規開発して，ゲスト OS の起動前でも NIC にアクセス可能にする．

初年度は，ATA マネージャのネットワークモードと NIC マネージャを先行して実装をおこない，ストレージサーバは既存の ATA over Ethernet の技術を利用する．これにより，ATA デバイスへのネットワーク越しでのアクセス機能の実現可能性について先行して検証する．

また，初年度にはストレージマネージャとストレージサーバの設計もおこなう．ストレージマネージャは，ストレージアクセス全体を管理するコンポーネントである．ストレージマネージャの役割は大きく分けると，(1)ゲスト OS からのディスクアクセスをストレージサーバに転送するストレージクライアント機能と，(2)ローカルのハードディスクの内容をゲスト OS に代わって管理するディスク管理機能がある．ストレージクライアント機能は，ゲスト OS の ATA ドライバからの I/O アクセスをサーバに対するアクセスへと変換するコンポーネントである．変換すべき I/O は，アクセス対象（セクタの LBA アドレス，セクタ数，転送方向など）を指定する I/O と DMA 転送のための I/O の２つがある．これらの I/O 要求をカプセル化してネットワークに転送することにより，サーバに対してセクタデータを読み書きできるようにする．転送

プロトコルの設計にあたっては，既存のカプセル化手法（iSCSI，ATAPI，ATA over Ethernet など）を参考にする予定である．ディスク管理機能は，サーバから送られてきたセクタデータをゲスト OS とは独立してローカルのハードディスクに書き込むコンポーネントである．これにより，セクタデータのキャッシュやハードディスクへのインストール作業をゲスト OS から独立におこなうこと等が可能になる．

ストレージサーバは，クライアントで動作する仮想マシンモニタからカプセル化された I/O 要求を受け取って，サーバのハードディスクからセクタデータの送受信をおこなうソフトウェアである．ストレージサーバ上では通常の OS を動作させられるので，実際のディスクイメージはファイルとして OS 内に格納することができる．ストレージサーバは I/O 要求の中の LBA やセクタ数などの情報からディスクイメージへのアクセスをおこない，結果をカプセル化してクライアントに返信する．ストレージサーバはストレージクライアントと対になるものであり，転送プロトコルの設計に当たっては，クライアント同様既存のシステムを参考にしながら行う予定である．

平成２３年度には，まず ATA マネージャのネットワークモードとパススルーモードの切り替え機能を実装する．次に，ストレージマネージャ及びストレージサーバの実装をおこなって，システム全体を構築する構築したシステムは，研究目的にも記載した以下の２つの利用形態を例として実際に実験をおこなって，実現可能性の検証や，性能・オーバーヘッドの測定などをおこなう．

（１） 任意の OS のネットワークブート

予めサーバに用意した様々な OS イメージから起動したい OS を選択することで，Windows XP や Linux，FreeBSD など異なる OS を起動したり，同じ OS でもディストリビューションやインストールされたプログラムなど環境が異なるものをいつでも起動したり出来るようにする．

（２） インクリメンタルインストール＆ゼロオーバーヘッド実行

OS の実行とインストールを同時におこなうことを可能にする．基本的には上記のディスクレスモードと同様に OS からのアクセスをサーバに転送して OS を実行しつつ，同時に仮想マシンモニタが裏でディスクへの書き込みをおこなってインストール作業をおこなう．

４．研究成果

提案方式が動作環境の点で透過的であることを示すために，既存の通常の PC 向け OS を一切改変せずにネットワークブートできることを確認する実験を行なった．実験では，Linux の標準的なディストリビューションである Debian 5.0 (Lenny) と，通常の PC 向け OS として一般的な Windows Vista (Business Edition)を使用した．OS イメージの作成には，PC エミュレータである QEMU を用い，通常の PC と同様の手順でイメージファイルにインストールを行なった．作成したイメージファイルを vblade に指定し，クライアント端末から提案方式でネットワークブートした．実験の結果，Debian 5.0 (Lenny)，Windows Vista (Business Edition) 共に，OS イメージそのものは一切改変せずに，提案方式によってネットワークブートできることを確認した．

また，提案方式が機能面でも透過的であることを示すために，ネットワークブートした OS からクライアント端末の周辺機器・内蔵機器が利用できることを確認した．USB デバイスであるフラッシュメモリや，端末に内蔵されている CD/DVD/Blu-ray ドライブからデータの読み書きができることを確認した．また，OS からグラフィックボードの機能を認識でき，デュアルディスプレイが可能なことも確認した．

性能面の透過性を測定する実験では，，キャッシュを有効にしてネットワークブートした場合，OS の起動時間は，ローカルディスクから起動した場合に比べて 14 秒短縮し，BitVisor 自体の起動時間を合わせても 8 秒短縮する結果となった．また，ディスクベンチマークソフトである Crystal Disk Mark を用いた実験では，サーバ OS のキャッシュ

を有効にした状態で OS をネットワークブートした場合は，ローカルディスクから起動した場合に比べて，シーケンシャルリードで 26.1%，シーケンシャルライトで 8.2%スループットが向上した．また，レコードサイズ 4K バイトの場合，ランダムリードで 6.7 倍，ランダムライトで 2.6 倍スループットまで向上した．

既存の仮想マシンモニタ方式との比較では，KVM と NFS を用いたネットワークブート方式と比較して，BitVisor 上でネットワークブートした場合のほうが，12 秒短いことがわかった．ディスクアクセスのスループットは，シーケンシャルリードでは，KVM 上に比べて 21.8MB/sec．ライトでは 46.6MB/sec スループットが大きいことがわかった．一方，レコードサイズ 512K のランダムリードでは，21.6MB/sec，ライトでは 56.1MB/sec スループットが大きいことがわかった．

また，ローカルストレージへのインストールをおこなうシステムも実装し，インストール完了後にはローカルブートと全く同等の性能が実現できることを確認した．

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１件）
① 表 祐志, 品川 高廣, 加藤 和彦. 仮想マシンモニタによる透過的ネットワークブート方式. 情報処理学会論文誌：コンピューティングシステム, 第 4 巻, 第 4 号, 228-245 頁, 2011 年 10 月. 査読有. http://id.nii.ac.jp/1001/00078067/

〔学会発表〕（計２件）
① 表 祐志, 品川 高廣, 加藤 和彦. 仮想マシンモニタによる透過的ネットワークブート方式. 第 22 回 コンピュータシステム・シンポジウム(ComSys 2010), 第 2010 巻, 第 13 号, 3-12 頁, 大阪, 2010 年 11 月. 査読有. 若手優秀論文賞受賞.
② 表 祐志, 品川 高廣, 加藤 和彦. VMM による透過的ネットワークブートシステム. 第 8 回先進的計算基盤システムシンポジウム (SACSIS2010), 奈良, 2010 年 5 月. ポスター発表. 優秀ポスター賞受賞.

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）

○取得状況（計 0 件）

〔その他〕
ホームページ等
http://www.os.ecc.u-tokyo.ac.jp/
http://www.bitvisor.org/

６．研究組織
(1)研究代表者
品川　高廣（SHINAGAWA TAKAHIRO）
東京大学・情報基盤センター・准教授
研究者番号：40361745