仮想化技術とオペレーティングシステムの研究

品川 高廣

1 概要

我々の研究室では、仮想化技術やオペレーティングシステム（OS）に関する研究をおこなっている。仮想化技術に関しては、仮想マシンモニタと呼ばれるシステムソフトウェアに関する研究をおこなっている。仮想マシンモニタとは、OS の下で動作してハードウェアを仮想化するソフトウェアであり、既存の OS に依存することなく新たな機能の追加を安全かつ容易に実現できるといった利点があるため、近年様々な応用に向けた研究が盛んにおこなわれている。本研究で主に対象としているのは、我々が考案した「準パススルー型」という新しいアーキテクチャの仮想マシンモニタである。準パススルー型とは、物理的なハードウェアを敢えて仮想化せずに可能な限り OS に対して直接そのまま見せる、すなわち OS からハードウェアへのアクセスをなるべく通過（パススルー）させつつ、必要最小限のアクセスだけを捕捉・変換する方式である。準パススルー型アーキテクチャは、(1) OS よりも高いセキュリティを実現できる、(2) OS の機能に依存せずに新たな機能を追加できるといった仮想マシンモニタの利点を保ちつつも、VMWare や Xen のような従来型アーキテクチャの仮想マシンモニタと比べて、(1) 仮想化のオーバーヘッドを大幅に削減できる、(2) ゲスト OS がハードウェアの機能を最大限活用できる、といった利点がある。我々の研究室では、準パススルー型アーキテクチャを実装した仮想マシンモニタ「BitVisor」をベースとして研究開発をおこなっている。OS に関しては、OS のカーネルを改変してセキュリティ機能を始めとした新しい機能を追加したり、アプリケーションの性能を向上させるための研究をおこなっている。

今年度は、主に以下のテーマに関して研究をおこなった。

- 仮想マシンモニタによるハードウェア抽象化レイヤ [査読付 1]

  デバイスドライバの維持管理コストを削減して新しい OS の開発コストを削減することを目的として、仮想マシンモニタのレイヤでデバイスドライバを実行するソフトウェアレイヤを構成し、広く一般に普及している virtio のインターフェイスでアクセスできるようにすることで、デバイスドライバの開発・維持管理を一元的におこなえるようにするシステムの研究開発をおこなった。

- ARM ベースのマイクロデータセンターにおけるライブマイグレーション [査読付 2]

  ARM マシンで構成されるマイクロデータセンターにおいて、仮想化のオーバーヘッドを低く抑えつつライブマイグレーションを実現するシステムに関する研究開発をおこなった。

- 仮想マシンモニタによるベアメタルクラウドのハードウェア保護 [査読付 3]

  ベアメタルクラウドのように物理マシンを提供している環境でハードウェア破壊やウィルス感染を防止することを目的として、仮想マシンモニタを用いて不揮発性メモリ領域へのアクセスのみを制限することで、性能を損なうことなくセキュリティを向上させることが出来るシステムに関する研究開発をおこなった。

- VM を意識した適応的キャッシュプリフェッチ [査読付 4]

  SSD と HDD を組みあせたような階層型のストレージ・システムにおいて、キャッシュすべきデータを適切に判断するために、VM 内のエージェントと連携して、ファイルシステムの構造に関する情報や、データベースのパフォーマンスに関する情報などを取得して、これからアクセスされるデータを予測したり、状況に応じてプリフェッチ速度を調整するなどの手法に関する研究をおこなった。

- macOS 上で Linux バイナリを動作させる技術に関する研究 [査読付 5]

  仮想化技術によりシステムコールやページフォルトなどをユーザレベルのプロセスで処理できるようにすることで、柔軟かつ堅牢な OS 互換レイヤを実現する手法に関する研究をおこなった。

- ファイルサーバの高速なマイグレーション [発表 1]

  ファイルサーバを新しくする際に、ファイルサービスを停止することなく中身のファイルをオンデマンドでコピーする手法に関する研究をおこなった。

- JIT ROP 攻撃を防止するための軽量な保護機構に関する研究 [発表 2]

  Google Chrome などの Web ブラウザに搭載されている JavaScript を Just-In-Time(JIT) コンパイルするエンジンに対して、情報漏えいのバグなどを利用してコードを読み込んで動的に Return-Oriented Programming(ROP) のチェーンを構築する JIT ROP 攻撃を防止するために、特定の値の定数を見えなくする Constant Blinding を軽量に実現する手法に関する研究をおこなった。

- 仮想マシンモニタによるベアメタルクラウドでのライブマイグレーション

  ベアメタルクラウドと呼ばれる物理マシンを貸し出す IaaS 環境においてライブマイグレーションを実現することを目的として、軽量な仮想マシンモニタによって物理デバイスの状態を取得・転送することにより OS からは透過的に物理マシンの状態を別のマシンに転送することが出来るシステムに関する研究をおこなった。今年度は主にマルチコア対応をおこなった。

- 仮想マシンモニタによる仮想マシンモニタのデバイスドライバの堅牢性検査

  仮想マシンモニタを二重化するネステッド仮想化の技術を用いて、VMWare など既存の仮想マシンモニタが持つ固有のデバイスドライバに対して、ハードウェアが想定外の動作をした場合に仮想マシンモニタがクラッシュしたりしないかどうかを検査するためのフレーム枠に関する研究をおこなった。

- 仮想マシンモニタによる DDoS 攻撃防止のためのパケットフィルタリング

  クライアント端末からの大量のパケット送信によるサーバに対する DDoS 攻撃を防止することを目的として、予め軽量仮想マシンモニタをクライアント側に導入しておくことで、DDoS 攻撃を受けているサーバ側からの要求に応じてパケット送信を停止できるようにすることで、攻撃を停止させる手法に関する研究開発をおこなった。

- ネステッド仮想化の動的 ON/OFF による仮想マシンモニタ若化

  クラウド環境において VMM がメモリリークなどによって徐々に性能が低下していくエイジングという問題に対して、稼働中の仮想マシン（VM）を止めることなく単一マシン上で高速に VMM を再起動できるようにすることで、VMM の若化を図る手法に関する研究をおこなった。

- ネットワークスタックと不揮発性メモリの統合による永続性 Key-Value Store の高速化

10 Gbit ネットワークや不揮発性メモリ（Non-Volatile Memory: NVM）のような高速な I/O デバイスの登場によって、OS のプロトコルスタックやファイルシステムなどのレイヤを経由するとハードウェアの性能を十分に活かしきれなくなっている問題に対して、Key-Value Store に特化してソフトウェアを極力バイパスしたシステムを構築することでハードウェア性能を最大限活かす手法に関する研究をおこなった。

- OS カーネルにおける BPF によるアクセス制御に関する研究

  OS カーネル内において柔軟かつ動的なアクセス制御をおこなうためのフレームワークの実現を目的として、Linux カーネルの Linux Security Module (LSM) と Berkeley Packet Filter (BPF) を組み合わせてプログラムによりアクセス制御の仕組みを記述する方式に関する研究をおこなった。

これらの研究は、主に以下の事業による支援により実施された。

- 公益財団法人三菱財団 平成 29 年度自然科学研究助成 研究代表者
- 日本学術振興会 科学研究費補助金 基盤研究 (B) 「準パススルー型仮想マシンモニタに関する研究」 研究代表者

研究成果としては、まず国際会議での査読付き論文の発表を 5 件（ACM SAC 2018, IEEE EdgeCom 2018, IEEE CloudCom 2017 × 2, ACM APSys 2017）おこなった [査読付 1, 査読付 2, 査読付 3, 査読付 4, 査読付 5]。他にも、査読なし国際会議での発表が 1 件 [発表 2]、国内シンポジウムでの発表を 1 件 [発表 1]、ポスター発表を 1 件 [発表 3] おこなった。2017 年 12 月には、第 29 回コンピュータシステム・シンポジウムの併設イベントとして、BitVisor Summit 6 を企画立案・開催した。1 件の招待講演と 8 件の一般発表で構成し、30 名前後の参加者があった。また、ソフトウェアとして BitVisor 2.0 [公開 1] を公開したほか、CPU の仮想化機能の性能を測定する VMXbench [公開 2] を公開した。また、共著で教科書 1 件の改訂を行った [著書 1]。

以下、2 章では、仮想マシンモニタによるハードウェア抽象化レイヤについて、3 章では、仮想マシンモニタによるベアメタルクラウドのハードウェア保護について、4 章では、仮想マシンモニタによる物理デバイスドライバ検証環境について、5 章では、macOS 上で Linux バイナリを動作させる技術に関する研究について、6 章では、ファイルサーバの高速なマイグレーションについて、7 章では、JIT ROP 攻撃を防止するための軽量な保護機構に関する研究について、それぞれ概要を述べる。

2 仮想マシンモニタによるハードウェア抽象化レイヤ

2.1 背景

デバイスドライバは、開発コストやコード品質を保つという観点からして、既存および新規のオペレーティングシステム（OS）における主要な関心事である。しかし、デバイスドライバは OS に強く依存しているため、デバイスドライバの実行環境が OS カーネルと密接に結びつき、デバイスドライバの再利用や統一が複雑になる。主要 OS からデバイスドライバを移植する手法では、さまざまな競合やエンジニアリングコストが発生してしまう。また、仮想マシン（VM）を用いてデバイスドライバを再利用する手法は、無視できないオーバーヘッドがかかる。

2.2 内容

本研究では、薄いハイパーバイザを使用した統一されたハードウェア抽象化レイヤの設計と実装を示す。本ハイパーバイザの上で動作する OS カーネルは、各デバイスクラスごとに 1 つのデバイスドライバを持てばよいため、OS のデバイスドライバの開発コストが削減される。本方式の主要な技術は、デバイスマスカレードである。デバイスマスカレードは、物理的なデバイスを最小限の労力で標準化

された抽象的なデバイスに変換する。ハードウェアデバイスに対するデファクト・スタンダードのインタフェースを利用することにより、抽象化レイヤの実装を OS カーネルからきれいに分離し、実際のシステムにおいて容易に利用できるようにする。仮想化のオーバーヘッドを削減するため、ハイパーバイザは単一の VM のみをサポートし、割り込みコントローラなど既に標準化されたデバイスへはパススルーアクセスでアクセスさせる。実験結果により、我々のシステムの性能は、ハイパーバイザのないベアメタルマシンの性能に匹敵することを確認した。

2.3 具体的成果

本研究の成果は、国際会議 The 33rd ACM/SIGAPP Symposium On Applied Computing (ACM SAC 2018) で論文として発表予定である。

3 仮想マシンモニタによるベアメタルクラウドのハードウェア保護

3.1 背景

従来の IaaS（Infrastructure-as-a-Service）クラウドは、仮想マシンをサーバーとして提供する。しかし、仮想化はパフォーマンスオーバーヘッドを招くほか、ハードウェア機能を最大限に活用することが難しい。そこでいくつかの IaaS ベンダーは、仮想マシンではなく物理的なハードウェアを提供するベアメタルクラウドと呼ばれる新しいサービスを開始した。しかし、物理ハードウェアをユーザーに公開すると、クラウドベンダーにとってはハードウェア保護の問題が発生する。物理ハードウェアは不揮発性メモリ（NVM）にファームウェアコードや設定データを保存する。したがって、NVM が悪意のあるユーザーによって変更された場合、ハードウェアはマルウェアによって永続的に破損したり他に感染したりする可能性がある。ベアメタルクラウドにはハードウェアを保護するための仮想化レイヤがないため、クラウドベンダーにとってはこれを防ぐのは難しい。

3.2 内容

本研究では、まずベアメタルクラウドで可能な攻撃の種類を説明し、次にベアメタルクラウドのハードウェア保護スキームである BMCArmor を提案する。BMCArmor は、ハードウェアを仮想化せず NVM へのアクセスを防ぐだけの薄いハイパーバイザを使用する。実験により、BMCArmor は実際にハードウェアの保護を実現できることや、パフォーマンスのオーバーヘッドがほとんど無いことが分かった。

3.3 具体的成果

本研究の成果は、国際会議 9th IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2017) で論文として発表した。

4 VM を意識した適応的キャッシュプリフェッチ

4.1 背景

ストレージ・キャッシュ・プリフェッチは、アクセス局所性に基づいてアクセス・パターンが予測可能である場合は、階層ストレージ・システムのアクセス・レイテンシを削減する有効な手法である。しかし、IaaS（Infrastructure-as-a-Service）クラウドでは、ストレージの仮想化によってデータが大幅に再配置され、ホストオペレーティングシステム（OS）で観測されるアクセスの空間的局所性が低下する。さらに、IaaS クラウドでは時間とともに変化する可能性のある様々なワークロードを持ったアプリケーションを一台のマシン上で動作させることになる。従って、アクセスパターンは空間的および時間的の両方で大きく変化する。

4.2 内容

本研究では、仮想マシン（VM）内の構造情報と統計情報を利用した適応型ストレージキャッシュプリフェッチ方式を提案する。ゲスト OS におけるアプリケーションのファイル使用状況および内部ファイルレイアウト情報の観察により、ホスト OS はストレージアクセス中に空間的および時間的局所性を取得することができる。さらに、アプリケーション・レベルのパフォーマンス統計を用いることで、過度のプリフェッチによるパフォーマンス低下を防ぎ、ホスト OS がプリフェッチ速度を適応的に調整することができる。我々は VM 内で動作する Linux 及び PostgreSQL と協調するプロトタイプのキャッシュプリフェッチシステムを実装した。TPCx-V ベンチマークを使用した実験では、VM を意識した手法が従来のプリフェッチと比較してパフォーマンスを 17.1 %向上させることが示された。また、我々のシステムは、既存の非プリフェッチキャッシングシステムよりも 3.15 倍優れたパフォーマンスを達成した。

4.3 具体的成果

本研究の成果は、国際会議 9th IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2017) で論文として発表した。

5 macOS 上で Linux バイナリを動作させる技術に関する研究

5.1 背景

Linux はプロダクション環境として人気のあるオペレーティングシステム（OS）であるが、一方で多くの開発者が macOS を日々の開発環境として利用することを好んでいる。この状況に対処する方法としては、仮想マシンで Linux を実行する方法や、アプリケーションを Linux から macOS に移植する方法がある。しかし、仮想マシンにはリソース共有の問題があるほか、アプリケーションの移植は非常にコストがかかるため、完全には移植できないことがしばしばある。低コストでかつシームレスなリソース共有を実現するための有望なアプローチとしては、MacOS 用の Linux 互換レイヤを開発する方法がある。しかし、OS 互換レイヤを実装する既存の方法は堅牢性または柔軟性に欠ける。

5.2 内容

本研究では、OS 互換レイヤの新しいアーキテクチャを提案する。この仕組みでは、ホスト OS のコアエミュレーションレイヤーをユーザー空間で実装することで堅牢性を向上させながら、仮想化技術を活用してホスト OS のカーネルに大きく依存することなく柔軟で強力なエミュレーション機能を実現する。本手法を実装し、Ubuntu のユーザランドが macOS 上で動作することを確認した。また、実験結果により、本手法が現実のアプリケーションに対して妥当な性能を有することを確認した。

5.3 具体的成果

本研究の成果は、国際会議 8th ACM SIGOPS Asia-Pacific Workshop on Systems (APSys 2017) で論文として発表した。

6 ファイルサーバの高速なマイグレーション

6.1 背景

ファイルサーバの移行は，格納ファイルの複製を伴うため長時間を要する．この間ファイルアクセスが停止すると，利用者の利便性を著しく損なう．

6.2 内容

本研究では，異機種間でのファイルサーバの移行時に，ファイルアクセスの停止時間を短縮するための移行方式を提案する．提案方式では，まず接続先サーバを移行先サーバに切り替える．その後アクセス要求時に応答に必要なファイルのみ移行元からオンデマンドで複製する方式と，低負荷時に未アクセスファイルをバックグランドで複製する方式を併用する．これにより，ファイルアクセスを停止すること無く時間のかかる複製をおこなえるようにする．性能評価の結果，移行に伴うアクセス停止時間を，一般的なファイルサーバのアクセス許容応答時間である 30 秒以内に短縮できることを確認した．

6.3 具体的成果

本研究の成果は、国内シンポジウム第 29 回コンピュータシステム・シンポジウム (ComSys2017) で論文として発表した。

7 JIT ROP 攻撃を防止するための軽量な保護機構に関する研究

7.1 背景

現代のブラウザには、JavaScript プログラムをネイティブバイナリコードにコンパイルする JIT（Just-In-Time）コンパイラがある。最近の JIT コンパイラは、JavaScript データを JavaScript 以外のメモリ領域に格納するため、シェルコードを JavaScript データとして入力するだけでは機能しない。この保護を克服するために、最近の攻撃では、コード領域に配置された命令に準拠した JavaScript プログラムの定数を利用し、ガジェットと呼ばれる小さなコードとして ROP（Return-Oriented Programming）でチェーン化する。この攻撃に対抗するために、最近のブラウザでは、JavaScript の定数を秘密鍵で暗号化し、実行時に復号化して攻撃者が任意のガジェットを挿入することを防いでいる。残念ながら、現在のブラウザ（Firefox、Google Chrome、および Microsoft Edge を含む）では、パフォーマンス上の理由から 2 バイト以上のブラインド定数しかないため、攻撃者は ROP 攻撃を行うのに十分な 1 バイトと 2 バイトのガジェットを生成することができる。

7.2 内容

本論文では、JIT コンパイラの高性能で安全なブラインド手法を提案する。この手法では、定数の値に基づいて定数をブラインドするかどうかを決定する。定数に制御フロー命令（たとえば texttt ret と texttt jmp）として解釈できる値が含まれている場合、2 バイト以下であってもその定数を無視する。それ以外の場合は、ガジェットとして使用できないため、定数をブラインドしない。この技術は、ブラインドされなければならない定数の数を減らすことによって、一定のブラインドのオーバーヘッドを効果的に削減する一方で、小さな定数でもガジェットとして利用される可能性を排除してセキュリティを向上させる。この技術を Microsoft Edge の JIT エンジンである ChakraCore で実装し、x64 システムで実行し、JIT エンジンのパフォーマンスを測定した。実験結果は、我々の技法がすべての定数をブラインドすることと比較して、最大 2.85%性能を改善することを確認した。

7.3 具体的成果

本研究の成果は、無査読国際会議 The 7th International Workshop on Networking, Computing, Systems, and Software で論文として発表した。

8 成果要覧

招待講演／招待論文

[招待 1] (APSys 2017) Bash on Ubuntu on macOS. Takaya Saeki, Yuichi Nishiwaki, Takahiro Shinagawa, Shinichi Honiden. 第 29 回コンピュータシステム・シンポジウム (ComSys2017), Dec 2017. 招待凱旋発表

著書／編集

[著書 1] 野口 健一郎, 光来 健一, 品川 高廣: IT Text オペレーティングシステム（改訂 2 版）, オーム社, 2018 年 1 月

査読付論文

[査読付 1] Iori Yoneji, Takaaki Fukai, Takahiro Shinagawa and Kazuhiko Kato. Unified Hardware Abstraction Layer with Device Masquerade. In Proceedings of the 33rd ACM Symposium On Applied Computing (ACM SAC 2018), Apr 2018.

[査読付 2] Ilias Avramidis, Michael Mackay, Posco Tso, Takaaki Fukai, Takahiro Shinagawa. Live Migration on ARM-based Micro-datacentres. In Proceedings of the 3rd Workshop on Edge Computing (EdgeCom 2018), Jan 2018.

[査読付 3] Takaaki Fukai, Satoru Takekoshi, Kohei Azuma, Takahiro Shinagawa, Kazuhiko Kato. BMCArmor: A Hardware Protection Scheme for Bare-metal Clouds. In Proceedings of the 9th IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2017), Dec 2017.

[査読付 4] Keiichi Matsuzawa, Takahiro Shinagawa. VM-aware Adaptive Storage Cache Prefetching. In Proceedings of the 9th IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2017), Dec 2017. Best paper candidate

[査読付 5] Takaya Saeki, Yuichi Nishiwaki, Takahiro Shinagawa, Shinichi Honiden. Bash on Ubuntu on macOS. In Proceedings of the 8th ACM SIGOPS Asia-Pacific Workshop on Systems (APSys 2017), Sep 2017.

公開ソフトウエア

[公開 1] BitVisor 2.0, http://www.bitvisor.org/, 2017 年 12 月

[公開 2] VMXbench, https://github.com/utshina/VMXbench.

その他の発表論文

[発表 1] 松沢 敬一, 早坂 光雄, 品川 高廣．オンデマンド及びバックグラウンド複製によるファイルサーバ移行時の停止時間削減．第 29 回コンピュータシステム・シンポジウム (ComSys2017)，2017 年 12 月

[発表 2] Tomoyuki Nakayama, Masanori Misono, Takahiro Shinagawa. High-performance and Secure Just-in-time Compiler Protection. The 7th International Workshop on Networking, Computing, Systems, and Software, Aomori, Oct 2017.

[発表 3] 忠鉢 洋輔, 品川 高廣, 後藤 厚宏, 加藤 和彦．カーネルに組み込まれた強制的な TLS 公開鍵ピンニング．コンピュータセキュリティシンポジウム 2017（CSS 2017），山形，2017 年 10 月．

特記事項

[特記 1] BitVisor Summit 6 開催（第 29 回コンピュータシステム・シンポジウム 併設イベント）、2017 年 12 月，http://www.bitvisor.org/summit6/