

卒業論文 2019 年度（平成 30 年度）

卒論

慶應義塾大学 環境情報学部

石川 達敬

徳田・村井・楠本・中村・高汐・バンミーター・植原・三次・中澤・武田
合同研究プロジェクト

2020 年 1 月

卒業論文 2019 年度（平成 30 年度）

卒論

論文要旨

アブスト

キーワード

OS

慶應義塾大学 環境情報学部

石川 達敬

Abstract Of Bachelor's Thesis Academic Year 2019

Title

Summary

abst

Keywords

OS

Bachelor of Arts in Environment and Information Studies
Keio University

Tatsunori Ishikawa

目次

第 1 章	序論	1
1.1	背景	1
1.2	着目する課題	1
1.3	目的	2
1.4	本論文の構成	2
第 2 章	関連技術	3
2.1	vm を用いた解析	3
2.1.1	qemu	3
2.1.2	libvmi	3
2.2	RDMA	3
2.2.1	Infiniband における RDMA 実装	4
2.3	オペレーティングシステム解析手段	4
2.3.1	kgdb	4
2.3.2	コアダンプを用いた解析	4
2.3.3	vm を用いた解析	4
第 3 章	アプローチ	5
3.1	オペレーティングシステムのコンテキスト	5
3.1.1	task_struct 構造体	5
3.1.2	オペレーティングシステムのビルドにおけるコンフィグ	6
3.1.3	ホスト自身によるレジスタやシンボルの参照	6
3.2	本研究で保持する情報	7
3.3	なければならぬ情報	7
第 4 章	実装	8
4.1	実装の概要	8
4.2	nettlp	8
4.2.1	process-list.c	8
4.3	実験環境	9
4.4	実装の前提情報	10
4.5	実装の全体	10
4.5.1	mem_dump	10

4.6	init_task およびカーネルコンフィグの取得	10
4.6.1	init_task に関する情報の復元	11
4.6.2	カーネルコンフィグの復元	11
4.7	Linux カーネルをプリプロセッサに通す	11
4.8	ていうか実装をカーネルに埋め込めばいいのでは	11
4.9	実装のまとめ	11
第 5 章	評価	12
5.1	評価手法	12
5.2	評価	12
第 6 章	まとめと結論	13
6.1	まとめ	13
6.2	結論	13
6.3	今後の課題	13
	謝辞	14

図 目 次

1.1	libvmi を用いる際のアーキテクチャ	1
2.1	PCI Express	3
3.1	Linux における mmu	6
4.1	全体	9

表 目 次

第1章 序論

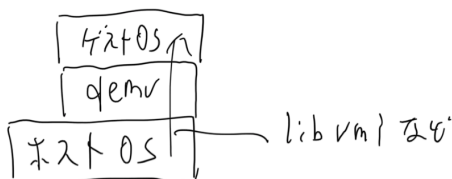
1.1 背景

コンピュータの管理者は、動作中、あるいはカーネルパニックなどによって停止したコンピュータの情報を監視・解析することが必要となる場面がある。動作中のコンピュータ自身に対しては、同一ホスト内の `top` コマンドや `ps` コマンドを用いて、プロセスの一覧を得たり、`gdb` コマンドを用いてプロセスをトレースし、プロセスの状態を把握する。論理的に停止したコンピュータに対しては、`kdump` と呼ばれる機構を通してメモリダンプを静的に解析し、原因の究明をする。また、状態を監視したいホストを物理的なマシンではなく、仮想マシンとして起動し、`qemu` や `Xen` などの基盤上で `libvmi` などを通して状態を解析する手法も存在する。

仮想マシン上に起動したホストに対する解析手段としては、上述の通り手段が豊富に揃っている。

これらの課題を解決する手法の一つとして、RDMA NIC を用いた解析手法が、(sora さんの論文で) 発表された。

図 1.1: `libvmi` を用いる際のアーキテクチャ



1.2 着目する課題

Linux カーネルのバージョン、`System.map` の情報および `config` の情報を知らなければ、メモリのみから正しくコンテキストを復元していくことはできない。

`libtlp` の論文（もうちょっとちゃんと書く）で紹介されている `process-list.c` は `System.map` を引数として渡し、`init_task` の行を read することで `init_task` の仮想アドレスを得ている。また

，カーネルコンフィグに依存するマクロの値や，使用する関数などもハードコーディングされており，論文の環境以外で動かすことが容易ではない。

1.3 目的

そこで本研究の目的として，問題の章であげた3つの情報，Linuxカーネルのバージョン，System.mapの情報およびカーネルコンフィグのうち，System.mapおよびカーネルコンフィグの情報をRDMA NICを用いて復元する．また，Linuxカーネルのバージョンを知ることさえできれば，どのようなカーネルコンフィグを持つコンピュータに対しても，RDMA NICを通してプロセスリストの一覧を取得できることを実証する．

この目的を達成するため，本研究では，いくつかの段階に分けてネットワーク越しにあるコンピュータに対して監視・解析を行っていく．第一の工程として，RDMA NICを用いて，監視対象ホストのメモリを全探索し，メモリに落ちているSystem.mapのうち，`init_task`が配置されている仮想アドレス空間に関する情報と，Linuxカーネルにおける`__phys_addr`関数，`task_struct`型を決定するためのカーネルコンフィグに関する情報を収集する．

次の工程として，収集したカーネルコンフィグを元に手元のコンピュータでLinuxカーネルのソースコードに対してプリプロセスの処理を行い，`task_struct`型を確定する．さらに，ソースコード上にある`__phys_addr`関数の実体を収集する．

最後に，この工程で得られた情報をもとに，libtlpで提唱されている手法を用いて，プロセスの一覧を正しく取得できることを確認する．

1.4 本論文の構成

こんなことは最後に書く．!!!!!!!!!!!!!!

2章では，本研究で使用するlibtlpと，その基盤として使用しているRDMAについて述べる．

??章では，RDMAを通してネットワーク越しにあるコンピュータを監視・解析を行うための実行環境の構成に関して述べる．

4章では，本研究で実装したものについて述べる（加筆）

5章では，本研究における評価として，Linuxカーネルのバージョンのみが与えられた状態でプロセスリストの一覧を取得できることを示す．

6章では，本研究に関する結論と，今後の課題について述べる．

第2章 関連技術

本章では、本研究で基盤技術として使用する RDMA (Remote Direct Memory Address) に関して述べる。

2.1 vm を用いた解析

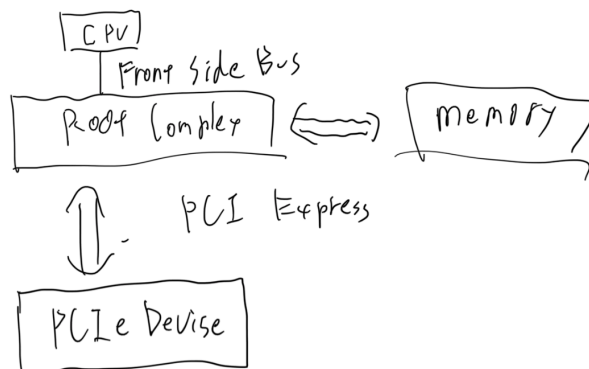
2.1.1 qemu

2.1.2 libvmi

2.2 RDMA

ここに RDMA の説明をかく

図 2.1: PCI Express



2.2.1 Infiniband における RDMA 実装

2.3 オペレーティングシステム解析手段

2.3.1 kgdb

2.3.2 コアダンプを用いた解析

2.3.3 vm を用いた解析

(1) libvmi

第3章 アプローチ

1.3で、本研究の目的を、動作中のコンピュータのメモリのダンプをリアルタイムで解析することで、コンピュータの状態をリモートホストから知ることができるようにする、と定義した。

そこで本章では、メモリのダンプをリアルタイムで取得・解析する上で前提となる情報と、この手法における課題について述べる。

3.1 オペレーティングシステムのコンテキスト

コンピュータの状態、すなわちオペレーティングシステムの動作中におけるホストの状態は、コンピュータ内部におけるレジスタの値および、内部から参照できる仮想アドレス空間上に保持されている。その例を下に示す。

あるプロセスを実行する際に、プロセッサはインストラクションポインタレジスタの命令を読み込み、逐次実行をしていく。call 命令などで別の関数を呼ぶ際には、その時点におけるインストラクションポインタレジスタの値をメモリ上に退避し、関数が終わった際に、呼び出し元に返るように設定されている。実行コードが整合性を保っているかは、実行可能ファイルを生成したコンパイラの責務なので、本論文では述べないが、プロセッサはプログラムの実行を行う際、レジスタの値を参照、退避、復帰、上書きさせることで、状態を保持、進行させていると言える。

プロセッサがレジスタの値を参照しそれを仕組みは、カーネルのコードを実行する際にも言える。ここでは、オペレーティングシステムから見たコンピュータの状態として、プロセスの切り替え処理、コンテキストスイッチにおける処理の流れを述べる。コンテキストスイッチとは、割り込み処理などによって定期的に呼ばれるプロセススケジューラから呼ばれる機構である。この機構は、実行中のプロセスの状態、すなわち、各レジスタの値および仮想アドレス空間に関する情報などをカーネルが管理しているメモリ上にあるデータ構造の中に退避する。

本セクションのまとめとして、コンピュータの状態は、ある瞬間においてはレジスタの値であり、この状態を保存する際は、メモリ上にレジスタの値を退避させていることを述べた。

3.1.1 task_struct 構造体

3.1でコンテキストスイッチにおいて、退避されるプロセスの情報は、対応したデータ構造に退避されると述べたが、この時に使用されるデータ構造がtask_struct 構造体である。task_struct 構造体には、一つのプロセスの情報の全て(?)が格納されている。その中には、仮想アドレス空間に関する情報を保持するmm_struct 構造体を参照するフィールドも存在する。

コンテキストスイッチ時には、`task_struct` 構造体に保持されている情報をレジスタに復帰させることで、中断される直前の情報を復元している。

3.1.2 オペレーティングシステムのビルドにおけるコンフィグ

`task_struct` 構造体をはじめとして、Linux カーネルの変数や型、関数は、様々なアーキテクチャやカーネルコンフィグに対応するため、マクロによって分岐されている。この分岐が確定するのは、Linux カーネルをビルドするときであり、構造体のメンバへのアクセス、関数のアドレスなどはコンパイラが保証している。

実際のカーネルのバイナリは、`vmlinux` としてコンパイルされた後、`strip` され `bzImage` となる。ユーザーが作成したカーネルモジュールなどで関数を呼び出す際は、シンボルとアドレスの変換表である `/boot/System.map` を参照し、仮想アドレスを得たのち、実際にメモリにアクセスする際に物理メモリアドレスを算出しアクセスしている。

3.1.3 ホスト自身によるレジスタやシンボルの参照

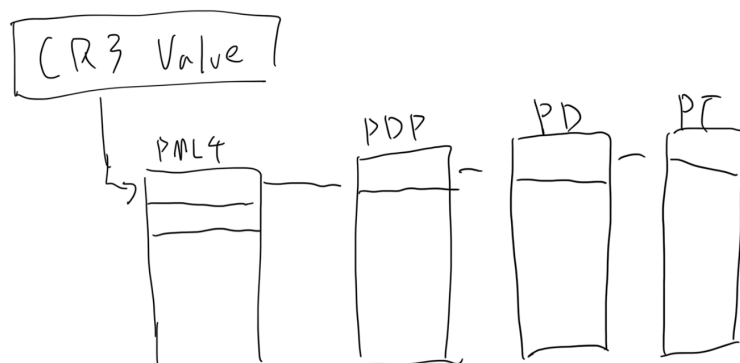
書き直し！！

3.1 で述べたように、オペレーティングシステムでは、レジスタの値などを退避する際、そのプロセッサ自身が `'push'` 命令などを用いてメモリにアクセスできる

レジスタを参照して、かつ `mmu` すら自分で参照ができる。

CPU レジスタの現在の値は直接知ることができないため、例えばプロセスの一覧を取得したい場合は、コンテキストスイッチ時に退避された値を辿っていく必要がある。しかし、上述の通り `task_struct` はビルドされた際のカーネルコンフィグによって、どのメンバが先頭アドレスからどのオフセットに保持されているかは変動する。

図 3.1: Linux における `mmu`



3.2 本研究で保持する情報

これまでで述べたように，本研究では，メモリダンプを局所的に取得し，リアルタイムで解析することで，物理的に異なるコンピュータからオペレーティングシステムのコンテキストを復元することを目的と設定した．

3.3 なければならぬ情報

オペレーティングシステムの状態を保持しているものとして，3.1 で述べたようにレジスタがある．しかし，3.2 で述べたように，現在のレジスタの値は，メモリから知ることはできない．

また，退避された値は，メモリから読み出しを行うが，その際には，

第4章 実装

4.1 実装の概要

本研究では，RDMA を用いて，動作中のマシンのメモリの値を取得していくことで，リモートホストから監視対象ホストのオペレーティングシステムのコンテキストを復元していくことを目指す．この目的を実現するために，本研究では，nettlp を用いて実験を行う．

4.2 nettlp

nettlp 本来の目的は，PCIe デバイスの開発プラットフォームである．（リファレンス）

その機能の一つとして，DMA message と ethernet パケットを相互変換する機能がある．2 で述べたが，RDMA の Infiniband 実装は，制限が多い．（もう少し詳しく）nettlp における RDMA では，物理アドレスを指定することで，任意のバイト数（なんバイトだっけ？）の値を取得することが可能である．また，アクセスできないメモリアドレスは存在しない，すなわち全メモリアドレス空間から値を取得することが可能となっている．

nettlp は FPGA ボード上で動作するものであり，これを利用するためのインターフェースとして，libtlp が用意されている．libtlp では，RDMA を用いてメモリダンプを取得するためのインターフェースが関数として用意されている．この関数を含んだヘッダファイルを include し，プログラムから呼び出すことで，メモリアドレスの値が返ってくる．

用意されている関数は，`dma_read` 関数と `dma_write` 関数の二つである．`dma_read` 関数は，値を読みだすための関数であり，呼び出す際に読みたいメモリアドレスを渡す．`dma_write` 関数は，値を指定した物理アドレスに書き込むための関数であり，呼び出す際に，書き込みたいメモリアドレスと値を渡す．

本研究では，`dma_read` 関数のみを用いる．

4.2.1 process-list.c

このファイルでやっていることをかく．本研究の実装が，この `process-list.c` を拡張したものであるということを書く．

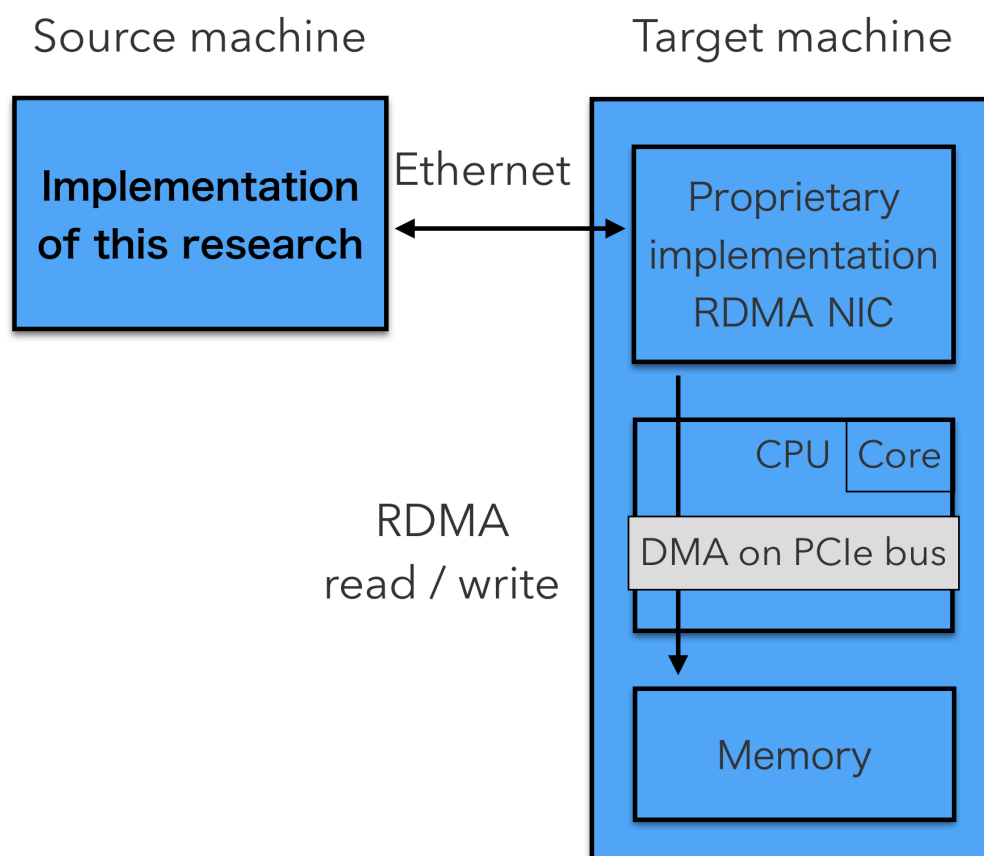
4.3 実験環境

本研究で実装を行う環境は，図 4.1 にあるように，nettlp が書き込まれた FPGA が刺さった監視対象ホストと，本研究における実装を実行するホストの 2 台で構成する．

監視対象ホストは，Linux 4.15.0-72-generic の ubuntu であり，PCIe デバイスとして，nettlp が書き込まれた FPGA ボードが刺さっている．本研究では，FPGA ボードとして，ザイリンクスのやつを使用している．また，この FPGA ボードは，ネットワークインターフェースでもあり，IP アドレスとして，192.168.10.1 を静的に振ってある．

実装を実行するホストは，Linux 4.19.0-6-amd64 の Debian buster であり，光ファイバーケーブル (名称はあとで修正) が刺さる NIC を刺している．以後，実装ホストと呼称する．この NIC には IP アドレスとして，192.168.10.3 を静的に振ってある．監視対象ホストに対して RDMA を実行する際は，`dma_read` 関数，あるいは `dma_wirte` 関数を通して 192.168.10.1 に対して IP パケットを送信している．

図 4.1: 全体



4.4 実装の前提情報

本研究では、Linux カーネルのバージョンは、実装ホストは知っている情報とする。で述べたもののうち、Linux カーネルのバージョンだけは一旦 Given でやる。

4.5 実装の全体

ダンプしてくるということを書く。物理アドレスのマッピングに関しても書く

実験における第一段階として、4.4 で述べたように、監視対象ホストのカーネルコンフィグおよび `init_task` の仮想アドレスを知ることを目指す。そこで、本研究では、この情報をメモリ上から探す。4.6.1 で述べる実装では、取得できるメモリダンプを全て取得し、解析する手法に関して述べる。

次の工程として、それをリストアップ。

次の工程として、収集したカーネルコンフィグを元に手元のコンピュータで Linux カーネルのソースコードに対してプリプロセスの処理を行い、`task_struct` 型を確定する。さらに、ソースコード上にある `__phys_addr` 関数の実体を収集する。

最後に、この工程で得られた情報をもとに、libtlp で提唱されている手法を用いて、プロセスの一覧を正しく取得できることを確認する。

4.5.1 mem_dump

第一の工程として、メモリの全ての情報を取得する。ソースコードは以下である。この実装を実装ホストで実行し、出力結果をファイルに格納する。この実装では、libtlp を通して、監視対象ホストのメモリを全探索する。この実装の実行には長い時間（何分？）かかるため、アトミックな情報ではない。そのため、ここで取得したメモリダンプは、解析には使えない。ここで取得したメモリダンプは、`System.map` のうち、`init_task` が配置されている仮想アドレス空間に関する情報および、Linux カーネル 4.15.0 におけるカーネルコンフィグに関する情報を収集するためのものである。

実行方法

```
./dump_mem > ~/Desktop/work9/dump
```

4.6 init_task およびカーネルコンフィグの取得

この章では、??で取得した、メモリダンプから、必要な情報を取得する。前処理として、以下に示すように、メモリダンプに対して `strings` コマンドをかけ、検索しやすくする。以後のファイル読み込みには、`str_list` を用いる。

strings

```
strings dump > str_list
```

4.6.1 init_task に関する情報の復元

で作成した `str_list` に対して, `init_task` に関する行を `grep` コマンドを用いて探す.

4.6.2 カーネルコンフィグの復元

4.7 Linux カーネルをプリプロセッサに通す

この工程では, 収集したカーネルコンフィグを元に手元のコンピュータで Linux カーネルのソースコードに対してプリプロセスの処理を行い, `task_struct` 型, および `__phys_addr` 関数など, `process-list.c` の影響のあるソースコードを確定する.

さらに, ソースコード上にある `__phys_addr` の実体を収集する部分に関する実装をより詳しく書く.

makeoldconfig

```
cd path/to/kernelsource
cp path/to/さっき拾った/config path/to/kernelsource/.config
make oldconfig
# save-temps オプションを設定
make -j9
```

4.8 ていうか実装をカーネルに埋め込めばいいのでは

最後に, 集められたデータをもとに, `process-list` を改造したものに関する説明をここに書く

4.9 実装のまとめ

この章は, 相当長くなるので, まとめを書く.

第5章 評価

5.1 評価手法

カーネルのバージョンのみわかる状態から，正しく `ps aux` と同じような出力を得られるかどうかを評価とする．

5.2 評価

未評価

第6章 まとめと結論

アブスト

6.1 まとめ

各章の振り返り

6.2 結論

結論は俺の知見，考察．結局こうだったを述べる．こういうところでは使えた，こういうところではダメだった．

6.3 今後の課題

課題はたくさんある

謝辞

アドバイスをくれた全員に感謝