シグナルの疑似送信による プロセスレベル障害からの迅速な復旧

木村　健人[1]　光来　健一[1]

1. はじめに

近年の大規模かつ複雑なシステムにおいてシステム障害を回避するのは難しい．しかし，システム障害が発生するとシステム上で動作しているサービスの品質が低下したり，完全に停止したりするため，サービスの利用者や提供者は大きな損失を被る．そのため，システムに障害が発生した場合には迅速に障害を検知して復旧を行うことが重要である．障害からの復旧を行うにはシステムにアクセスして作業を行う必要があるが，システムが応答しなければハードウェアリセットによって復旧するしかない．しかし，強制的にシステムのリセットを行うとデータが失われ，復旧に時間やコストがかかる可能性がある．本稿では，GPU 上の復旧システムが OS を間接的に制御することでシステム障害からの復旧を行う GPUfas を提案する．

2. GPUfas

GPUfas では，障害発生時に GPU からメモリ上の OS データを書き換え，OS 自身の機能を用いて障害の原因を取り除く．GPUfas のシステム構成を**図 1** に示す．GPU は OS が動作する CPU やメインメモリから物理的に隔離されており，復旧対象システムの障害の影響を受けにくい．GPU は専用の演算コアやメモリを持っているため，復旧対象システムのリソース不足の影響を受けない．また，システム起動時に GPU を占有して GPUfas を実行することにより，システム障害発生後に復旧システムが実行できないという事態を避けられる．

一例として，GPUfas はシステム障害の原因となっているプロセスへのシグナルの疑似送信により，プロセスレベル障害からの復旧を行うことができる．OS はシグナルを送ることによってプロセスを制御することができるが，

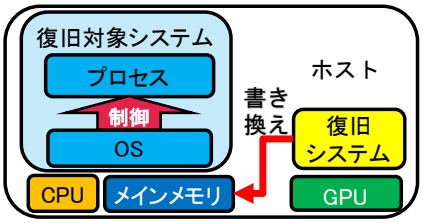

図 1　GPUfas のシステム構成

GPU から直接プロセスにシグナル送信を行うことはできない．そこで，GPUfas がメインメモリ上にあるプロセス情報を書き換えてシグナルが送られた状態に変更することで，プロセスへのシグナル送信を疑似的に実現する．疑似送信されたシグナルは OS によって処理され，プロセスに配送される．

また，仮想マシン（VM）内のシステム障害に対しても，VM 内のゲスト OS を間接的に制御することによって復旧を行う．この場合には，VM イントロスペクションを拡張して VM 外から OS データの書き換えを行う．VM イントロスペクションは本来，VM のメモリにアクセスして VM 内の情報を取得する技術であるが，GPUfas ではさらに VM 内の情報の書き換えも行う．このように，VM 内のシステムの復旧には GPU は不要である．

3. 実験

GPUfas の有用性を確かめるためにメモリが不足する障害からの復旧時間を測定した．比較として，別の端末から SSH で接続し，pkill コマンドを実行した時の復旧時間も測定した．復旧対象ホストには**表 1** のマシンを用い，VM を復旧させる実験では**表 2** の VM を用いた．

大量のメモリを使用するプロセスを強制終了させるのにかかった時間は**図 2** のようになった．GPUfas はシステムのメモリ不足を検知して障害の原因となったプロセスを特定した後，624ms で KILL シグナルを疑似送信してそのプ

[1] 九州工業大学
KyushuInstituteofTechnology

表 1　復旧対象ホスト

OS	Linux 4.18.0
CPU	Intel Core i7-8700
メモリ	DDR4-2667 16GB
スワップメモリ	5GB
GPU	NVIDIA GeForce GTX 960
ソフトウェア	NVIDIA GPU driver 430.40 CUDA 10.0.130 LLVM 8.0

表 2　復旧対象 VM

ゲスト OS	Linux 4.15.0
割当メモリ量	2048MiB
CPU 数	1 個
割当ディスク量	50GiB
ソフトウェア	QEMU-KVM 2.11.2

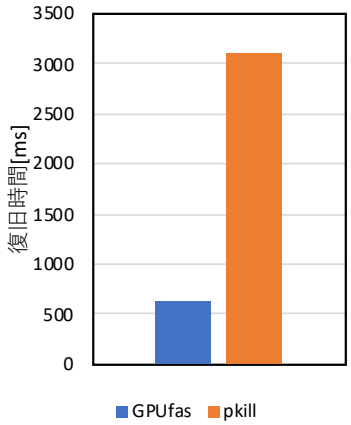

図 2　メモリ不足からの復旧時間

ロセスを強制終了させることができた．これによりシステムのメモリ不足が解消され，障害からの復旧を行うことができた．一方，pkill コマンドを用いた場合，スラッシングが発生してプロセスを強制終了させるのに 3 秒以上の時間がかかった．このことから，GPUfas のほうが 5 倍高速に復旧を行えることが分かった．

次に，大量のプロセスに疑似的にシグナルを送信して終了させるのにかかる時間を測定した．シグナルの疑似送信を GPU から行った場合の時間を**図 3**，VM イントロスペクションを用いて行った場合の時間を**図 4** に示す．SSH でログインして pkill コマンドを実行する方法と比較すると，定期的に実行するスリープ時間が短いプロセスの場合には高速化したが，スリープ時間が長いプロセスの場合には大幅に遅くなった．これは，スリープしている間はプロセスを終了させることができていないためである．一方，VM 内のプロセスの場合にはスリープ時間が長い場合でも GPU からシグナルを送信するより大幅に高速に終了させられることがわかった．

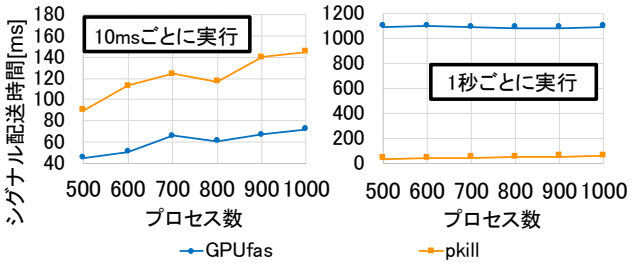

図 3　GPU からプロセスへのシグナル配送時間

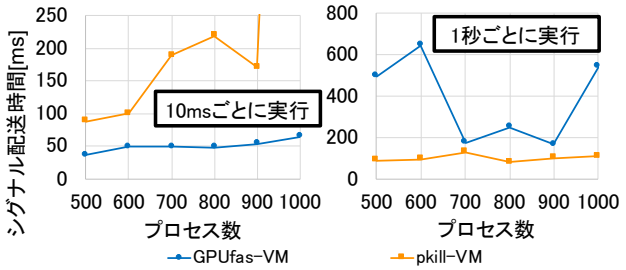

図 4　VM 内のプロセスへのシグナル配送時間

4. まとめ

本稿では，GPU 上で OS の挙動を間接的に変更することでシステム障害からの復旧を行う GPUfas を提案した．GPUfas では，GPU 上で動作する復旧システムがメインメモリ上の OS データを書き換え，OS 自身の機能を用いて障害の原因を取り除く．一例として，プロセスにシグナルを疑似的に送信することで，プロセスによって引き起こされた CPU 過負荷やメモリ不足などの障害からの復旧を可能にした．

今後の課題は，スリープしているプロセスを即座に終了させられるようにすることである．OS データの書き換えだけでは難しいため，OS 内に用意した最小限の復旧支援機構を呼び出すことによって実現することを計画している．また，シグナルの疑似送信以外のプロセスレベル障害からの復旧手法やカーネルレベル障害からの復旧手法を検討する必要がある．