

LKSTによる Linuxカーネルの評価と ボトルネック解析

(株)日立製作所
システム開発研究所

平松 雅巳, 杉田 由美子





目次

- カーネル性能評価手法
- LKSTとは
- LKST Log Toolsとは
 - 性能解析ツール
 - 可視化ツール
 - 使い方
- IOボトルネック解析例
 - lozoneによるボトルネック例
 - リクエストキュー長解析
 - スピンロック解析
- LKSTのオーバヘッド



カーネル性能評価手法

■ 性能評価手法

□ ベンチマーク

- アプリケーションレベルの情報のみ

□ サンプリング

- 実行頻度のみ(前後情報なし)
- サンプリングに漏れる場合が存在

→カーネル内部の詳細な情報が欲しい

■ カーネルトレーサを利用

□ 主な処理を確実に記録可能



LKST (Linux Kernel State Tracer) とは

■ Linuxカーネル向けイベントトレーサ

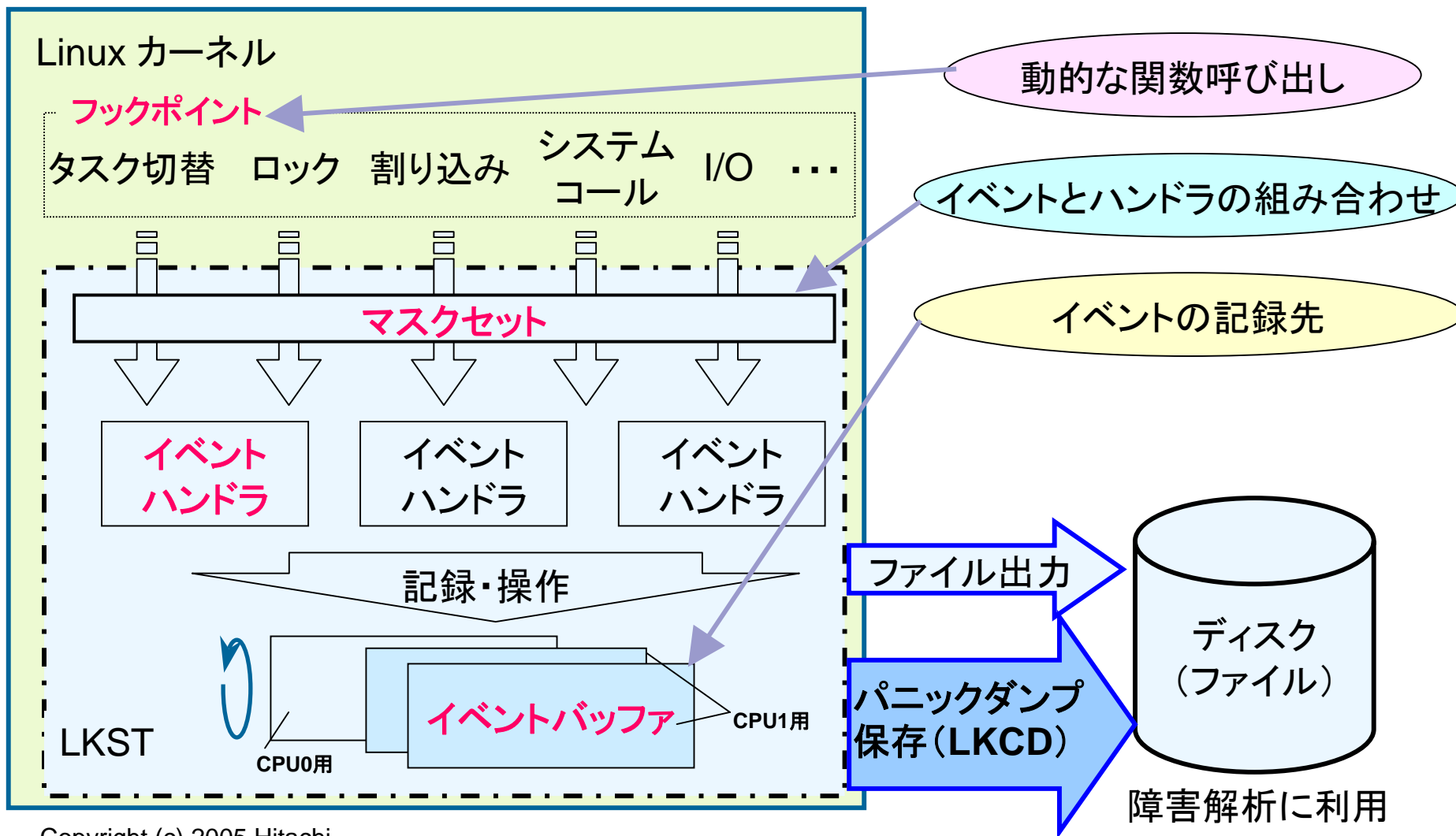
- カーネルドライバ(本体と組み込みイベントハンドラ)
- カーネルコードに埋め込まれたフックポイント
- ユーザコマンド群
- 拡張イベントハンドラ

■ 特徴

- イベントトレース時の柔軟な**カスタマイズ設定**
 - 記録するイベントの種類
 - イベント記録時の処理内容
 - イベントの記録内容
 - 記録領域



LKSTの動作図





LKST Log Toolsとは

■ LKSTによるカーネル性能評価機能

- LKSTの機能拡張部分
- 性能データの解析ツール
- 解析結果のプロットツール

■ 特徴

- 性能情報ごとに**個別の性能指標値**を算出
 - 単一の情報だけではない
- データの**切り出し・表示形式の変更**が可能
 - フィルタ・フォーマッタの切り替え
- 解析結果を**プロット**可能
 - PS/PDF形式でプロットデータを保存

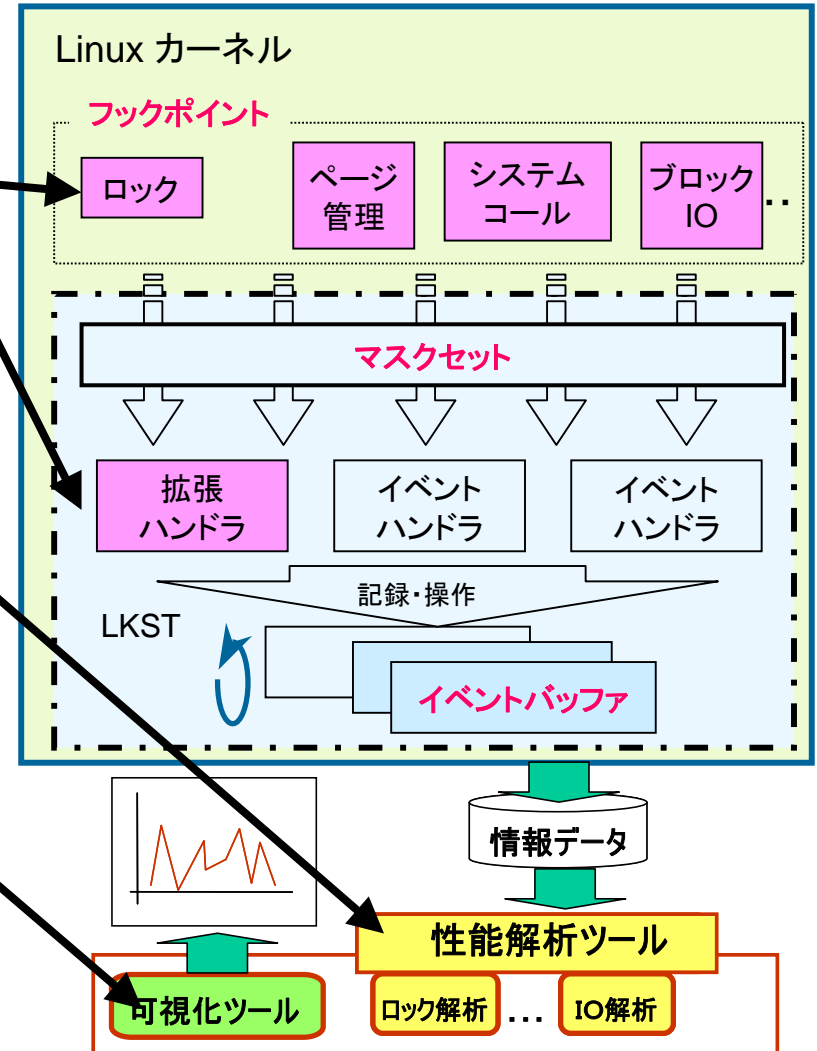


LKSTによる性能評価機能

■ 性能評価ポイントで情報を取得する
「LKST機能拡張」

■ 取得した情報を解析する
「性能解析ツール」

■ 解析した情報を可視化する
「可視化ツール」

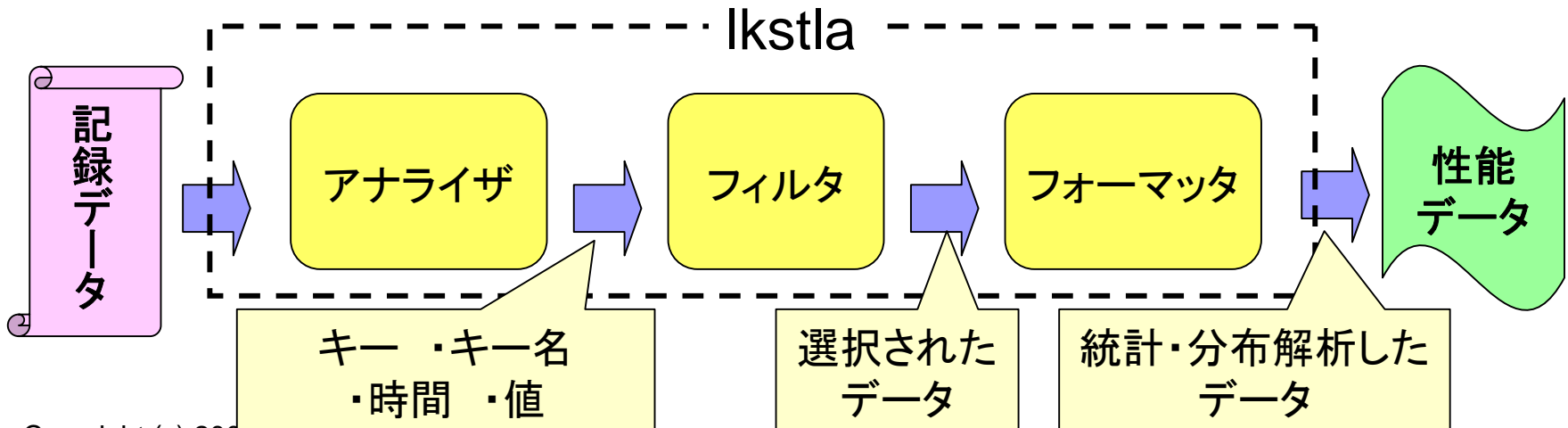




性能解析ツール(ikstla)

■ 3つの機能を持っている

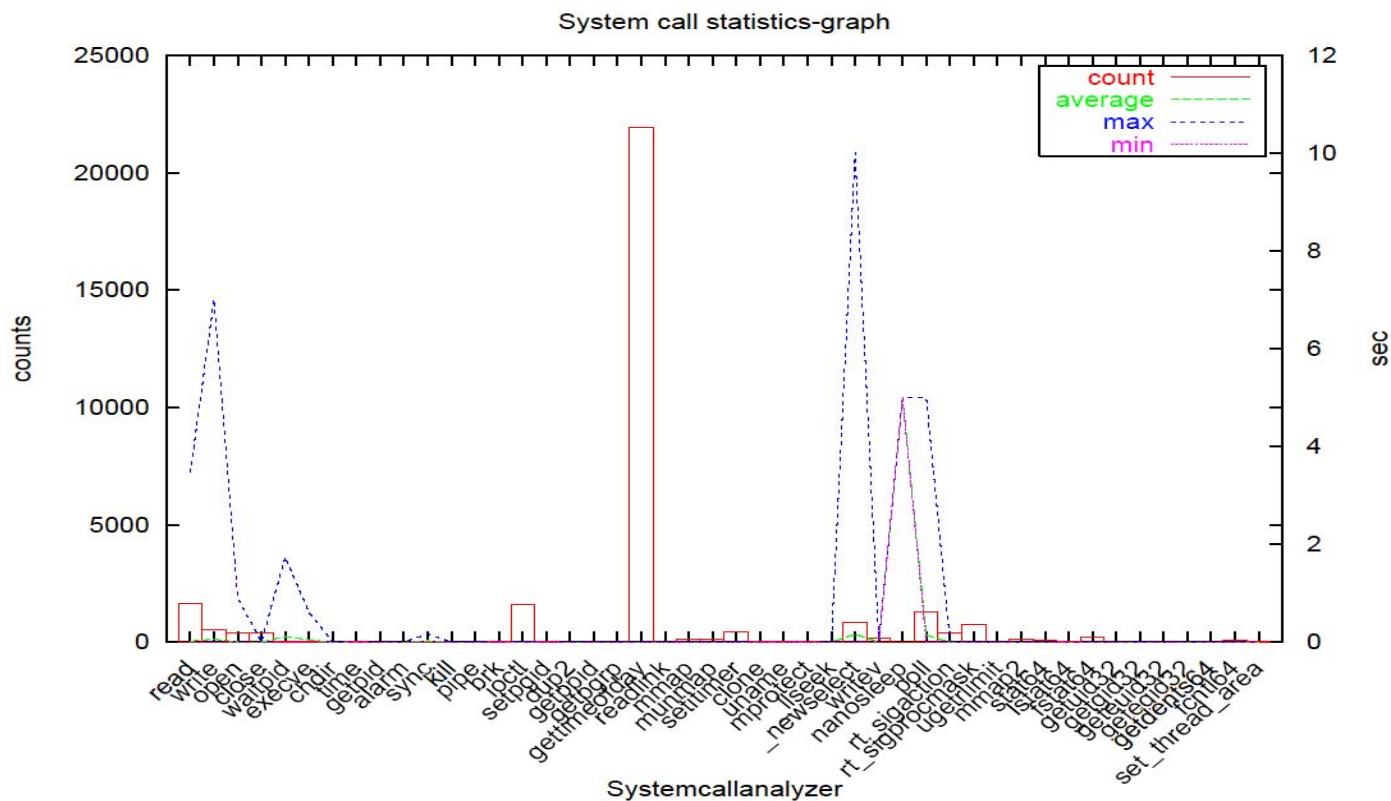
- アナライザ
 - LKSTのログファイルを解析して、性能を示す値を抽出する。
- フィルタ
 - いくつかの条件によって、結果を選り分ける。
- フォーマッタ
 - 解析結果の統計や分布などを表示する。





可視化ツール

- lkst_plot_log/lkst_plot_dist/lkst_plot_stat
 - lkstlaの解析結果をpdfにしてプロット





使い方

■ デモ

- マスクセットによる記録設定
- バッファからのデータ保存
- lkstlaによるデータの解析
- lkst_plot_statを使ったデータの表示



IOボトルネック解析例

- lozoneベンチマークによるMIRACLE LINUXの性能評価
- OProfileを使ったボトルネック箇所の特定制
- LKST Log Toolsを使ったボトルネック解析

※ミラクル・リナックス社発表の資料に基づく



IOボトルネック測定

■ 目的

- システムの限界性能調査
- ボトルネックの発見と原因解析

■ Iozone

- ファイルIO性能測定ベンチマーク
 - 並列化処理(プロセス / スレッド)
 - IO方法(read/write/random/mmap/etc.)
 - 同期方式(通常(非同期) / O_SYNC(同期))
 - 処理ディスク数 1~4
 - 処理プロセス数 1~4



測定対象

■ PCサーバ

- Dual Xeon (Hyper-Threading)

 - 論理CPU数: 4個

- 6GBメモリ

- SCSIドライブ

■ ディストリビューション

- MIRACLE LINUX V3.0

 - カーネル2.4ベース

■ ファイルシステム

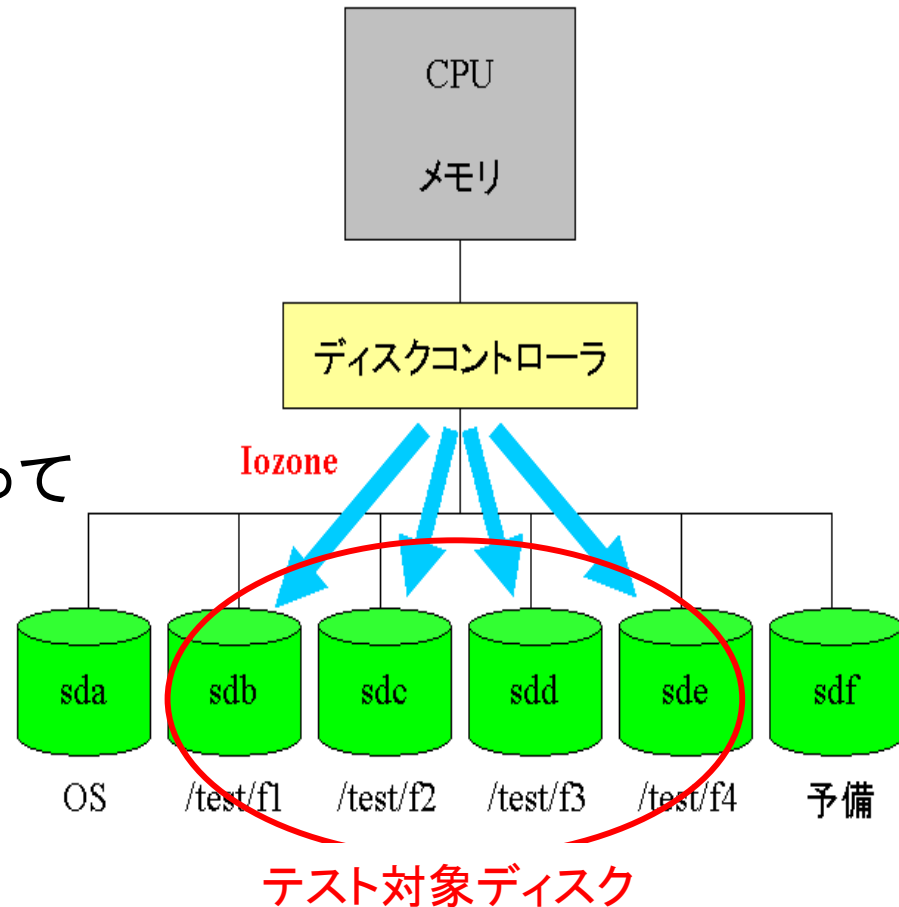
- Ext3ファイルシステム

 - 事実上、Linuxの標準ファイルシステム



ベンチマーク環境構成

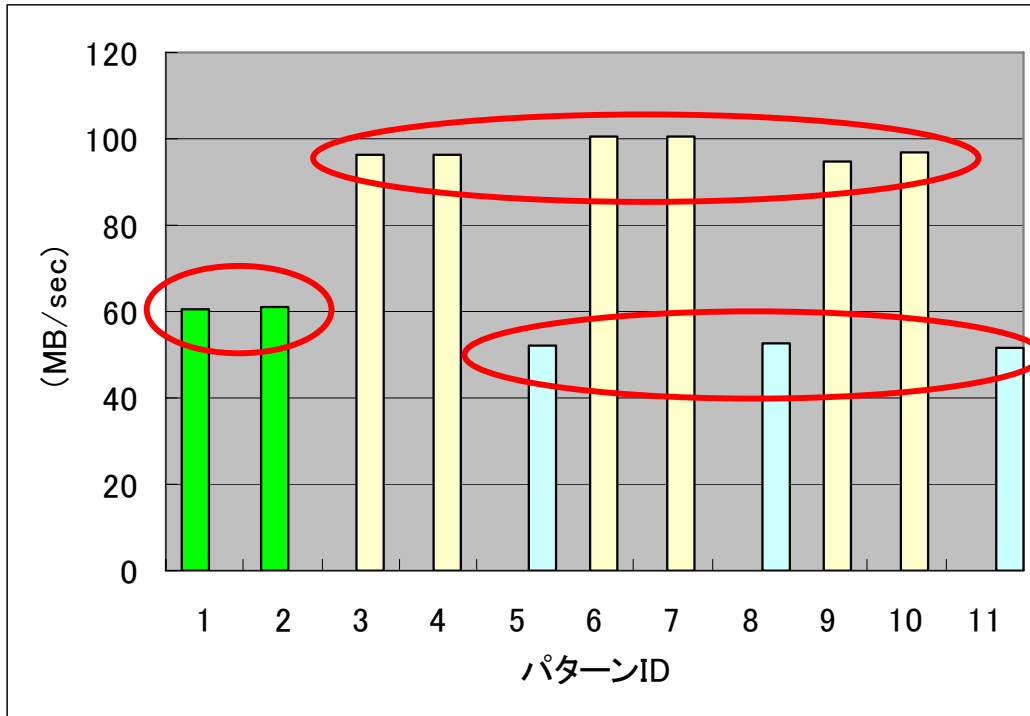
- ディスクコントローラ
 - Ultra320 SCSI
- SCSI HDD
 - 1～4台使用
 - システムHDDと別
 - テストパターンによって増減





ベンチマーク結果

■ writeスループット測定結果



- グループA:**
 CPU使用率が高い
 最大スループット

- グループB:**
 CPU利用率が低い
 低いスループット

- グループC:**
 CPU利用率が高い
 最低スループット

→ **グループBとCにおいて性能が劣化した原因を探る**



lozoneパラメータ

| パターンID | ファイルサイズ | 使用ディスク数 | スレッド使用 | lozoneコマンド数 | |
|--------|---------|---------|--------|-------------|-------|
| 1 | 8G | 1 | × | 1 | グループB |
| 2 | 8G | 1 | ○ | 1 | |
| 3 | 4G | 2 | × | 1 | グループA |
| 4 | 4G | 2 | ○ | 1 | |
| 5 | 4G | 2 | × | 2 | |
| 6 | 2731M | 3 | × | 1 | |
| 7 | 2731M | 3 | ○ | 1 | グループC |
| 8 | 2731M | 3 | × | 3 | |
| 9 | 2G | 4 | × | 1 | |
| 10 | 2G | 4 | ○ | 1 | グループC |
| 11 | 2G | 4 | × | 4 | |



解析手順

■ ボトルネック解析の手順

1. プロファイリングによる **大まかな絞り込み**
 - OProfile(*)を利用
2. LKSTを用いて **詳細なデータ**を取得
3. LKSTLogToolsを用いて **データを解析**
4. 解析結果からの考察

* OProfile:

サンプリングベースのプロファイリングツール
設定したサンプリング間隔での実行箇所の特定を行う



グループBの解析(1)

OProfileによる解析

■ グループBにおける実行箇所への絞り込み

| 順位 | 関数 | カウント | 占有率 | 累積占有率 |
|----|---------------------|-------|-------|-------|
| 1 | default_idle | 1,993 | 24.37 | 24.37 |
| 2 | try_to_free_buffers | 532 | 6.504 | 30.87 |
| 3 | launder_page | 532 | 6.504 | 37.38 |

| アドレス | サンプル数 | default_idle内占有率 | 命令 |
|------------|-------|------------------|----------------------|
| 0xc0109254 | 3 | 0.1505 | je 0x0109280 |
| 0xc0109260 | 1 | 0.05018 | movl 0x14(%edx),%eax |
| 0xc0109268 | 1945 | 97.59 | hlt |
| 0xc0109269 | 44 | 2.208 | ret |

hlt命令が多くサンプリングされている→OSの待ち時間が多い

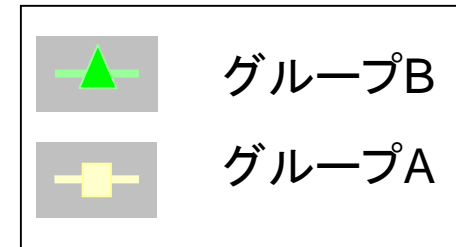
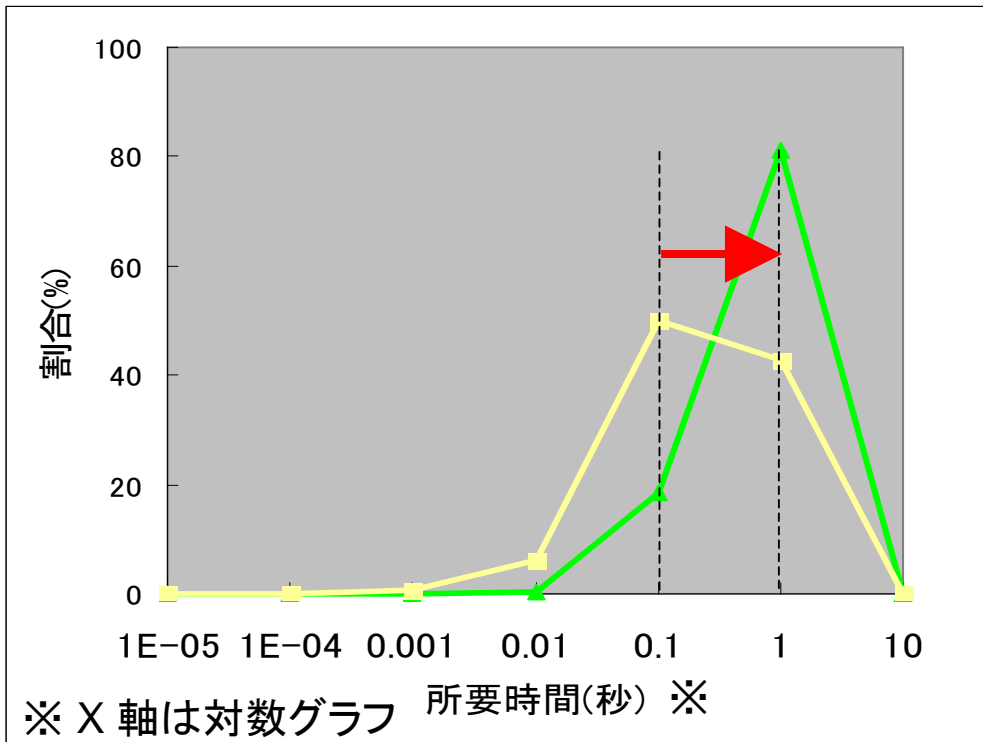


グループBの解析(2)

LKSTによるIO処理の解析(1)

■ 待ち時間処理→IO処理

□ IO処理時間をLKSTで解析



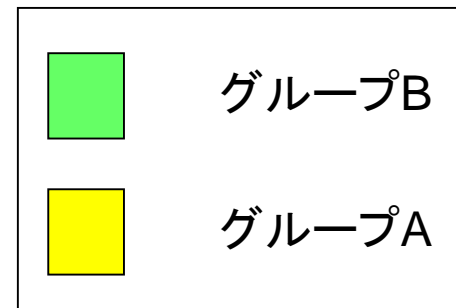
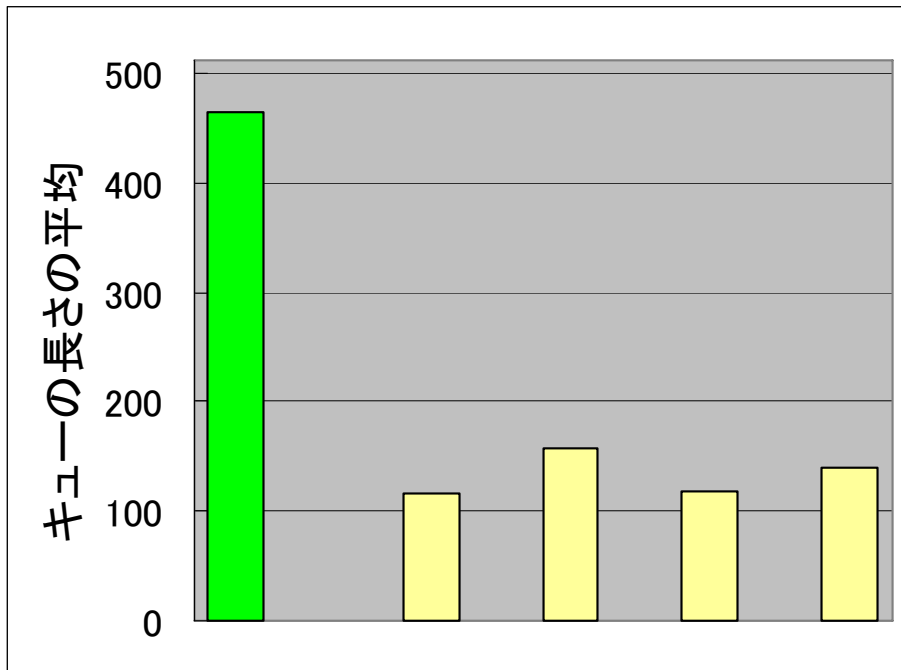
→グループBのIO処理時間がグループAより一桁大きいところにピークがある



グループBの解析(3)

LKSTによるIO処理の解析(2)

■ IOリクエストキューの長さをLKSTで解析



→グループBのキューの長さが
ほとんど限界値(=512)



グループBの解析結果

- Oprofileによる解析
 - hlt命令が多い
- LKSTによるIOリクエストキューの解析
 - ディスクが少ない
 - **並列処理が出来ず、キューが長くなる**
 - IO処理が遅い
 - CPU処理に空き時間発生 → **hlt命令実行**



グループCの解析(1)

OProfileによる解析

■ グループCにおける実行箇所の絞り込み

| 順位 | 関数 | サンプル数 | 占有率(%) | 累積占有率(%) |
|----|-----------------------|--------|--------|----------|
| 1 | .text.lock.ioctl | 10,310 | 28.74 | 28.74 |
| 2 | .text.lock.buffer | 8,257 | 23.02 | 51.76 |
| 3 | default_idle | 2,166 | 6.038 | 57.80 |
| 4 | do_generic_file_write | 2,048 | 5.709 | 63.50 |

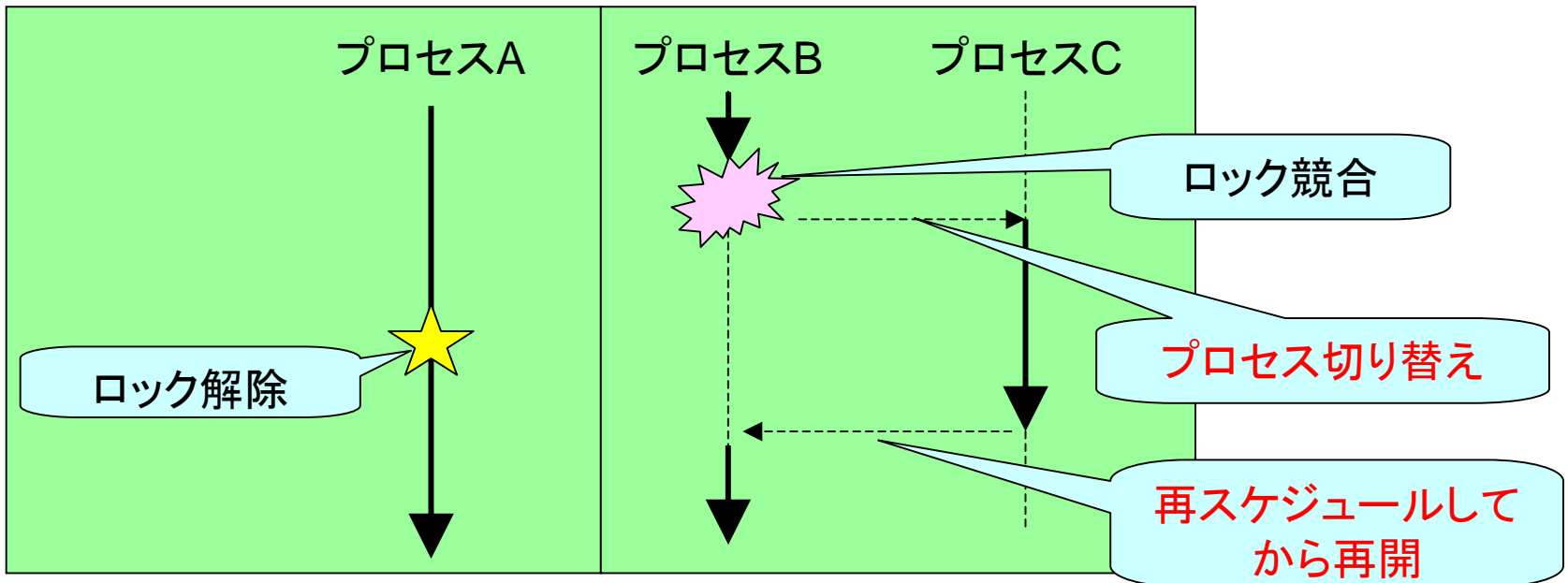
ビジーウェイト処理が多くサンプリングされている → **ロックに問題**



ビジーウェイト処理とは(1)

■ 通常のロック待ち処理

□ 例) セマフォ、mutexなど



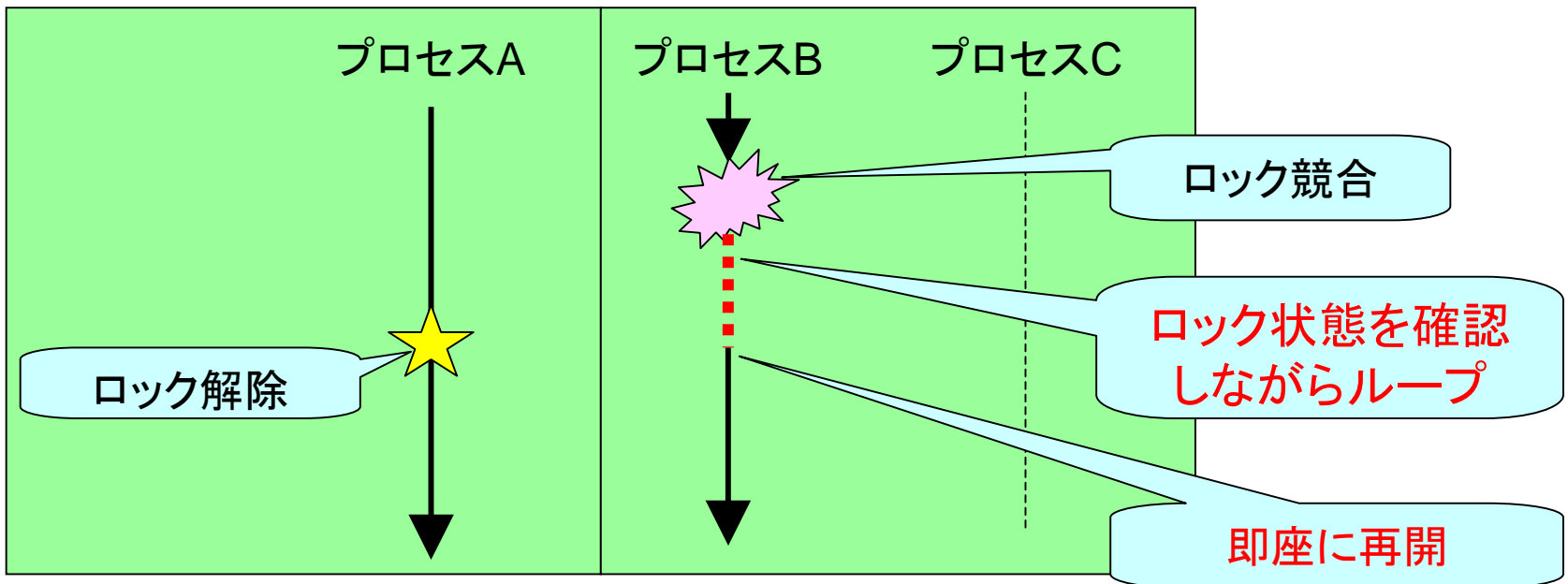
プロセス切り替えを行う→CPU処理時間を無駄にしない



ビジーウェイト処理とは(2)

■ ビジーウェイト処理

□ 例) カーネル内スピンロック



プロセスを切り替えずに待つ→CPU処理時間が無駄になる



グループCの解析(2)

LKSTによるビジーウェイト解析(1)

■ ビジーウェイト処理時間をLKSTで解析

| 順位 | アドレス | 回数 | 平均(秒) | 最大(秒) | 合計(秒) | 割合(%) |
|----|------------|--------|----------|----------|--------|--------|
| 1 | 0xc019f83e | 85 | 0.3777 | 16.14 | 32.1 | 41.1 |
| 2 | 0xc01977d1 | 1,028 | 0.02984 | 15.51 | 30.7 | 39.3 |
| 3 | 0xc018ad07 | 4 | 3.786 | 15.14 | 15.1 | 19.4 |
| 4 | 0xf8863ba3 | 61,477 | 1.8E-07 | 2.49E-05 | 0.0111 | 0.0142 |
| 5 | 0xf88637bb | 61,476 | 1.79E-07 | 7.42E-06 | 0.0110 | 0.0141 |

全ビジーウェイト処理の99.8%が3箇所集中



グループCの解析(3)

LKSTによるビジーウェイト解析(2)

■ ビジーウェイト処理時間の分散を見る

| アドレス | ~1 マイ クロ 秒 | ~10 マイク ロ秒 | ~100 マイク ロ秒 | ~1ミリ 秒 | ~10ミ リ秒 | ~100 ミリ秒 | ~1秒 | ~10秒 | 10秒以 上 |
|------------|---------------------|------------------|-------------------|-----------|------------|-------------|-----|------|-----------|
| 0xc019f83e | 82 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 0xc01977d1 | 963 | 48 | 11 | 3 | 1 | 0 | 0 | 0 | 2 |
| 0xc018ad07 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

非常に長いビジーウェイトが何度かおきている ←



グループCの解析結果

■ Oprofileによる解析

- グローバルカーネルロックのサンプリング頻度が高い→ロック回数が多い or ロック時間が長い

■ LKST

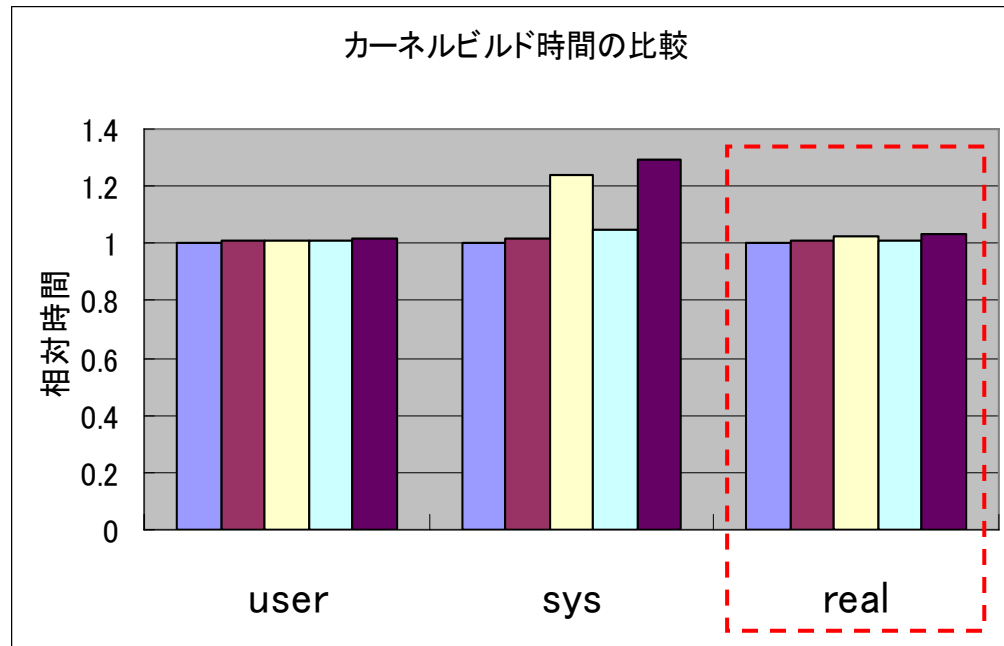
- 実際に10秒以上のビジーウェイト処理の発生を確認→ロック時間が長い



LKSTのオーバヘッド(kernel build)

■ 評価環境

- Pentium4 Xeon 2.8GHz (L2:1MB) (Hyper Threading有効)
- Memory: 4GB
- Linux 2.6.9+LKST 2.2.1 on Fedora Core 2



- 元カーネル
- 記録しない
- 性能記録
- 性能記録-spinlockなし
- 全て記録

(測定環境: IPA提供)

アプリケーション実実行性能(real)への影響は0.5~3.5%程度



まとめ

■ まとめ

- LKSTを使用した性能評価機能の開発
 - LKSTの機能を拡張し、ログから性能情報を解析する機能を開発
- OProfileとLKSTを連携したカーネルボトルネックの解析
 - IO高負荷時のカーネル処理のボトルネックを解析

■ 今後の予定

- LKST
 - サンプリング機能をつけるなどして、収集情報の増加を抑えたい
 - lkstlaを強化し、解析できるイベントを追加したい

■ 関連URL

OSS推進フォーラム <http://www.ipa.go.jp/software/open/forum/>

LKST <http://lkst.sourceforge.jp/>

iozone <http://www.iozone.org/>

OProfile <http://oprofile.sourceforge.net/news/>



商標

- ❑ Linuxは、Linus Torvaldsの米国およびその他の国における登録商標あるいは商標です。
- ❑ MIRACLE LINUXはミラクル・リナックス株式会社が使用権許諾を受けている登録商標です。
- ❑ Intel, Intel Xeon, Pentium は、アメリカ合衆国およびその他の国におけるインテル コーポレーションまたはその子会社の商標または登録商標です。
- ❑ その他の記載されている社名および製品名は各社の登録商標また商標です。

その他

- ❑ この発表内容は「独立行政法人 情報処理推進機構 オープンソースソフトウェア活用基盤整備事業」に係る委託業務の調査・開発に基づくものです。



Happy Evaluating!



補足: ビジーウェイトプロセスの解析

■ ロックプロセスの特定

| 時刻 | cpu | コマンド名 | lock/unlock | 関数 |
|---------------------------|-----|-----------|-------------|--------------|
| 10:59:59.109884679 | 0 | kdm_greet | lock | |
| 10:59:59.109884888 | 0 | kdm_greet | unlock | |
| 11:00:15.543851385 | 3 | iozone | unlock | schedule() |
| 11:00:15.543853871 | 1 | crond | lock | permission() |
| 11:00:15.543862258 | 1 | crond | unlock | |

約15秒もの間ビジーウェイト状態にあった

- iozoneプロセスが他のアプリケーションの動作を邪魔していた
- **スループット低下の原因?**