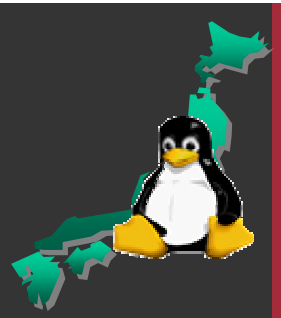


Linuxの障害解析・性能評価を支援するツール: LKST/LKSTLogTools、DAV

2005/06/02

日立製作所 システム開発研究所
杉田由美子
sugita@sdl.hitachi.co.jp

HITACHI
Inspire the Next

Contents

HITACHI
Inspire the Next

1-1 開発の背景

●ミッションクリティカル・システムでは、適切な性能保証、迅速な障害対応が必須
●Linuxには、性能評価ツール、ダンプ、トレース、ディスク状態の可視化といった、障害解析を支援する標準的なツールが無い
●十分なデータが得られず、現象調査や障害解決に時間がかかったり、原因の特定に至らないケースも多々ある

対応手段を開発し、ノウハウを蓄積することにより、障害解析、性能評価時間の短縮を図る必要あり

1-2 具体的には…

現状の問題点

- 生死監視で応答が遅れ、系切り替えが発生する等、性能遅延に起因する障害が発生。しかし原因特定できる情報が無い
- ファイルの更新を繰り返すと性能劣化が発生。データの断片化が関係していると思われるが情報が無い

情報収集機能の充実が必要

LKST
(Linux Kernel State Tracer) /
LKSTLogTools

DAV
(Disk Allocation Viewer)

Alicia

Contents

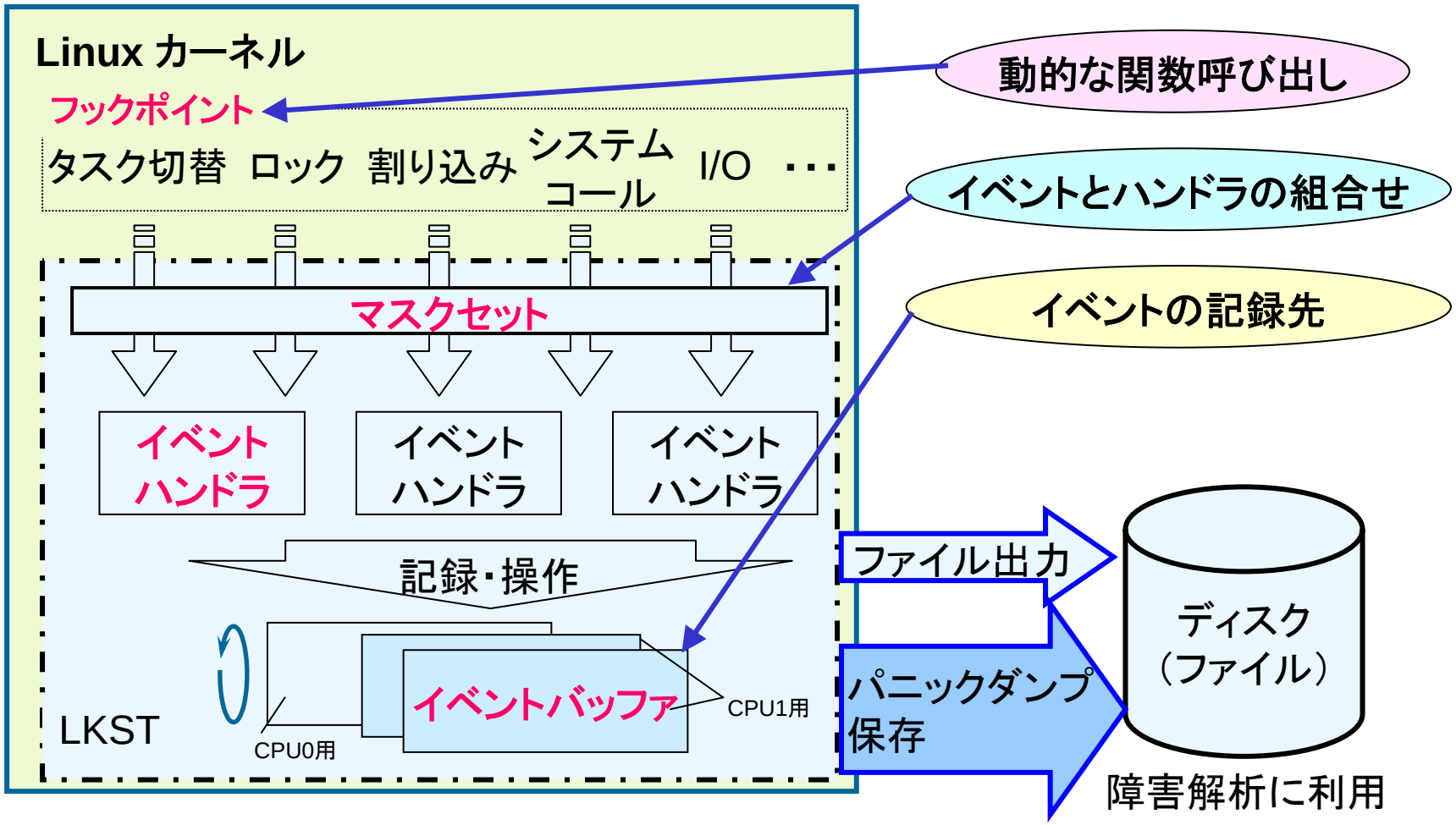
HITACHI
Inspire the Next

2-1 LKST(Linux Kernel State Tracer)とは

- Linuxカーネル向けイベントトレーサ
  (カーネルの挙動情報を収集)
  - カーネルコードに埋め込まれたフックポイント
  - カーネルドライバ(本体と組込みイベントハンドラ)
  - 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

2-2 LKSTの動作図



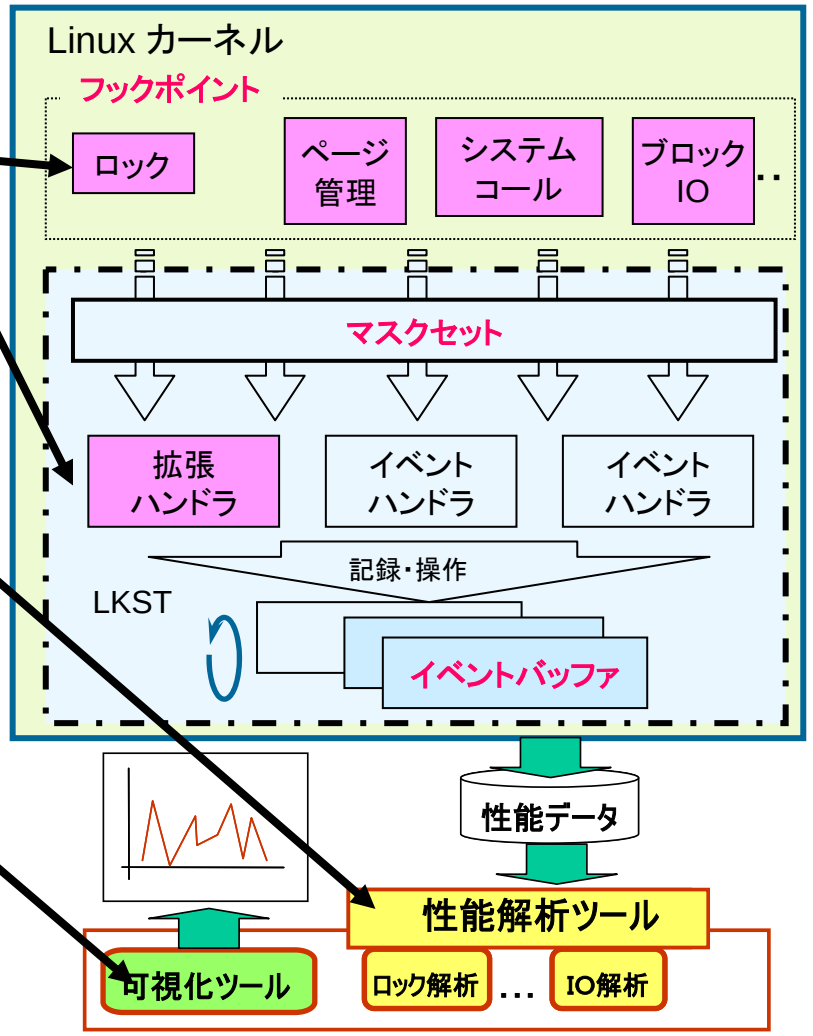
2-3 LKST Log Toolsとは

- **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**
  - 性能の特徴ごとに個別の性能指標値を算出
    - 単一の情報算出だけではない
  - 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

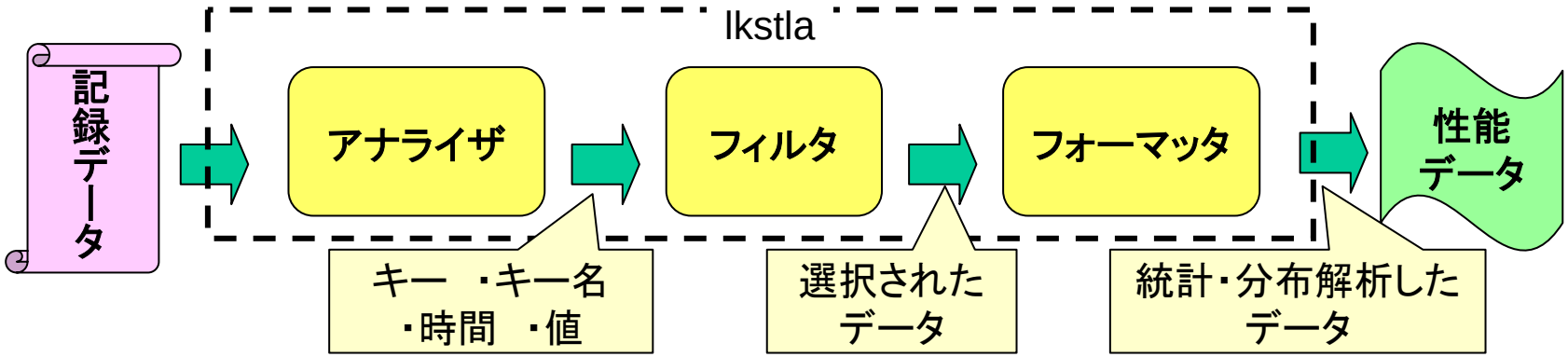
2-4 LKSTによる性能評価機能

- 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」



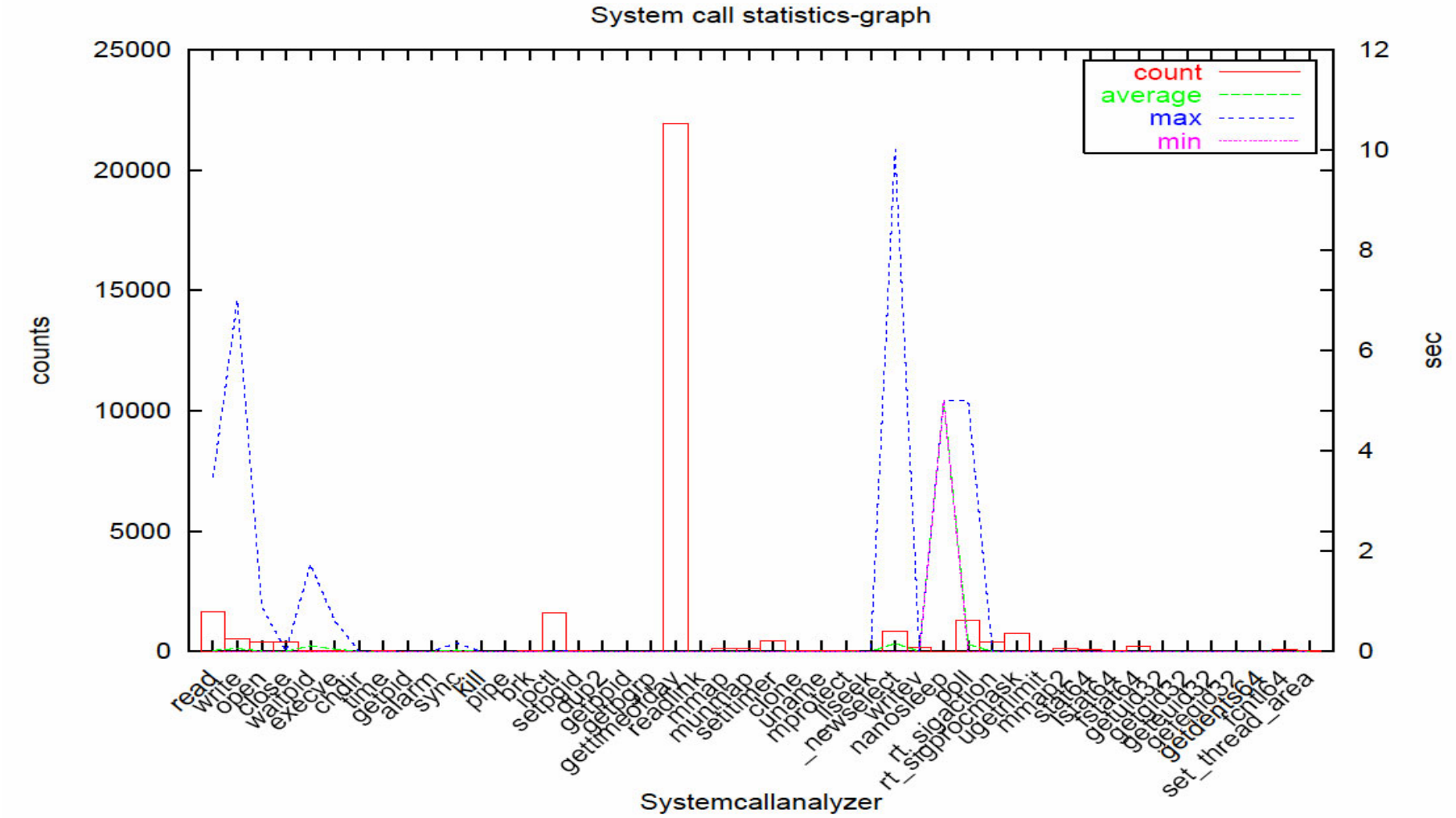
2-5 性能解析ツール(lkstla)

- 3つの機能:
  - 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。



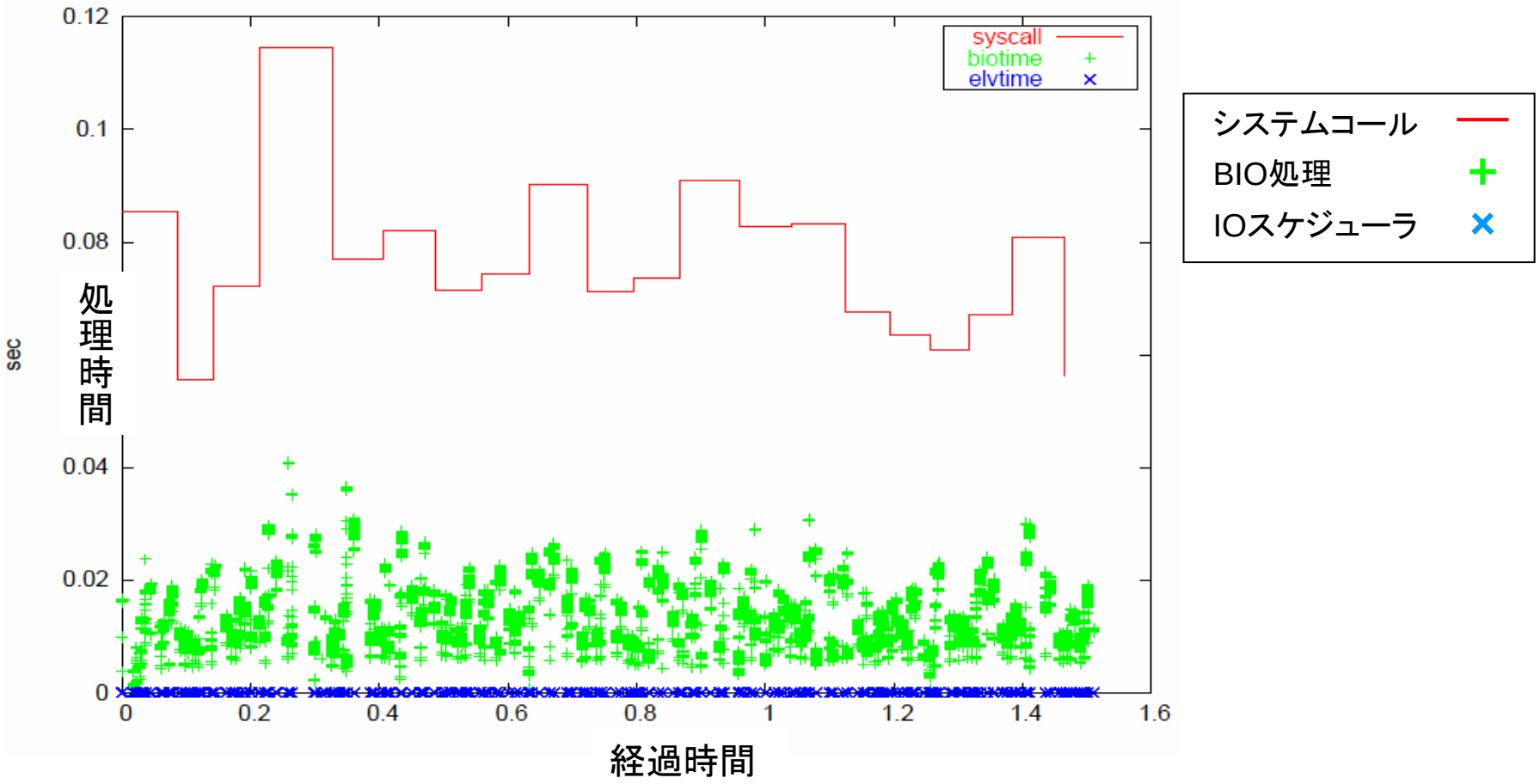
2-6 可視化ツール実施例(1)

システムコールの実行状況評価



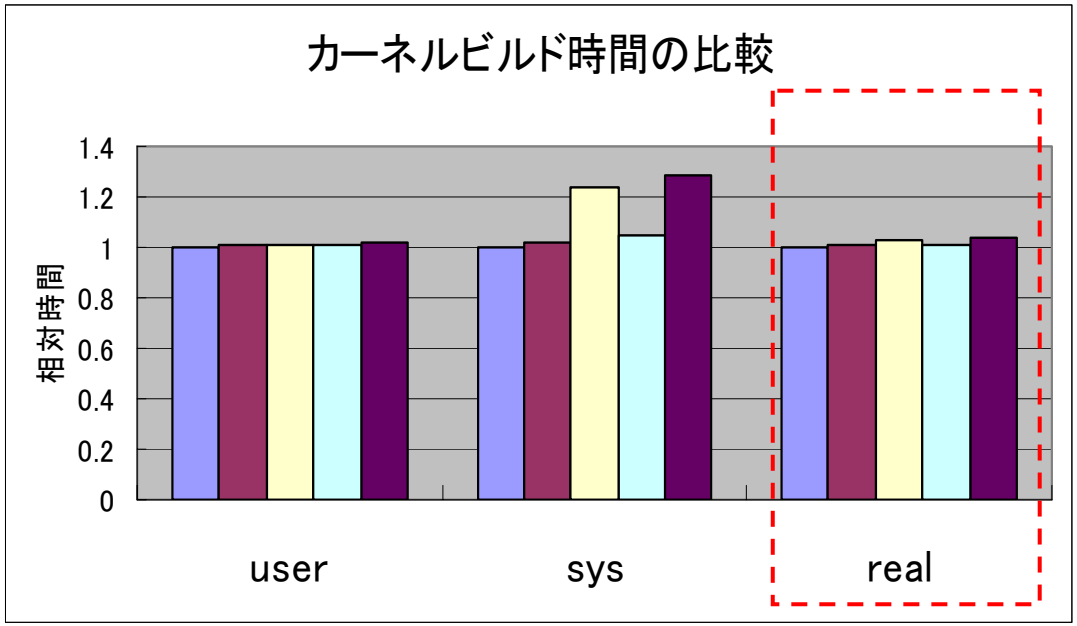
2-7 可視化ツール実施例（2）

システムコール、BIO処理、IOスケジューラの性能評価



2-8 LKST/LKSTLogToolsのオーバヘッド

- 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



(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

2-9 効果と課題

LKSTLogTools効果

- カーネルのモジュール単位の性能評価が可能
- カーネルの処理階層ごとの性能の傾向や統計情報を得ることにより、検出しにくい性能遅延箇所を検出可能
- 検出した各種データを重ね合わせて可視化することにより、多角的な性能解析が可能

潜在的な性能遅延箇所を検出可能

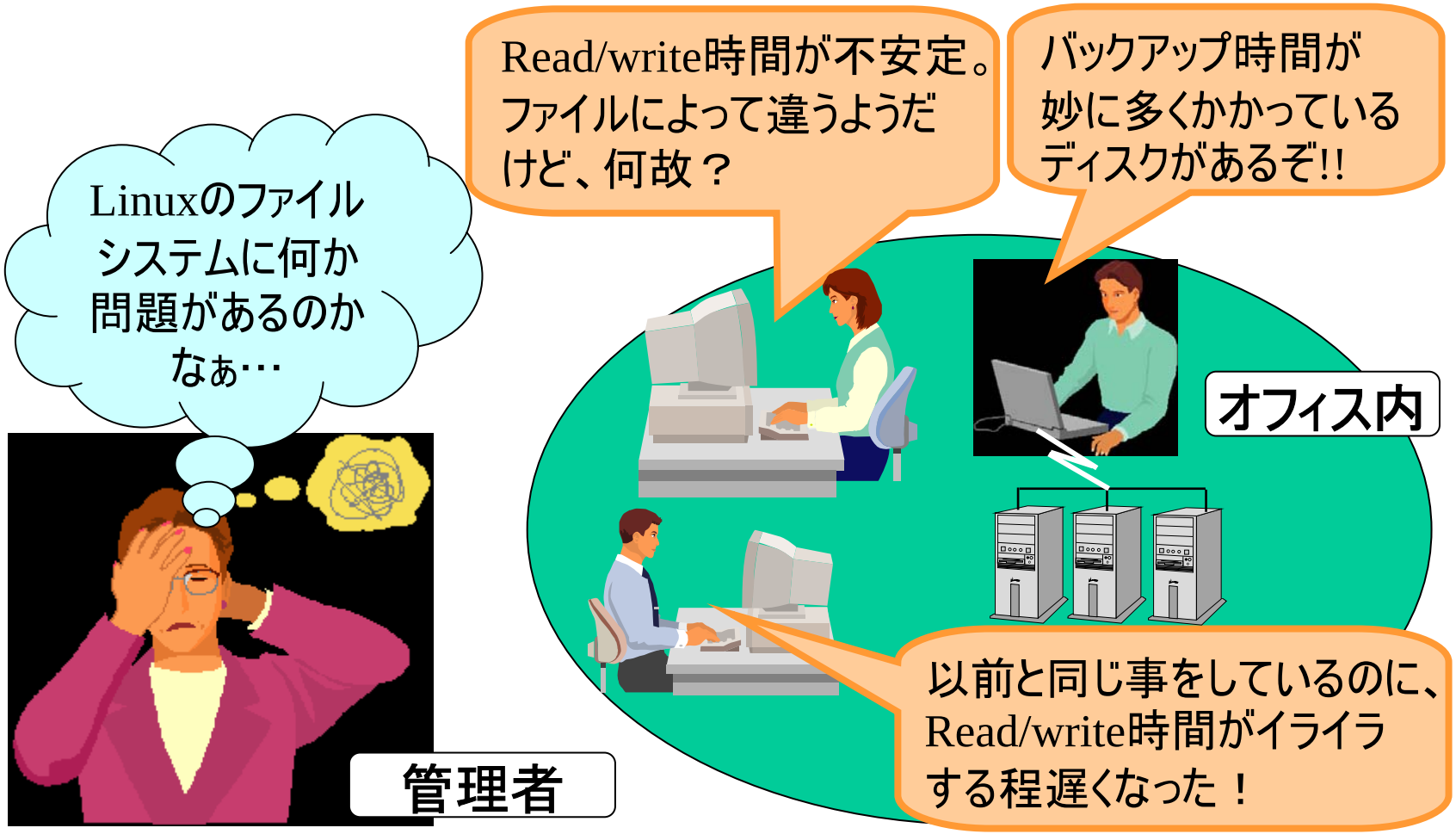
今後の課題

- 64ビット対応
- 評価実績を増やし、評価項目の充実と評価ノウハウの蓄積を図る
- GUIの充実
- オーバヘッド削減

Contents

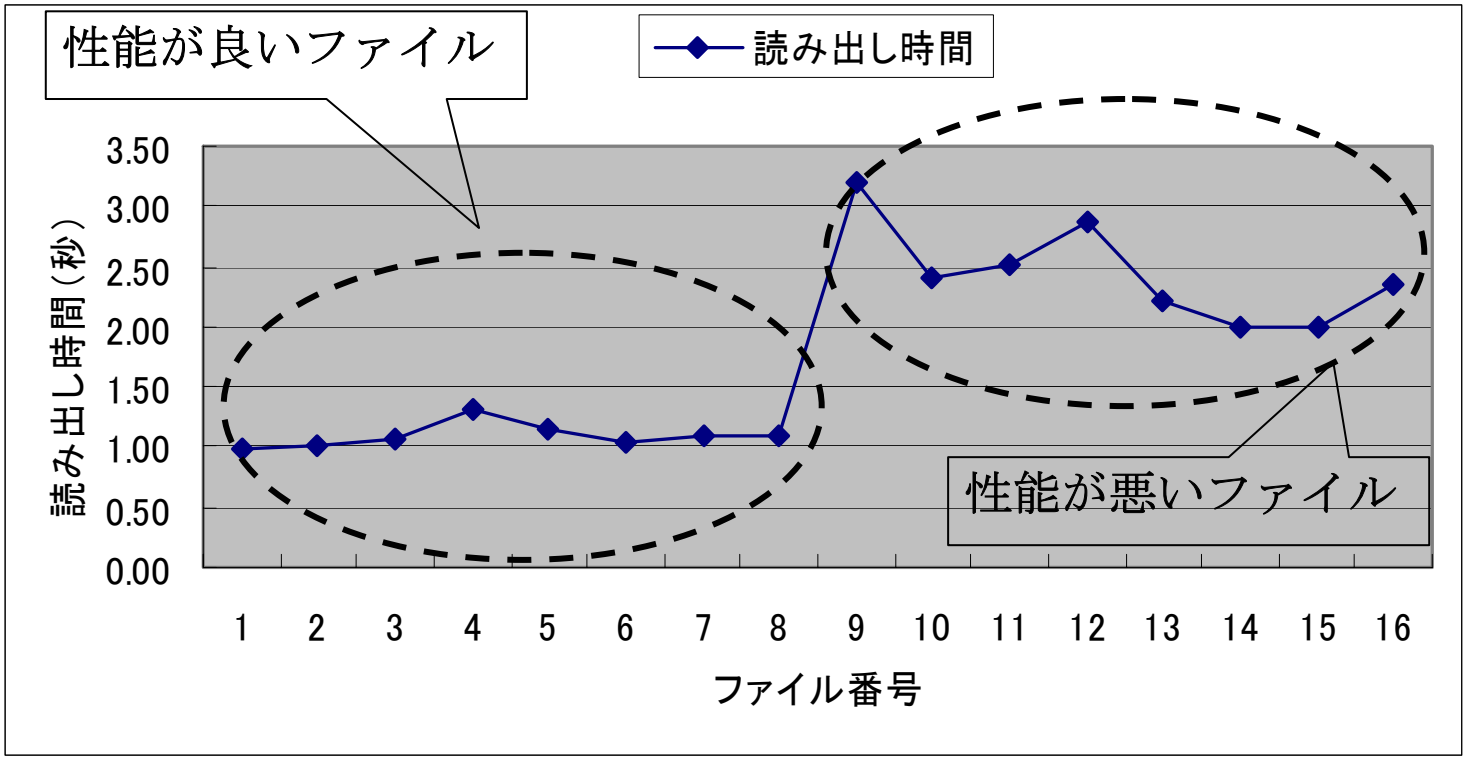
HITACHI
Inspire the Next

3-1 動機と背景


Linuxのファイルシステムに何か問題があるのかなぁ…
管理者
Read/write時間が不安定。ファイルによって違うようだけど、何故？
バックアップ時間が妙に多くかかっているディスクがあるぞ!!
オフィス内
以前と同じ事をしているのに、Read/write時間がイライラする程遅くなった！

3-2 現象事例

●同じパーティション内に、読み込み性能が早いファイルと遅いファイルが混在



3-3 DAVとは

●DAVとは

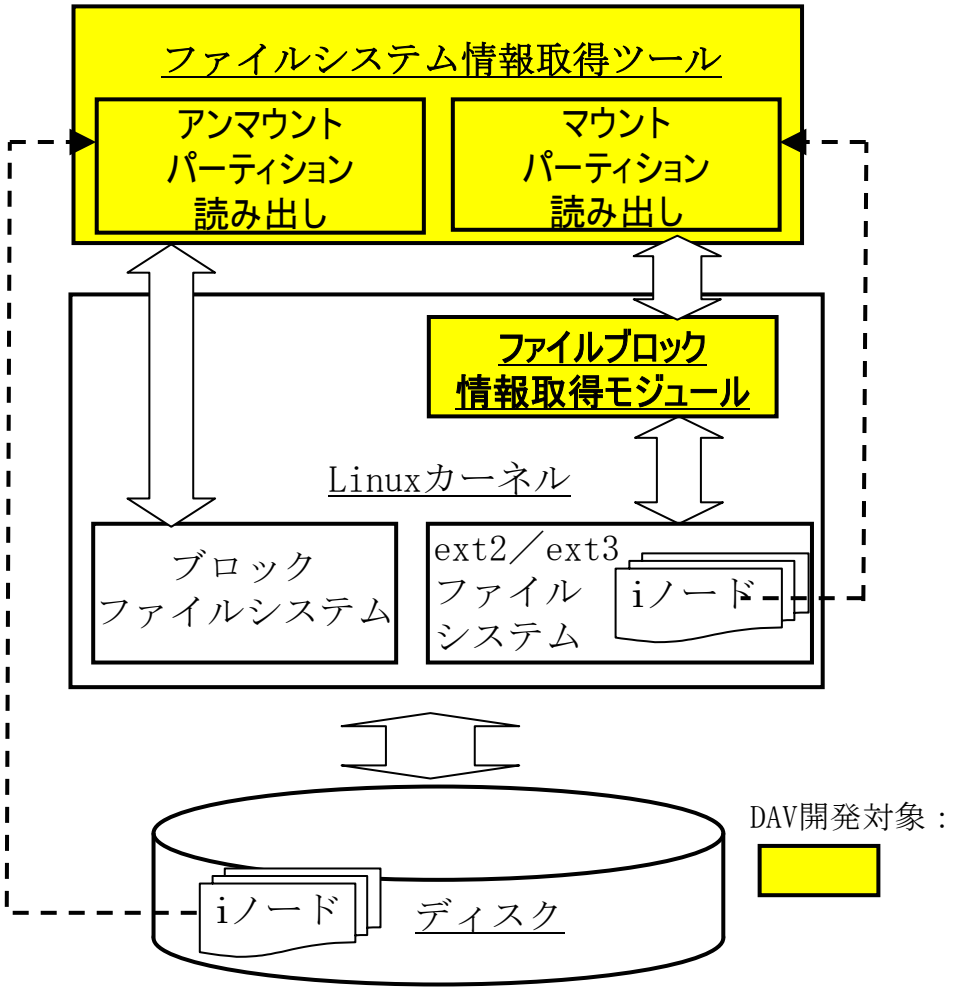
Linuxのext2／ext3ファイルシステムのフラグメンテーション情報を取得して視覚化するツールであり、下記の機能を持つ。

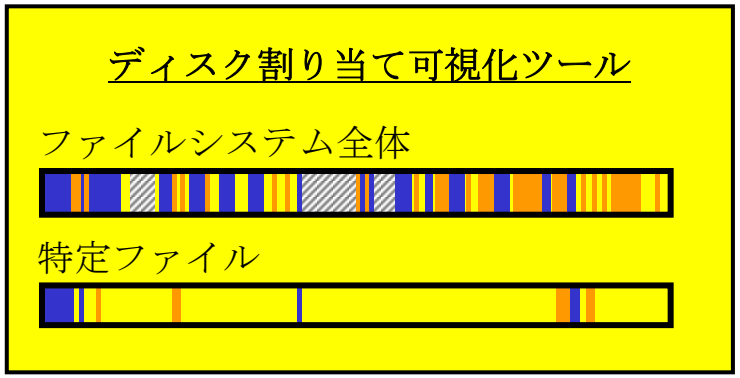
- ファイルシステムのマウント状態に無関係に、フラグメンテーション状況を取得可能
- パーティション全体／任意のディレクトリ下の全ファイル／単一ファイルのフラグメンテーション情報を取得可能

●DAVのプログラム構成

- フラグメンテーション情報を取得し、テキスト情報を出力するプログラム
- 取得した情報をGUI表示するプログラム
- dav_liveinfoモジュール（マウント状態時にパーティションのフラグメンテーション情報を取得するドライバ）

3-4 DAVの構成





●DAV構築・実行に必要な条件

- **カーネルバージョンが2.4以上であること**
- **カーネルソースが展開されていること**
- **GTK+1.2がインストールされていること**

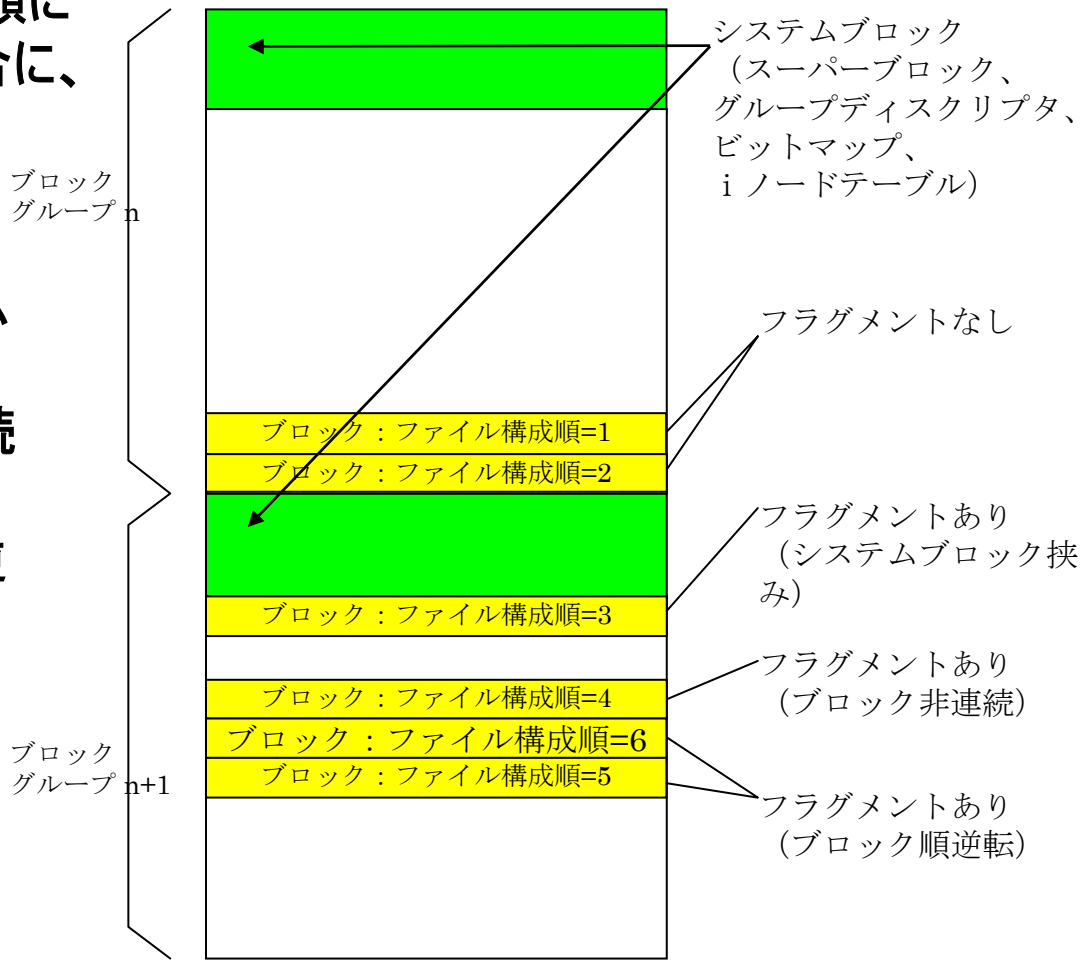
●動作の確認済み環境

- Upstreamカーネル2.6.9
  （Fedora Core 2で再構築）
- Miracle Linux V3.0
- Fedora Core2, 3

3-5 フラグメンテーションの判定仕様

ファイルブロックをファイル構成順に見て行き、下記の状態の場合に、フラグメントブロックと判断。

(1). ファイルブロックが、システムブロックを挟んでいる。
(2). ファイルブロック番号が連続していない。
(3). ファイルブロックの順番が逆転している

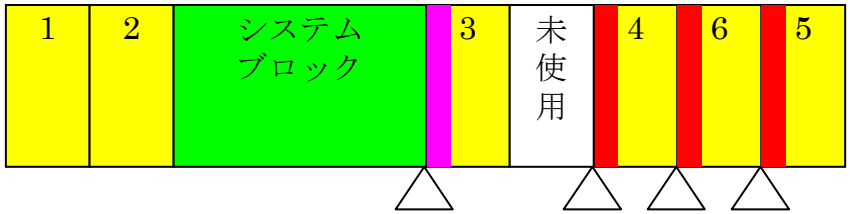


3-6 フラグメンテーションの表示仕様

フラグメンテーション状況の出力（テキスト形式／GUI表示）は2種類

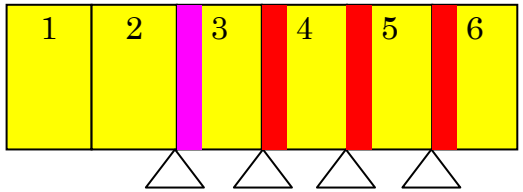
(1).ブロック番号順表示

物理的なブロック番号順でシステムブロックを含めて表示する表示



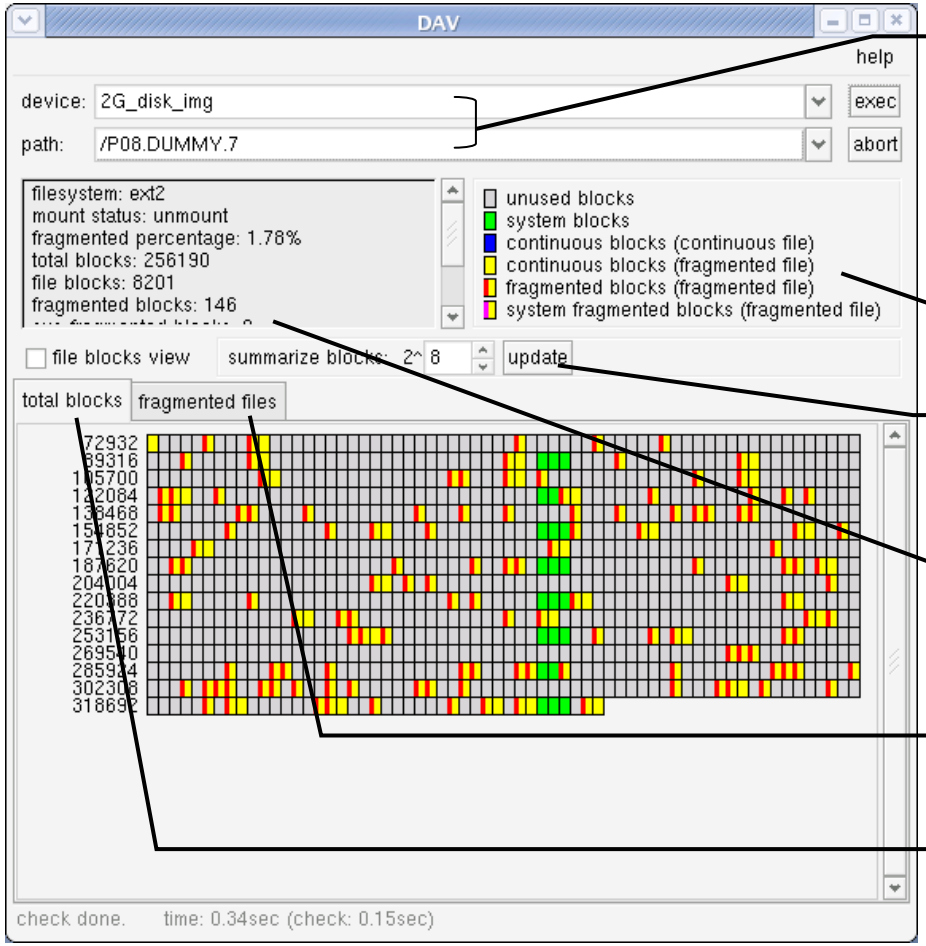
(2).ファイル構成順表示

確認対象のファイルを構成するブロックだけを、その構成順に表示



※数字は、ファイル構成順番
△は、フラグメント位置

3-7 DAVの出力形式（GUI）



フラグメンテーション状況確認対象の指定

フラグメンテーション状況取得の開始／中止を指定

ブロック表示の色分けの凡例

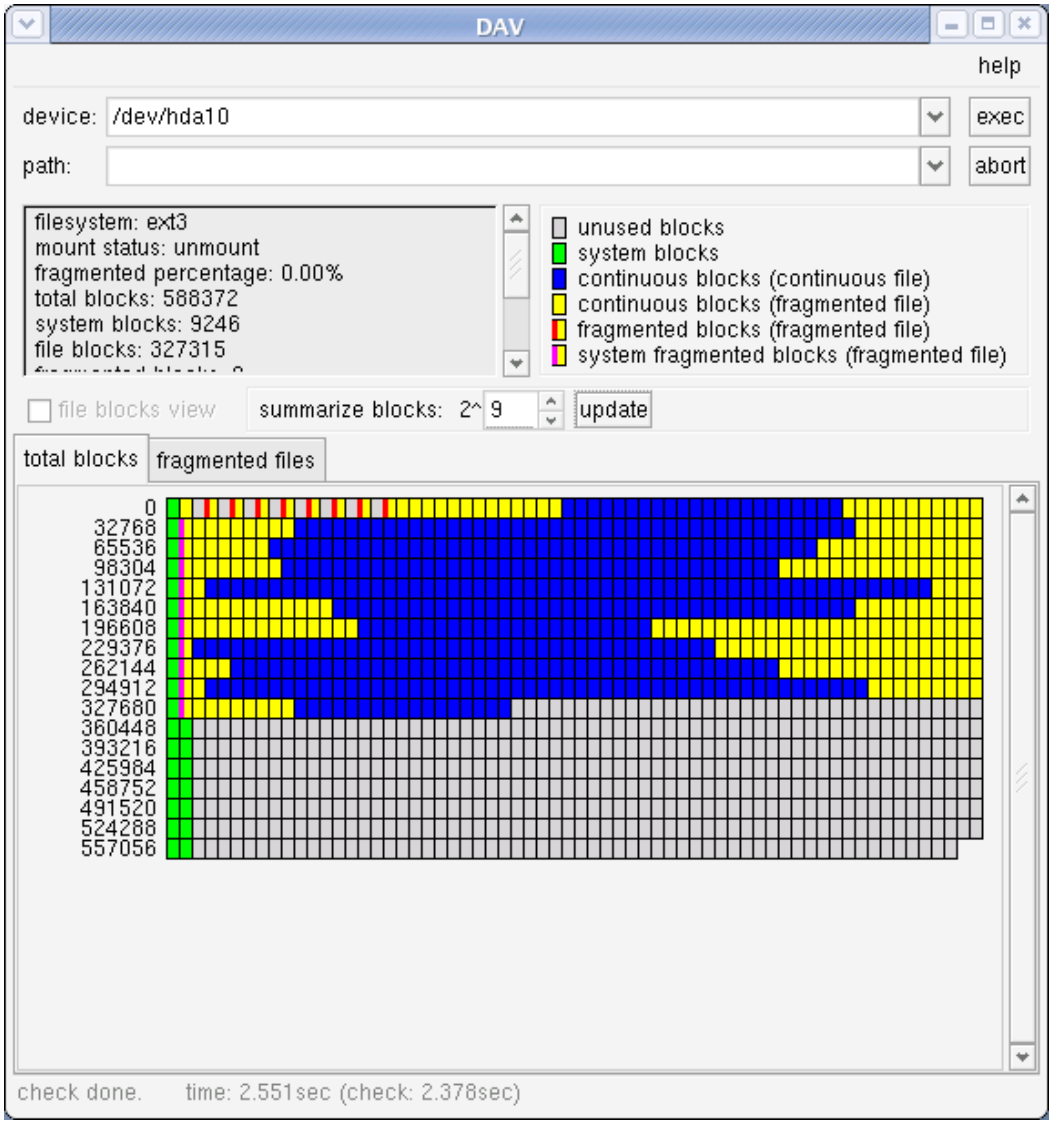
何ブロック分を集約して１つのブロックに表示するかを指定

フラグメンテーション状況のテキスト出力部分

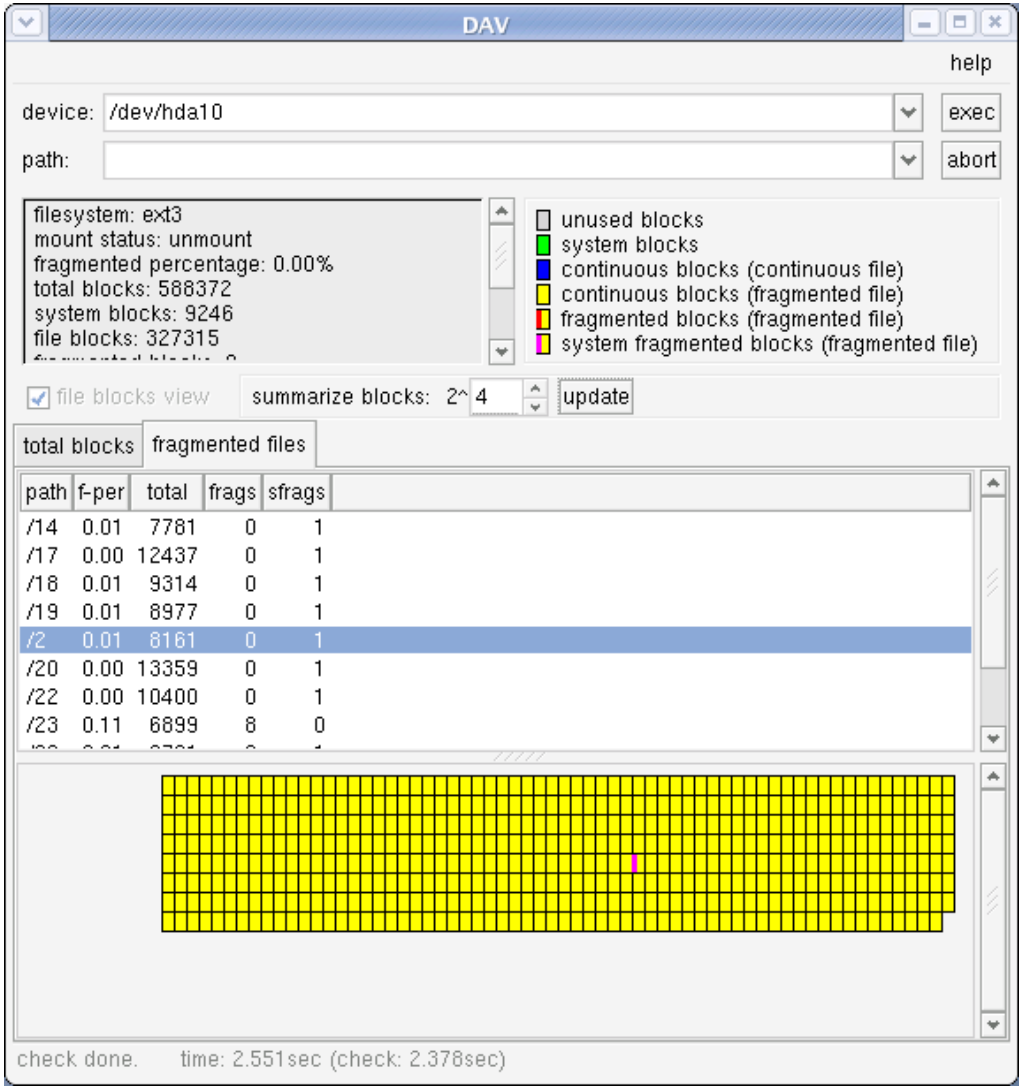
フラグメントファイル表示タブ

フラグメンテーション状況のブロック全体表示タブ

例：集約ブロック数を変更した後の画面



例：フラグメントファイル一覧から、ファイルを選択して表示



例：デフラグツールの効果を視覚的に確認

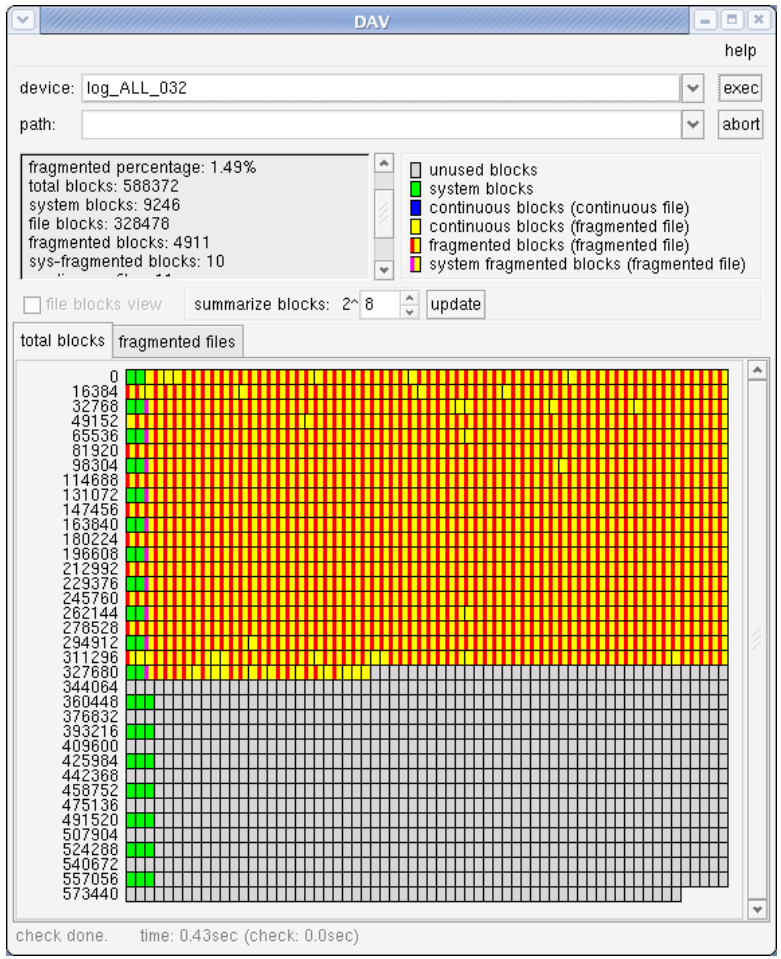
1．フラグメンテーション状態のスナップショットを取得
```
dac -Tv /dev/hdaXX > log_ALL_032
```

2．デフラグ実行

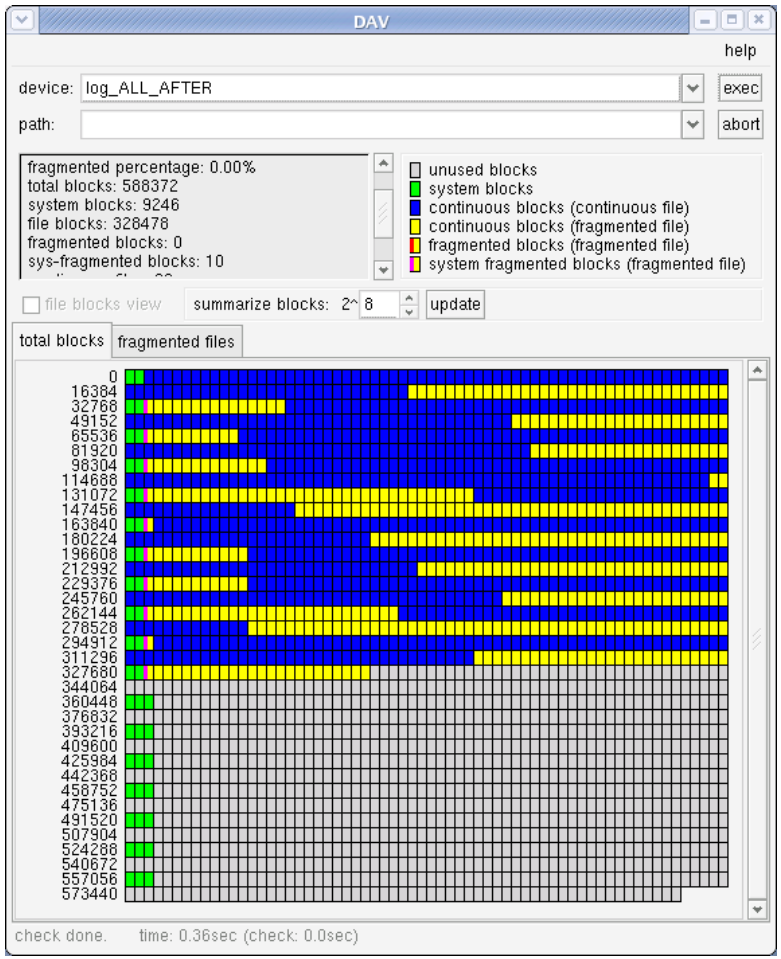
3．デフラグ実行後のスナップショットを取得
```
dac -Tv /dev/hdaXX > log_ALL_AFTER
```

4．それぞれのスナップショットファイルをGUI表示すると…

デフラグ前



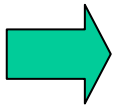
デフラグ後



3-8 効果と課題

DAV効果

- アクセス性能劣化時にディスクのフラグメンテーションがどう生じているかを視覚的に確認可能
- パーティション単位とファイル単位のフラグメント状態を取得可能
- フラグメンテーション情報を保存しておき、後でGUI表示可能
- カーネルプログラムを変更することなく実行可能



性能劣化の、より詳細な分析を行うためのヒントを、容易に取得

今後の課題

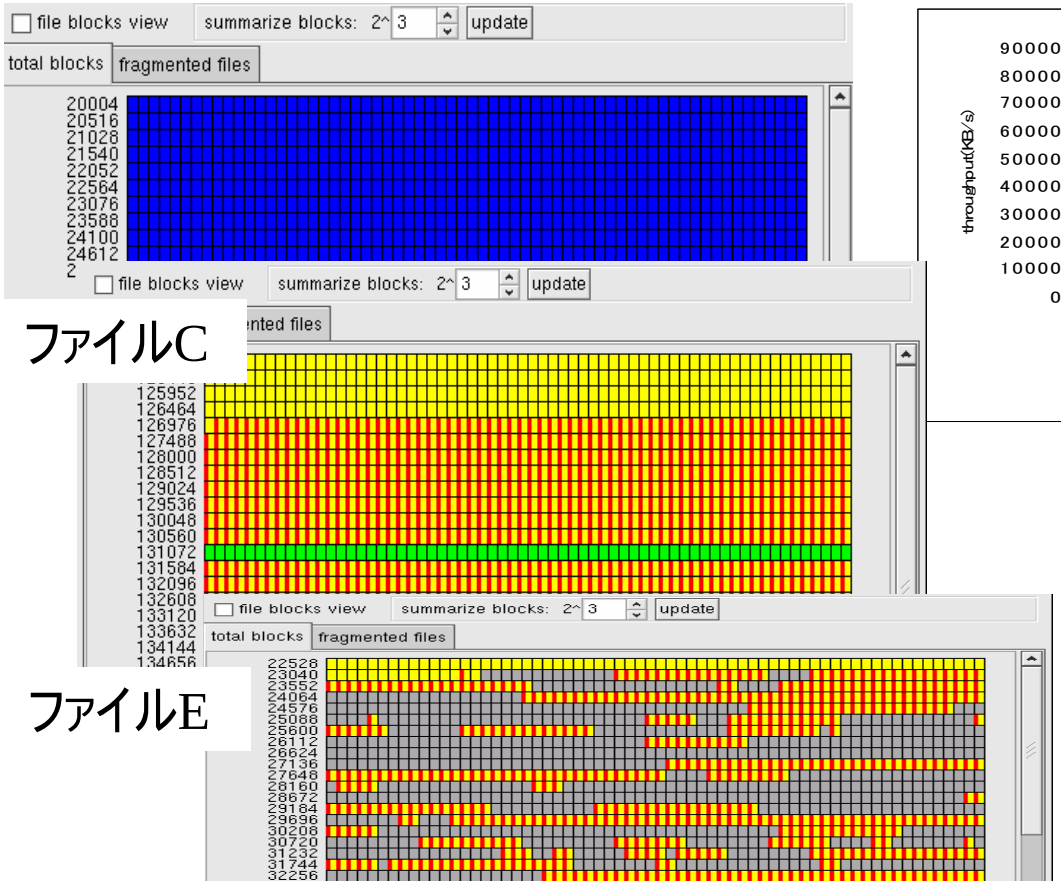
- Gtk+2への対応
- 2つのDAV結果の差分の取得
- 他ファイルシステム（XFS, JFS等）対応
- GUIの充実

Contents

HITACHI
Inspire the Next

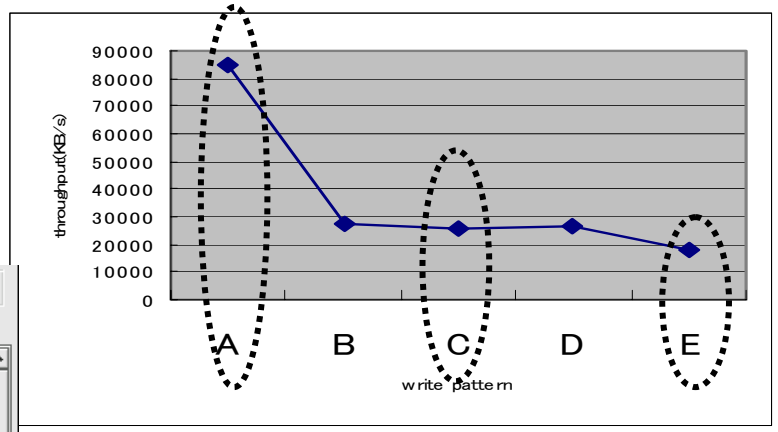
フラグメンテーション発生状況(DAV)と性能

ファイルA



ファイルC

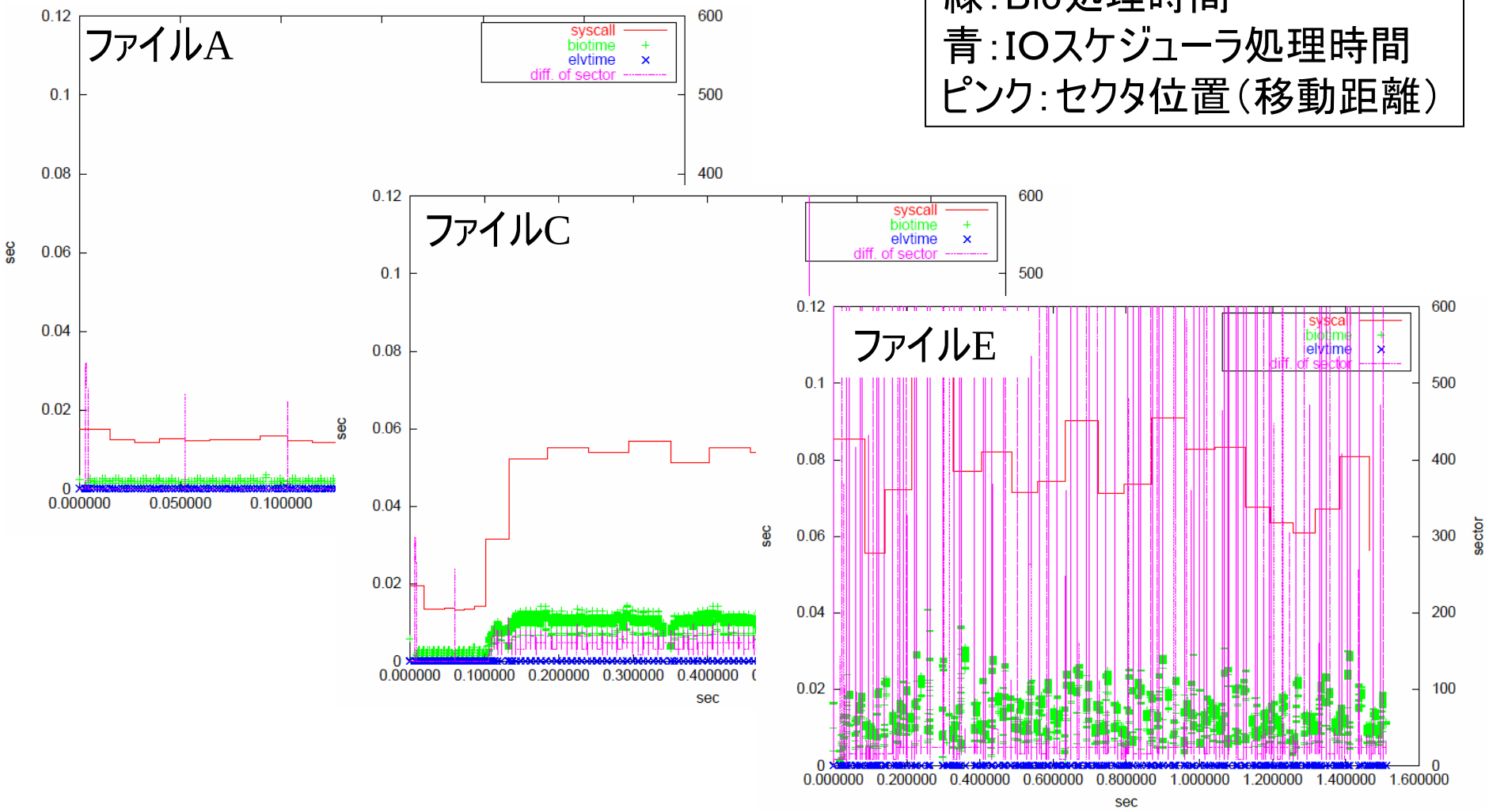
ファイルE



データ読み込み性能
（スループット）

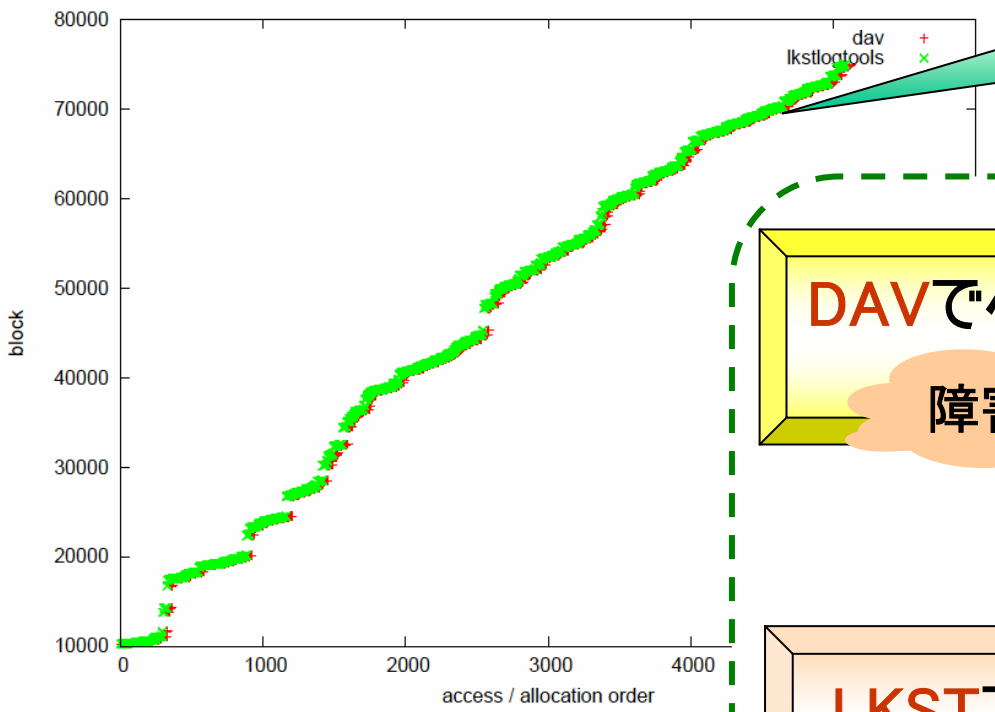
LKST/LKSTLogToolsを使った分析

赤：Syscall処理時間
緑：Bio処理時間
青：IOスケジューラ処理時間
ピンク：セクタ位置（移動距離）



Linux Conference 2005(6/1)発表
「フラグメンテーション解析を支援するDAVとLKSTLogToolsの開発」より

LKST/LKSTLogToolsとDAVの情報比較



DAVとLKSTの情報は
ほぼ一致

DAVでヘルスチェック

障害の兆候

LKSTで詳細解析
LKSTLogTools

運用改善

確認

コード改善

Contents

HITACHI
Inspire the Next

5-1 URLとメーリングリスト

URL

◆日本OSS推進フォーラム（IPAページより）：報告書公開
http://www.ipa.go.jp/software/open/forum/

◆英語：ソースコードとマニュアル公開
http://sourceforge.net/projects/lkst　　http://sourceforge.net/projects/davtools

◆日本語：同上
http://sourceforge.jp/projects/lkst　　http://sourceforge.jp/projects/dav

メーリングリスト

◆英語
lkst-users@lists.sourceforge.net　　davtools-users@lists.sourceforge.net
lkst-develop@lists.sourceforge.net　　davtools-develop@lists.sourceforge.net

◆日本語
lkst-users@lists.sourceforge.jp　　dav-users@lists.sourceforge.jp
lkst-develop@lists.sourceforge.jp　　dav-develop@lists.sourceforge.jp

5-2 商標、その他

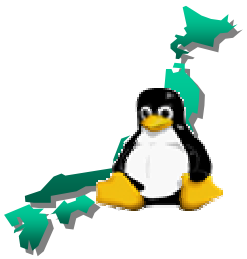
- Linuxは、Linus Torvaldsの米国およびその他の国における登録商標または商標です。
- MIRACLE LINUXは、ミラクル・リナックス株式会社が使用許諾を受けている登録商標です。
- その他、記載されている会社名、製品名は、各社 の登録商標または商標です。

このプログラムの一部は、
「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」
に係る委託業務の一環として開発しました。

5-3 お願い

LKSTLogToolsやDAVを使って、
あなたのサーバやPCのカーネルの動きやディスクの
フラグメンテーション状態を確認してみて下さい。

LKSTLogTools/DAV機能へのご意見・ご要望は、
メーリングリストへ。　開発にもご参加ください！



（株）日立製作所
http://www.hitachi.co.jp

日本OSS推進フォーラム 開発基盤WG
http://www.ipa.go.jp/software/open/forum