Copy On Writeを用いた オーバーレイブロックデバイスの実装

丹 英之[†] , 須崎 有康[‡] , 飯島 賢吾[‡] , 八木 豊志樹[‡]

tanh@alpha.co.jp, {k.suzaki, k-iijima, yagi-toshiki}@aist.go.jp

株式会社 アルファシステムズ[†]
独立行政法人 産業技術総合研究所[‡]

発表の概要

- 背景: ライブCDの現状
- 目的: 次世代ライブCD
- 目的を達成するための手段の検討
- 関連研究
  - Loopデバイス, Device-mapper, Unionfs, User-mode Linux
- 本研究では
  - 方針, Overlay Block Deviceのソフトウェア階層, 処理シーケンス, 差分情報を保持するファイル, 関連研究との違い
- OBDの実装とライブCDへの組み込み
  - KNOPPIX, 圧縮ループデバイス, OBD-KNOPPIXの起動シーケンス, ディレクトリ構成
- OBDの評価
  - スループットの確認, Unionfsとの比較, 組み込んだライブCDの評価
- まとめ
- 今後の検討/展開

背景: ライブCD(CD起動のOS)

以前からCD起動のOSが用いられてきた

- 常用OSをHDにインストールするため
- HDメンテナンス用

概して非・常用OS

Linux ConferenceなのでCD起動Linux

背景: ライブCDが広まりつつある

- プレスで安価に大量複製
  - 雑誌の付録/自由なソフトウェアの配布手段
- メディアが持つロバスト性
  - ユーザが想定外の操作をしても再起動で復旧

- 教育分野などへの適用
  - KNOPPIX-Edu, etc.
- アプリケーションに特化したPCへの適用
  - PCベースアプライアンス/ゲーム専用機？

背景: ライブCD普及への阻害要因

- 起動が遅い
  - この課題の解決方法は，機会があったらお話します．
- 書き込みできない
  - メディア自体が書き込み不可能
  - メディア上にのるファイルシステムは変更できない
  - ユーザはそのディレクトリツリーに変更を加えることができない

システムを弄り倒したいユーザ/改変していく用途には，
やっぱりOSがHDにインストールされている必要がある？

ロバストである長所は，適用先によっては短所となる

利点を生かしつつ適用範囲を広げたい

目的: 次世代のライブCD

利点を生かしつつ適用範囲を広げるには

ディレクトリツリーをユーザが自由に変更できるようになれば, (ある層に)受けが良くなるかもしれない.

他にも沢山あるはず. 面白いネタがあったら一緒にやりましょう.

目的を達成するための手段の検討

ユーザが自由にファイル変更できるようにしたい

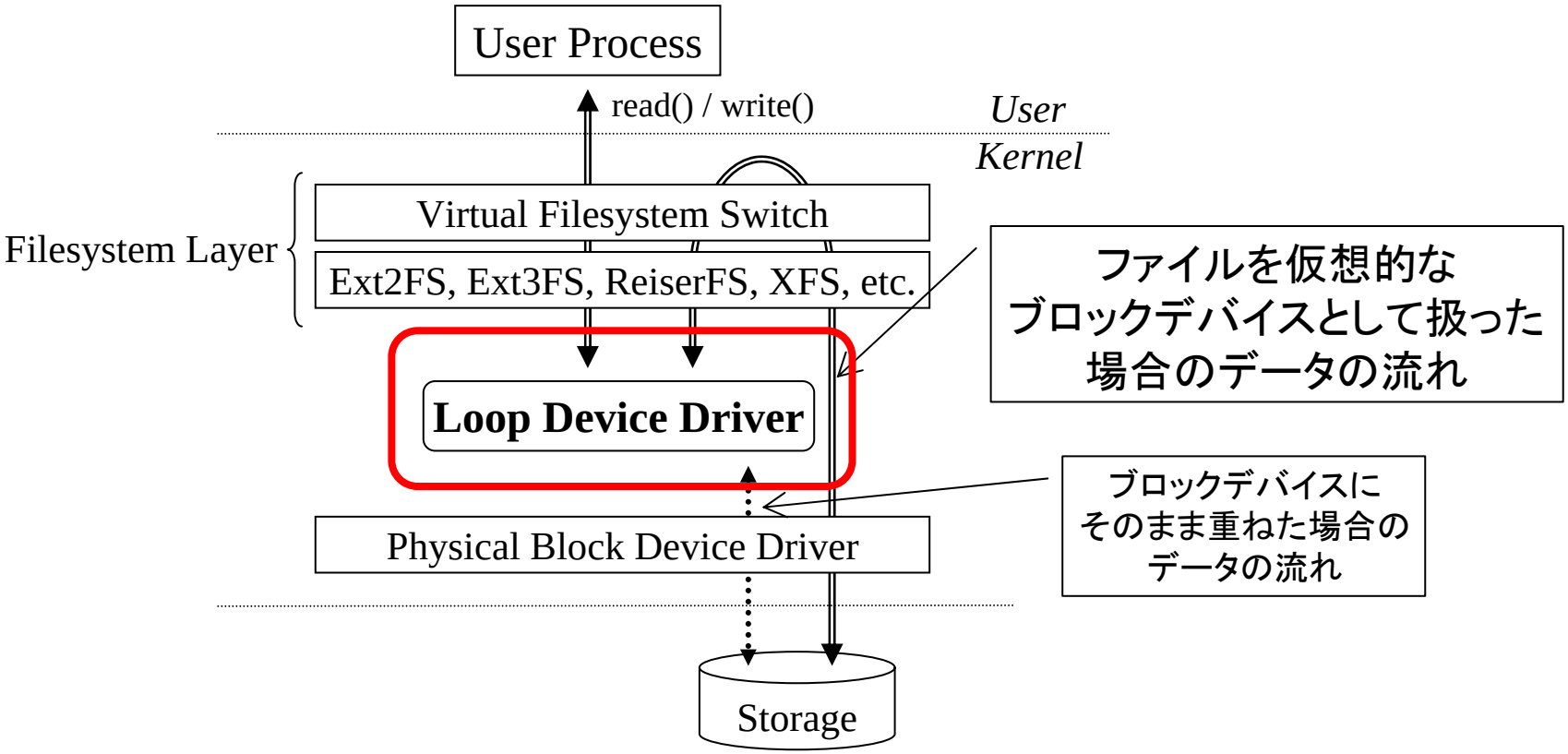
- CD-ROMへの書き込みは不可能.
- ユーザには書き込みできているように見せ掛けるしかない.
- ユーザの見えない所で擬似的な書き込みを実現する.
- どのソフトウェア階層で実現するか？

関連研究

- ブロックデバイスの仮想化
  - Loopデバイス
  - Device-mapper
- 擬似的書き込みを実現する仮想FS
  - Unionfs
- 仮想OSのデバイス
  - User-mode Linux(UML)のUML Block Device

関連研究(1)
Loopデバイス

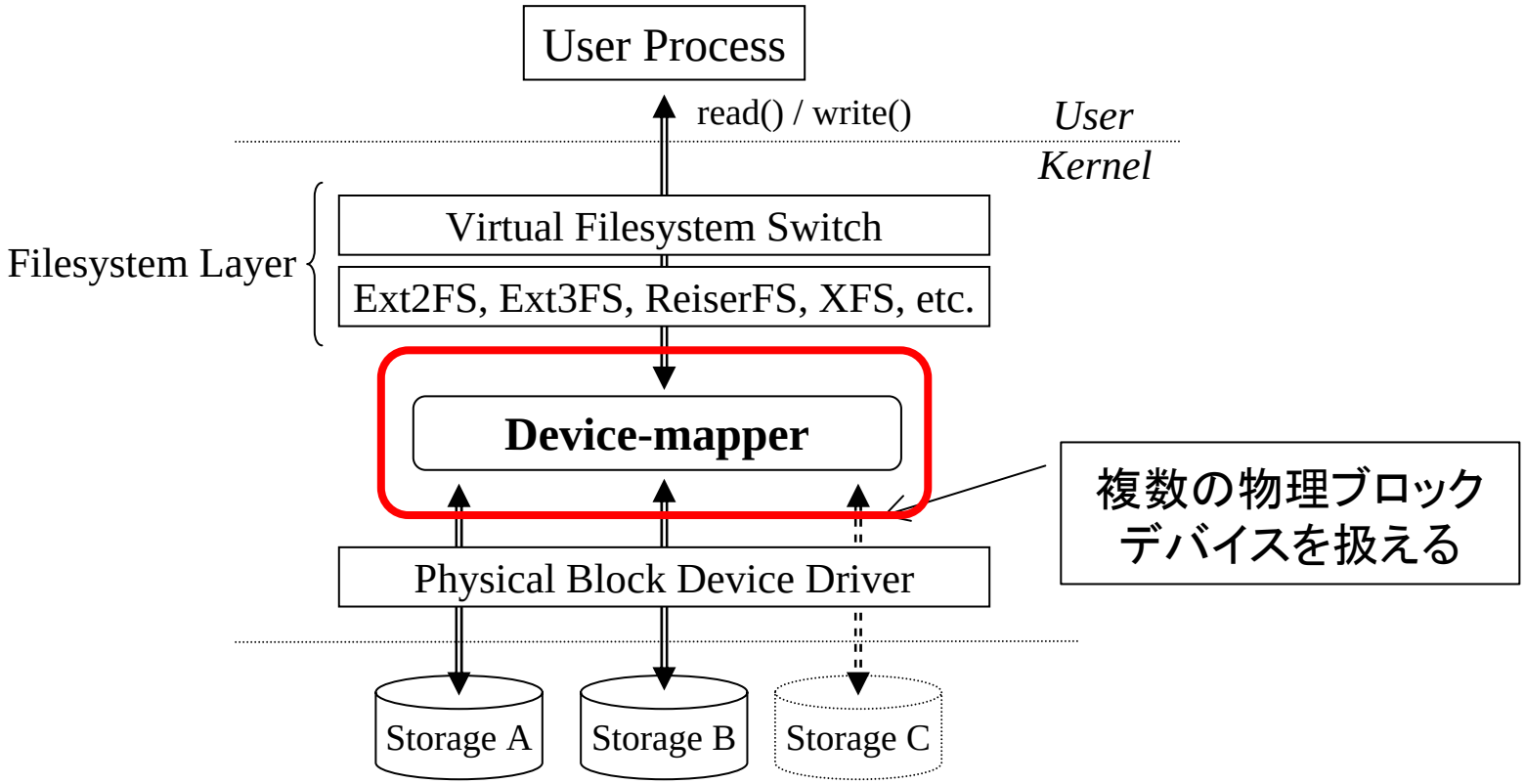
ファイルをブロックデバイスのイメージとして扱う
仮想ブロックデバイス



関連研究(2)

Device-mapper

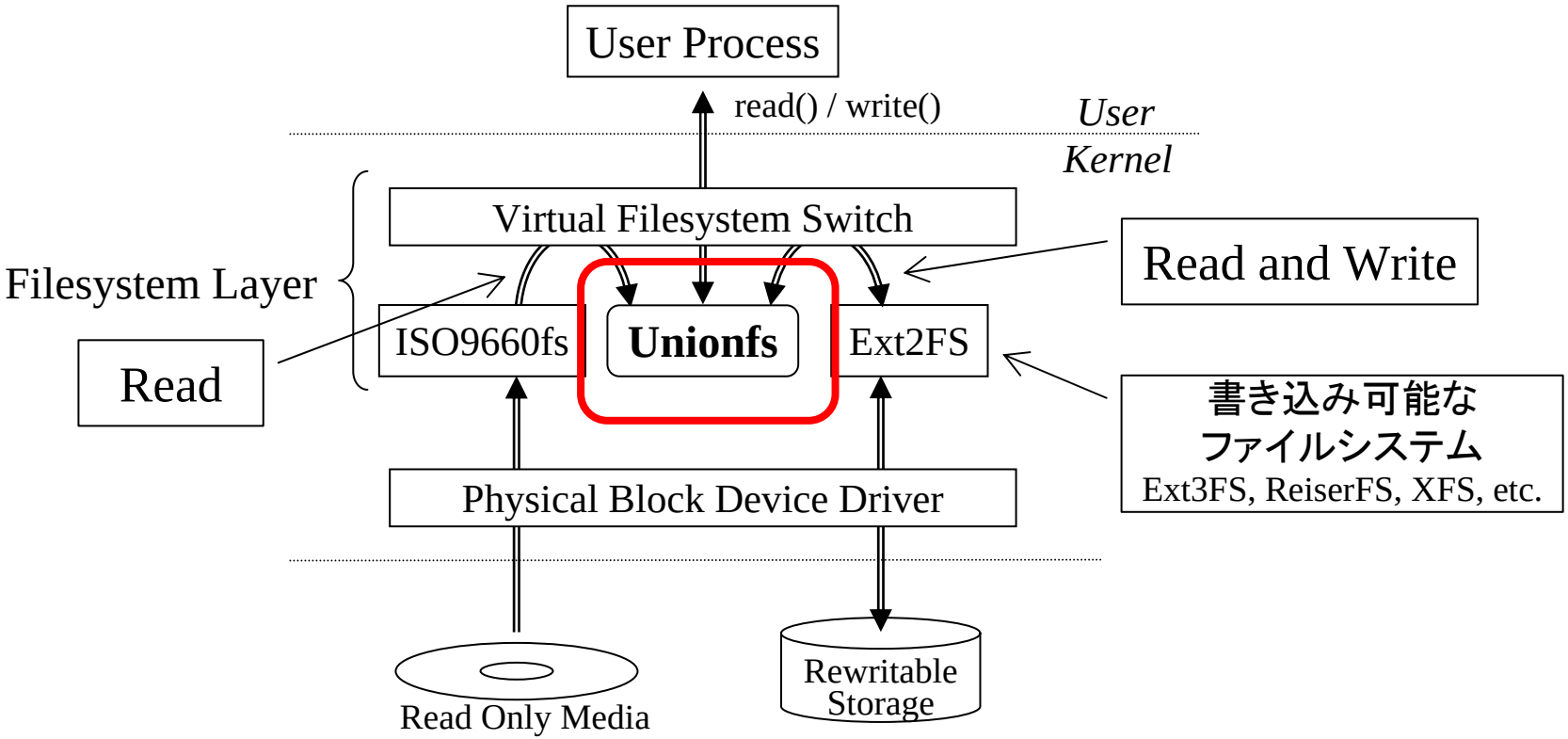
複数の物理ブロックデバイスを統合し
論理ブロックデバイスとして提供するソフトウェアレイヤ



関連研究(3)

Unionfs

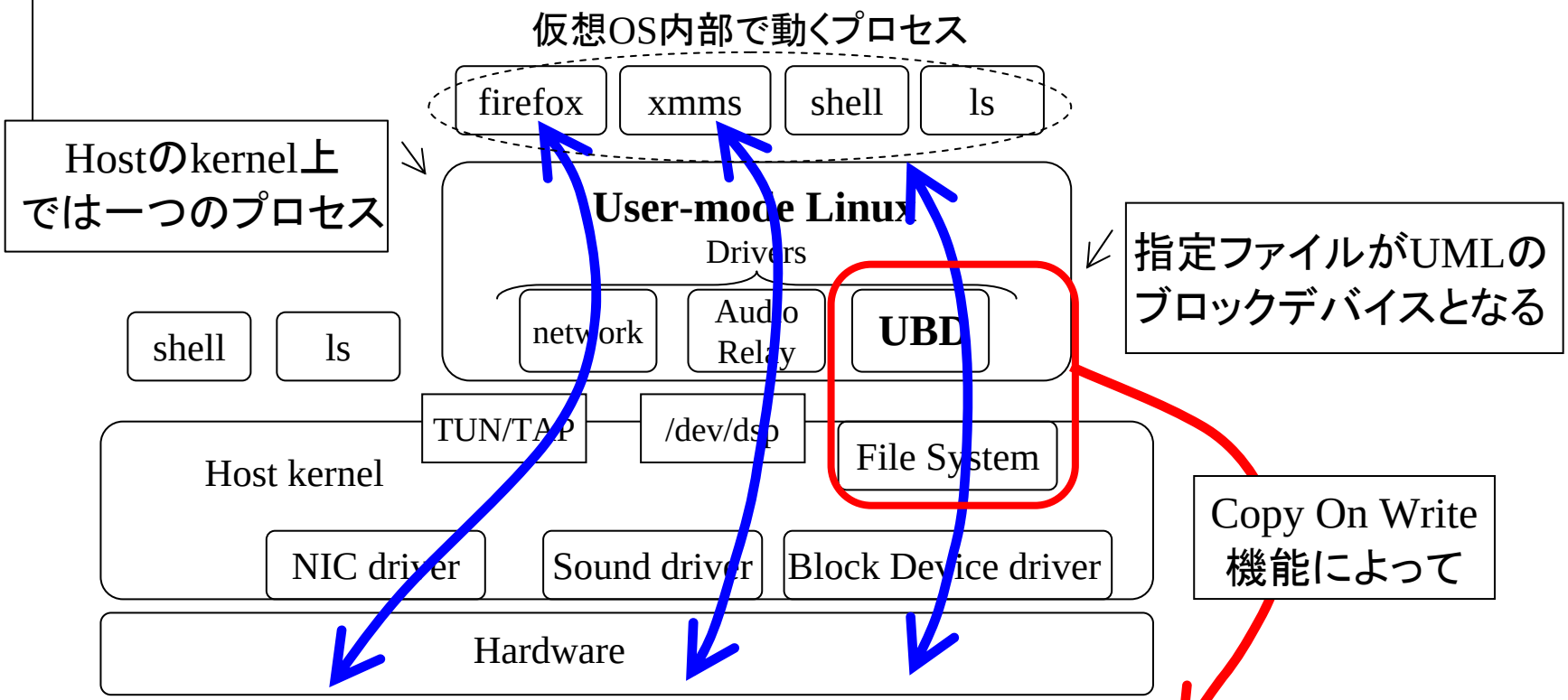
二つのディレクトリ(マウントポイント)を
一つに束ねる仮想ファイルシステム



関連研究(4)

User-mode Linux (UML)のUML Block Device (UBD)

仮想OS内部から参照する仮想ブロックデバイス



ブロックデバイスの更新分を参照元の差分としてCOWファイルに保持できる

関連研究のまとめ

- Loopデバイス
  - ファイルをブロックデバイスのイメージとして扱う仮想ブロックデバイス.
  - 仮想ブロックデバイスとそのデータ元(ファイル/物理ブロックデバイス)は1対1.
- Device-mapper
  - 複数の物理ブロックデバイスを統合し論理ブロックデバイスとして提供するソフトウェアレイヤ.
  - Copy On Writeで差分を他のブロックデバイスに保持できる.
  - 扱う対象はブロックデバイスのみ.
- Unionfs
  - 二つのディレクトリを一つに束ねる仮想ファイルシステム
- User-mode LinuxのUML Block Device
  - 仮想OS内部で参照するブロックデバイス.
  - Copy On Writeで指定ファイルからの差分をCOWファイルに保持できる.

本研究では

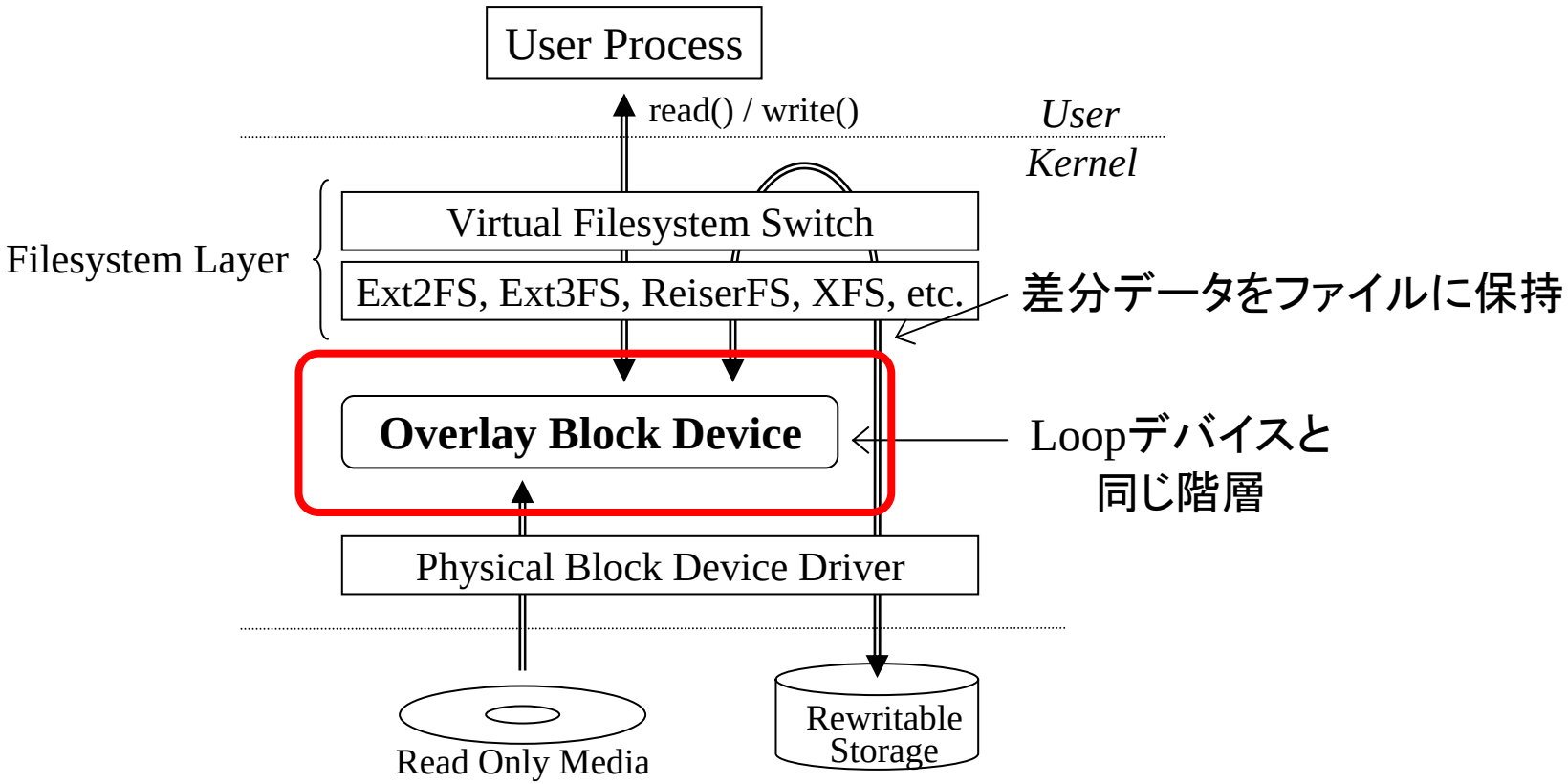
- どのレイヤで擬似的書き込みを実現するか？
  - 楽な実装/他のライブCD応用系(UML-KNOPPIX, HTTP-FUSE-KNOPPIX, etc.)との連携

ブロックデバイスレベルでの実現
対象とするブロックデバイスに重ねる仮想デバイス

- 差分はどうするか？
  - 差分データは扱い易く

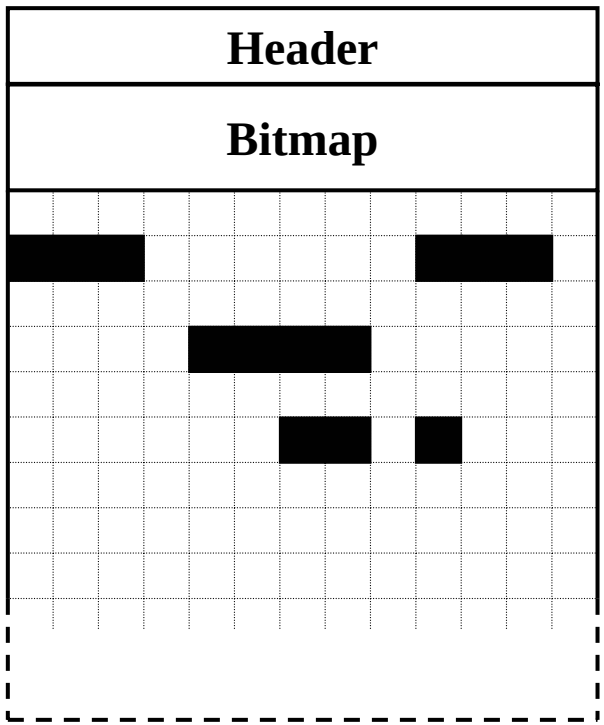
仮想デバイス内でCopy On Write処理
差分データはファイルに出力

Overlay Block Device (OBD)のソフトウェア階層



差分情報を保持するファイル

User-mode LinuxのUBDで用いる
COWファイル(v3)と同じフォーマット



magic, バージョン,
対象デバイスのサイズ,
最終更新時刻, セクタサイズ

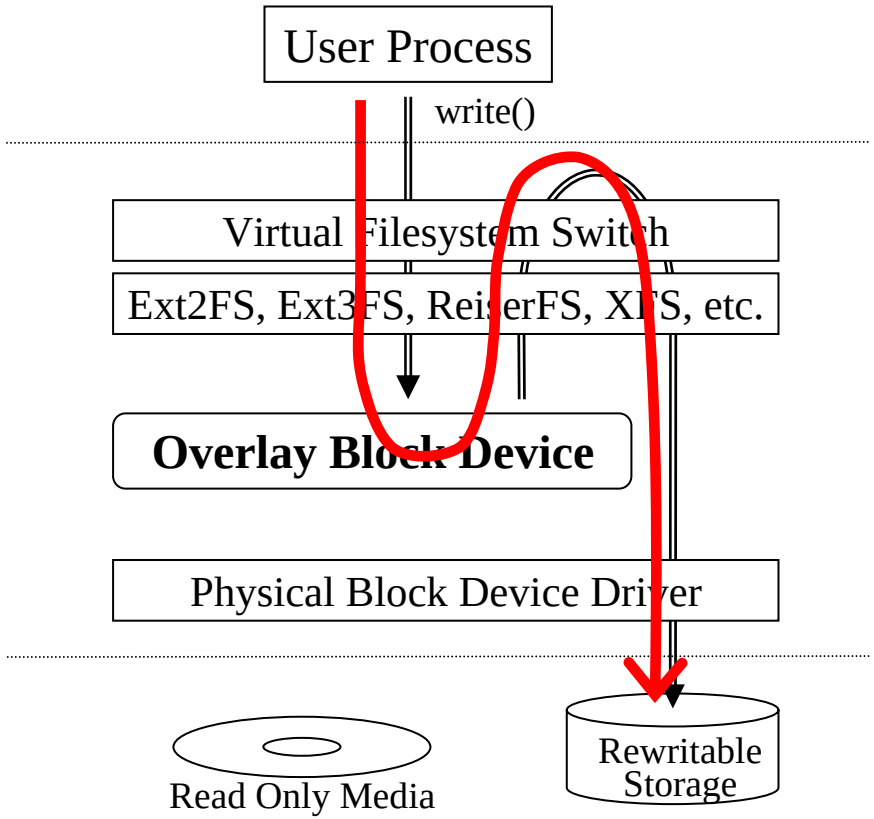
更新状況を保持
(1 bit/sector)

更新されたセクタの内容

Bitmapと更新内容の部分は
スパース

未使用(零)部分は，物理的な
ディスク領域を消費しない.

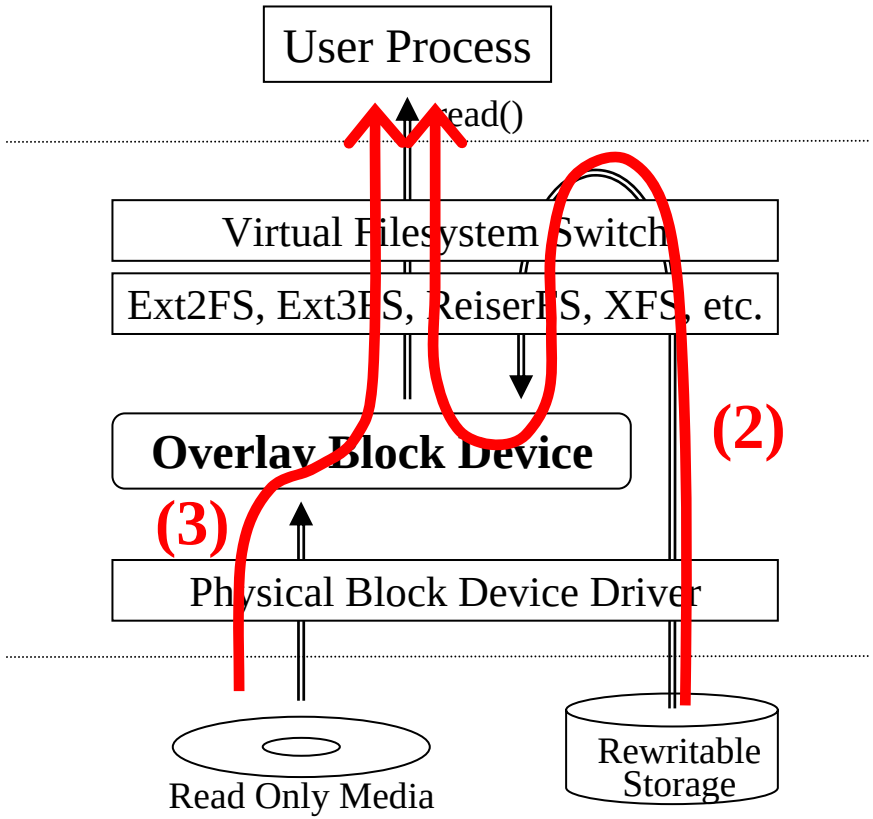
ストレージへ書き出す場合



(1) 更新されるセクタのフラグを立てる.

(2) 書き込みデータを差分ファイルに出力.

ストレージから読み込む場合



(1) 読み込むセクタのフラグを確認.

(2) フラグありでは，差分ファイルから読み込む.

(3) フラグなしでは，元のブロックデバイスを参照する.

関連研究と本研究(OBD)の違い

- Device-mapperもCOWによって差分を扱える
  - Device-mapperはデバイスのみ
  - OBDはデバイスとファイルが対象
- Unionfsも擬似的書き込み機能を提供する
  - Unionfsは仮想ファイルシステム
  - OBDは仮想ブロックデバイスでレイヤが異なる

Overlay Block Device
User-mode LinuxのUBDと同等の機能を,
リアルなハードウェアで動くKernelで実現する
仮想ブロックデバイス.

OBDの実装

- デバイスドライバ
  - Loopデバイス(kernel/device/block/loop.c)を元に実装(kernel 2.6.8.1, kernel2.6.7.11)
    - Copy On Writeの機能
    - COWファイル設定のioctl追加
    - procfsでの状況表示
- COWファイル設定コマンド
  - util-linuxのlosetupを元に実装
    (util-linux-2.12b)

ライブCDへの組み込み

ユーザがディレクトリツリーを自由に変更できるライブCDを作ろう.

ライブCDといえばKNOPPIX

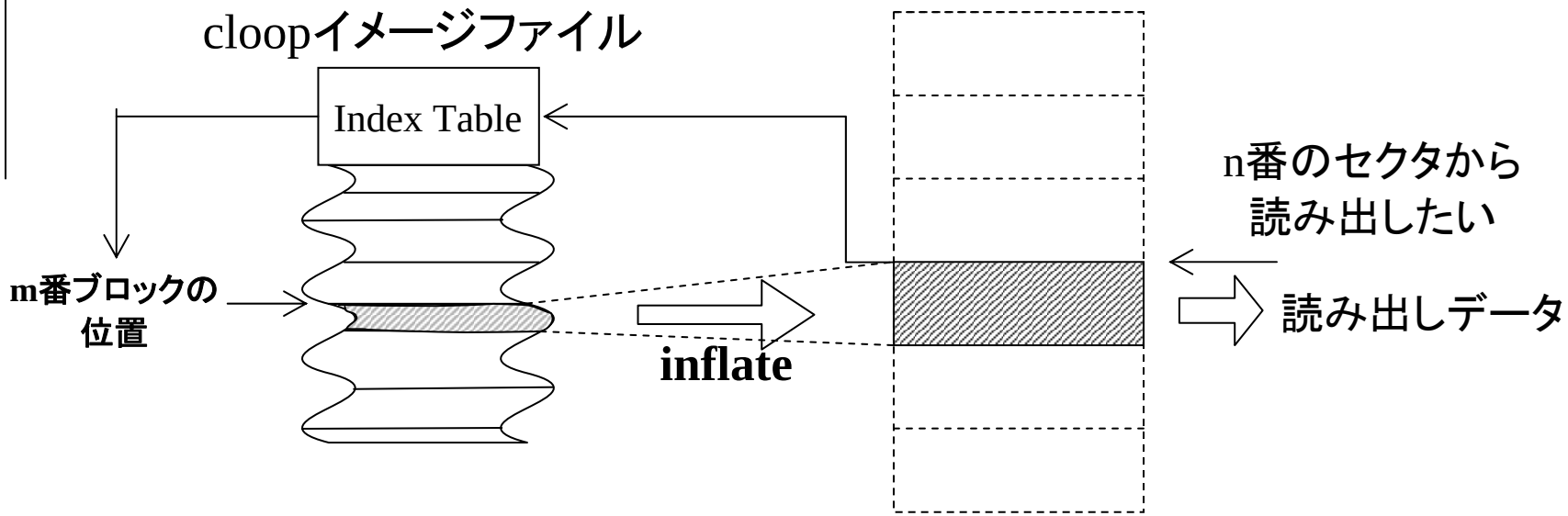
ライブCD KNOPPIX

- CD起動のDebian GNU/Linuxディストリビューション
  http://www.knoppix.net
  - Knopper氏が開発，AISTで日本語化
- AutoConfigによる強力なデバイス認識機能
  - Linuxで使用可能なハードウェアは自動設定
- cloop/zlibを用いた圧縮ループデバイス
  - 展開すると1.8GB，OS + AppliでCD一枚

Windowsマシンをすぐに
Linux Boxにできる

圧縮ループデバイスcloop

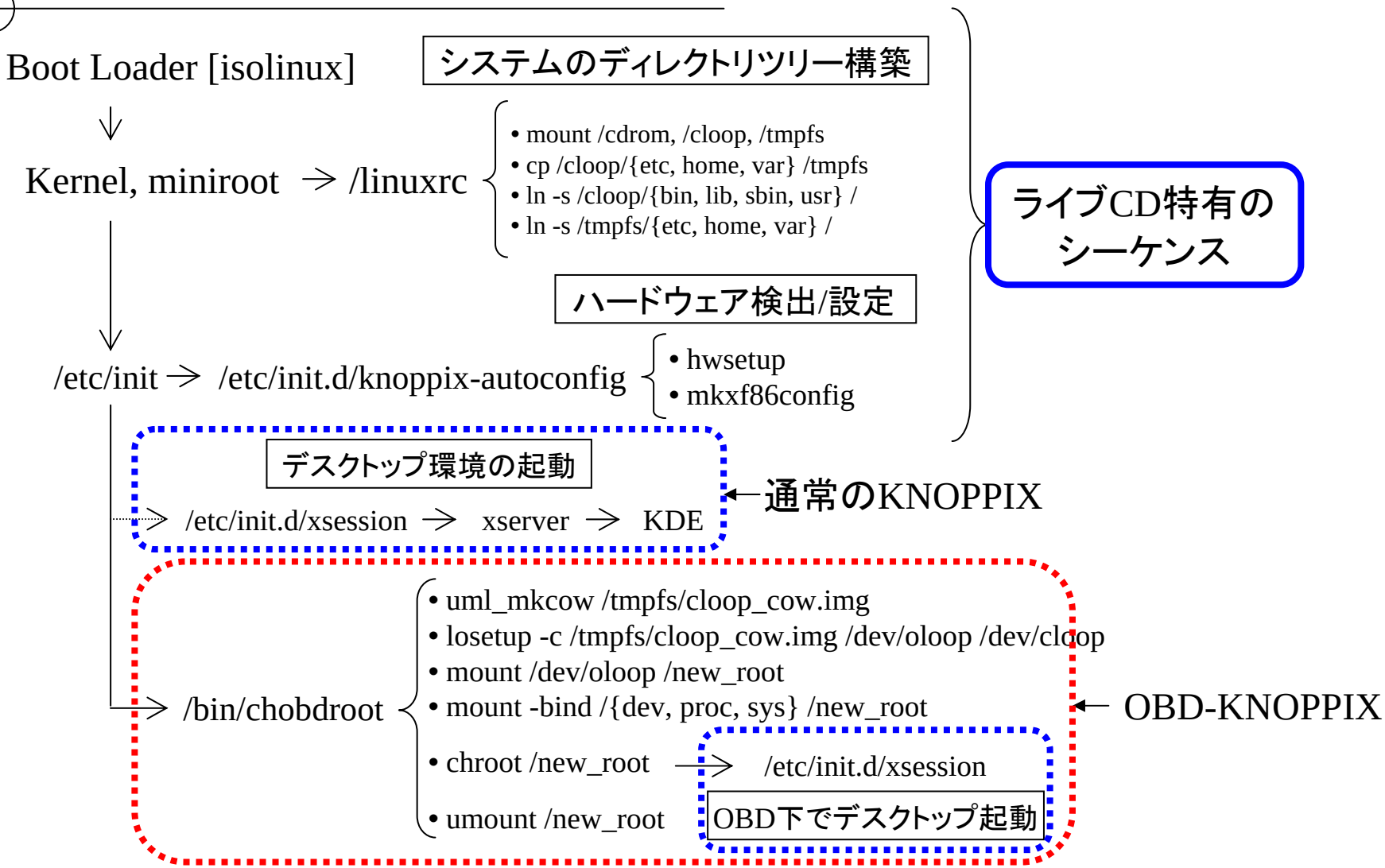
ループデバイス ＋ ブロック伸張機能



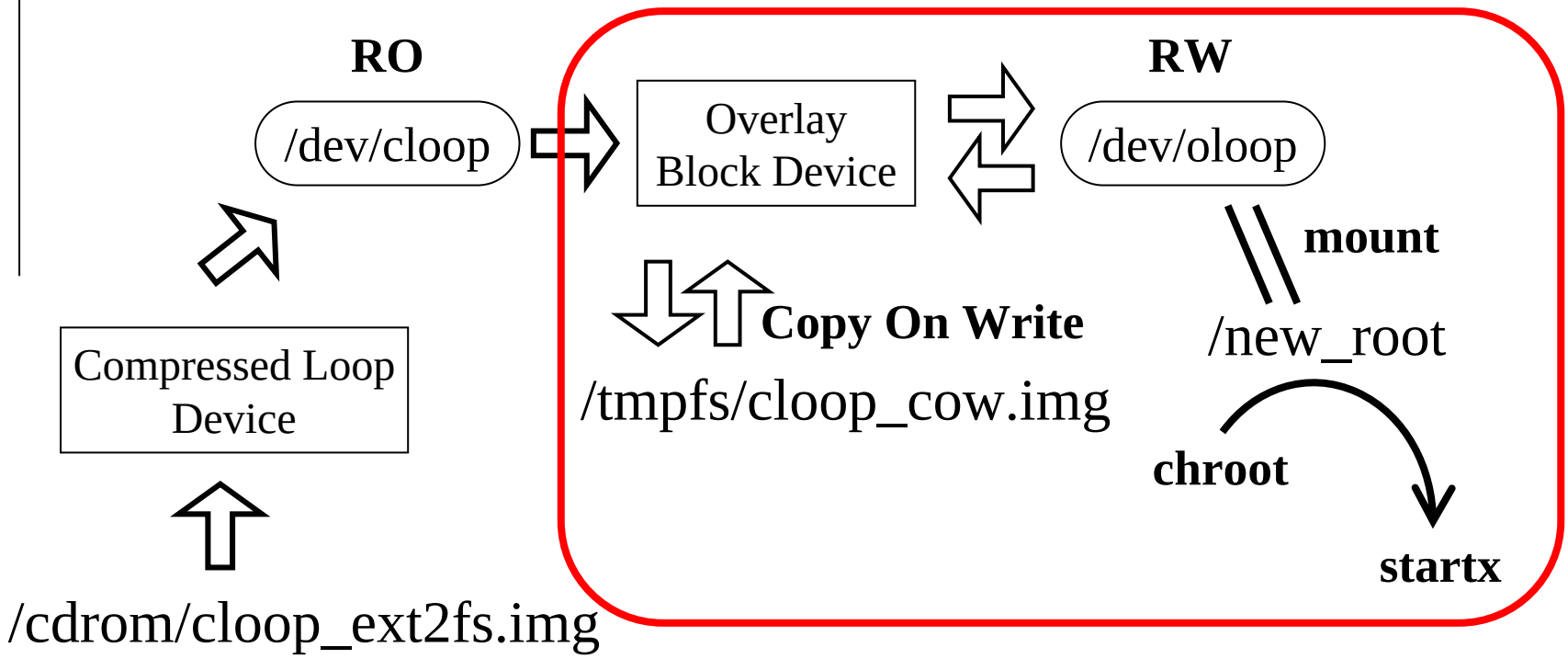
伸張(Read)だけで，圧縮(Write)できないデバイス

これにOverlay Block Deviceを適用する

OBD-KNOPPIXの起動シーケンス



OBD-KNOPPIXのディレクトリ構成



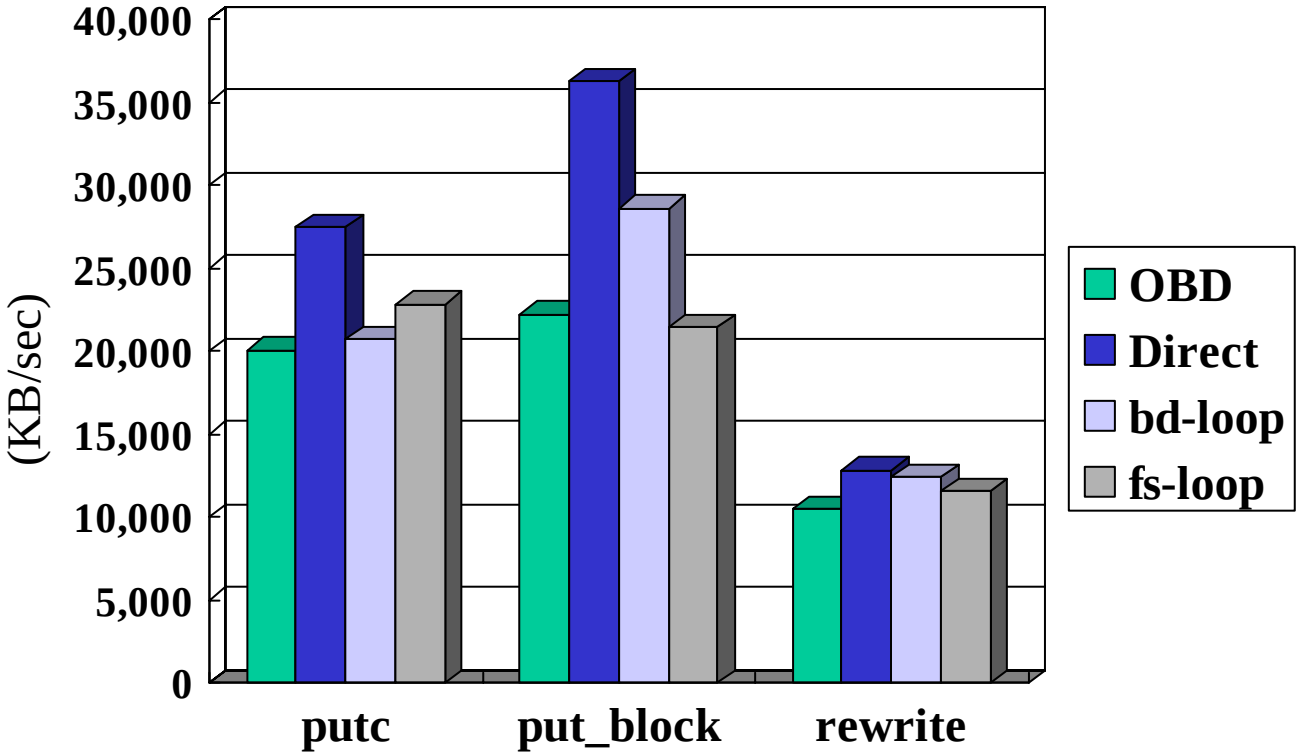
OBDの評価

- 仮想化レイヤを挟むことによるスループット劣化の確認
- 異なる階層での実装で擬似書き込み機能を提供するソフトウェア(Unionfs)との比較

OBDの評価(1)

- 仮想化レイヤを挟むことによるスループット劣化の確認
  - OBD, デバイスの直接参照(Direct), デバイスをloop経由で参照(bd-loop), ファイルシステム上のイメージをloop経由で参照(fs-loop)の4パターンに，読み書きのスループットを測定.
  - ファイルシステムはExt2FS
  - bonnie++1.03a, linux-2.6.8.1
  - CPU:P4-2GHz, MEM:512MB, HD:MAXTOR 5T040H4 (40.9GB, 7200RPM, 2MB CACHE, ATA-100), UltraDMA Mode 2

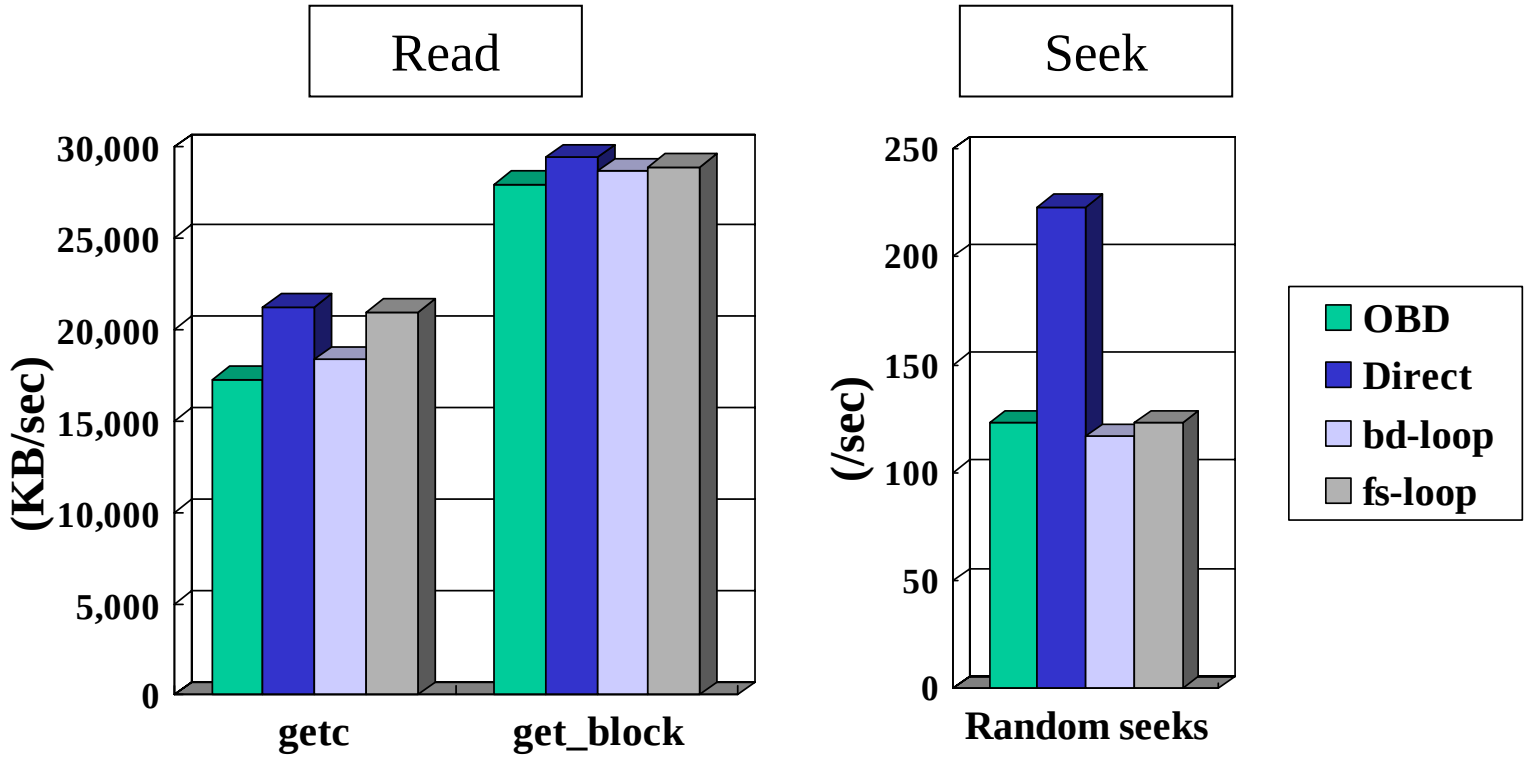
スループットの確認(Write)



デバイスの直接参照(Direct), デバイスをloop経由で参照(bd-loop),
ファイルシステム上のイメージをloop経由で参照(fs-loop)

OBDは，他の階層の組み合わせと比較しても遜色ない.

スループットの確認(Read & Seek)



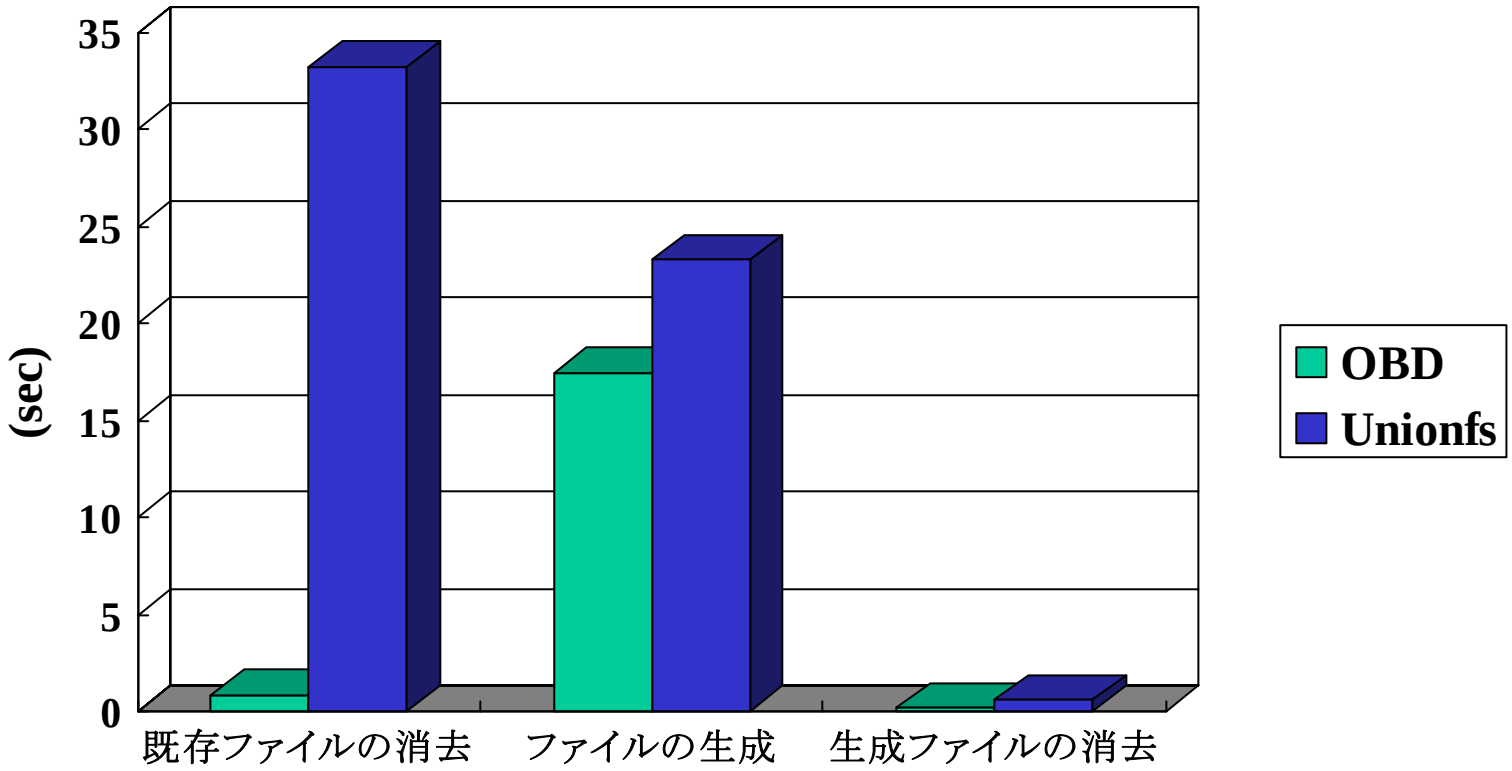
デバイスの直接参照(Direct), デバイスをloop経由で参照(bd-loop),
ファイルシステム上のイメージをloop経由で参照(fs-loop)

OBDは，他の階層の組み合わせと比較しても遜色ない.

OBDの評価(2)

- 異なる階層での実装で擬似書き込み機能を提供するソフトウェア(Unionfs)との比較
  - サイズ0のファイル50,000個を生成，消去に要する時間を測定.
  - CPU:P4-2GHz, MEM:512MB, HD:MAXTOR 5T040H4 (40.9GB, 7200RPM, 2MB CACHE, ATA-100)，UltraDMA Mode 2
  - tmpfs上に100MBのExt2FSイメージ
  - Unionfs 1.0.11, linux-2.6.11.7

Unionfsとの比較



OBDは，Unionfsよりも高速である.

OBDを組み込んだライブCDの評価

- パッケージの追加や削除など，HDにインストールされたLinuxディストリビューションと同様の操作が可能となった.
- Xのセッション終了後，COWファイルを退避し，次回起動時に適用することで，前回の更新内容が反映されたファイルシステムを再現できた.
- 差分ファイルが扱いにくい
  - cloopはブロックサイズに，伸張後のサイズを返す.
  - COWファイルはスパースで差分が零でも，論理的には1.8GB.

まとめ

- Copy On Write機能を実現する仮想デバイスOverlay Block Deviceを実装した.
- パフォーマンスを測定した結果，Loopデバイスと同程度のスループットを得た.
- ライブCDに組み込むことで，ファイルシステムをユーザが自由に変更できるようになった.

今後の検討/展開(1/2)

- ファイルフォーマットの検討
  - COWファイルはスパースでも，論理的にはサイズが大きく扱いにくい.
  - 差分情報を時系列で管理すると，ブロックデバイスのスナップショット機能が実現できる.
- cloopとの統合で透過的に圧縮される書き込み可能なloopデバイス?

今後の検討/展開(2/2)

- OBDを組み込んだライブCD公開？

slashdot.org, 2005/03/02, http://linux.slashdot.org/article.pl?sid=05/03/01/2329225

Knoppix 3.8 at CeBIT w/ Kernel 2.6, FF, and More

Operating Systems | **Posted by timothy on Wednesday March 02, @04:10AM from the slicing-edge dept.**

clsc writes *"The German tech news site Heise Online reports that **Knoppix 3.8** is being presented at CeBIT (Hall 9, Stand C39). Knoppix 3.8 has kernel 2.6 as default, KDE 3.3.2, OpenOffice 1.1.4, as well as... Firefox 1.0 and Thunderbird 1.0. There's also a really neato new thing involving **unionfs**. It seems to imply that you can change most anything on the running system, even as it is running from CD - and changes can be stored too (even on NTFS)."*

- Kernel PatchでKernel-mlデビュー？

From: roland <for_spam@gmx.de>

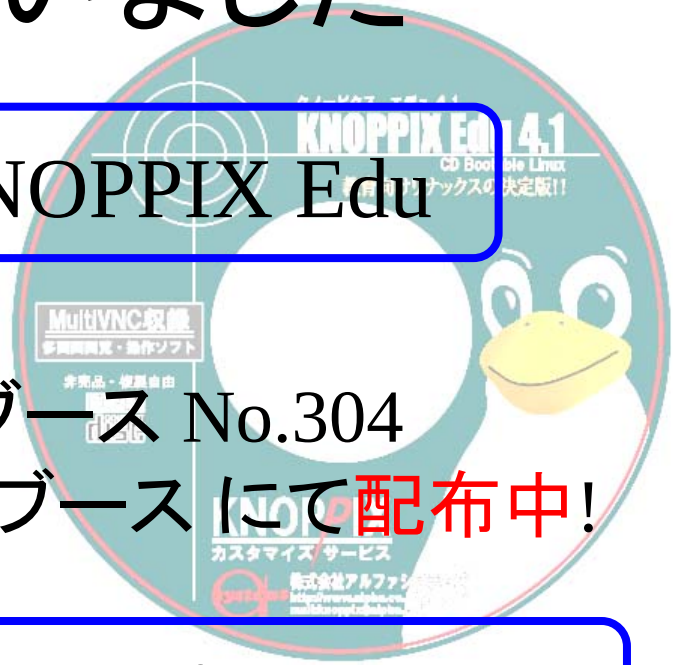
Date: 2005/05/26 9:52

Subject: cowloop - copy-on-write loop driver

To: linux-kernel@vger.kernel.org

ご清聴ありがとうございました

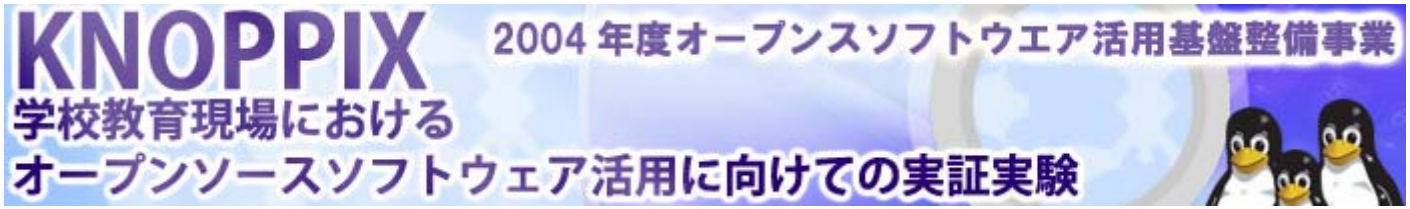
教育用途向けライブCD, KNOPPIX Edu



LinuxWorld Expo/Tokyo 2005

- 弊社(アルファシステムズ)出展ブース No.304
- .org パビリオン，KNOPPIX-Eduブース にて配布中!

教育におけるオープンソースデスクトップの実証実験



成果報告をもうすぐ公開 http://www.alpha.co.jp/knoppix/osse/