Empowered by Innovation NEC

ext4オンラインデフラグ実装に関する研究開発

藤田　朗 <a-fujita@rs.jp.nec.com>
NECソフトウェア東北株式会社
2009/9/18

目次

- Linuxにおけるファイルシステムの開発状況
- ext4,フラグメンテーションの問題
- ext4オンラインデフラグ、処理手順①ー⑤
- EXT4_IOC_MOVE_EXT ioctl,ブロック交換処理①ー⑤
- 性能測定結果
- ext4オンラインデフラグがマージされるまで
- 開発で大変だったこと
- 今後の予定

Linuxにおけるファイルシステムの研究開発状況

最近Linuxにマージされたファイルシステムと特徴

ext4 (2.6.28からLinuxに採用)
- ext2/3との互換性
- ファイルサイズ、FSサイズ拡大
- 遅延、連続ブロック割当て　など

(http://ext4.wiki.kernel.org/index.php/Main_Page より)

Btrfs　(2.6.29-rc1からLinuxに採用)
- コピーオンライトファイルシステム
- オンラインファイルシステムチェッカー
- マルチデバイスサポート　など

(http://btrfs.wiki.kernel.org/index.php/Main_Pageより)

NILFS2　(2.6.30-rc1からLinuxに採用)
- ログ構造型ファイルシステム
- 連続スナップショット
- システムクラッシュ後の高速リカバリ など

(http://www.nilfs.org/en/about_nilfs.html より)

ext4

ext3,ext4の諸元比較

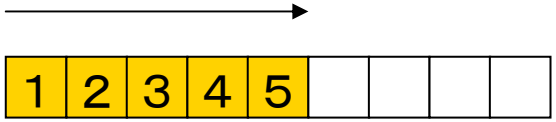
FS種別	最大ファイルサイズ	最大FSサイズ
ext3	2TB	16TB
ext4	16TB	1EB
	8倍！	65536倍！！

ext4で強化された機能

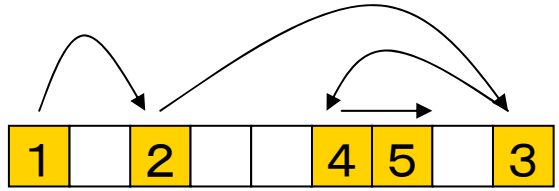
- エクステント形式によるデータ管理
- 複数ブロック割当て
- 遅延書き込み
- ジャーナルチェックサム

フラグメンテーションの問題

フラグメント(断片化)とは
　データがディスク上に分散して配置されていること



フラグメント無し



フラグメント有り

フラグメントによるデメリット

- データを読み込むときにディスクヘッドのシークが発生しI/O性能が劣化する
- 無駄なシークによりディスクの稼動部に負荷がかかり、耐久性が低下する

従来のext系のファイルシステムではオンラインでフラグメントを解消する方法がない

ext4オンラインデフラグ

- ext4オンラインデフラグはユーザー空間プログラムe4defragコマンドとカーネル空間のEXT4_IOC_MOVE_EXT ioctlを使って実現する

- e4defragコマンドはe2fsprogsパッケージで管理されている
- EXT4_IOC_MOVE_EXT ioctlは2.6.31からLinuxにマージされた

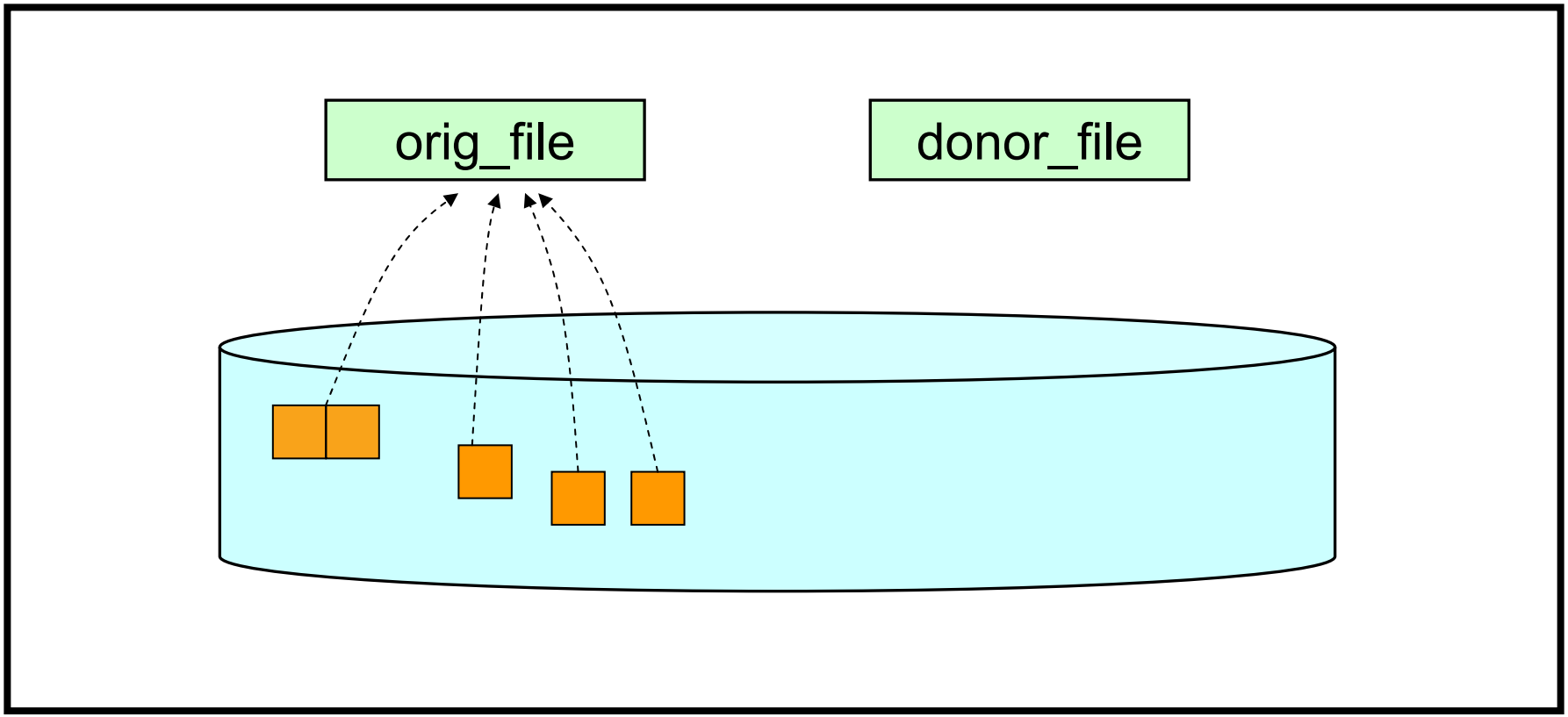
- e4defragの使用方法

  e4defrag [-c] [-v] file, directory, device

実行例

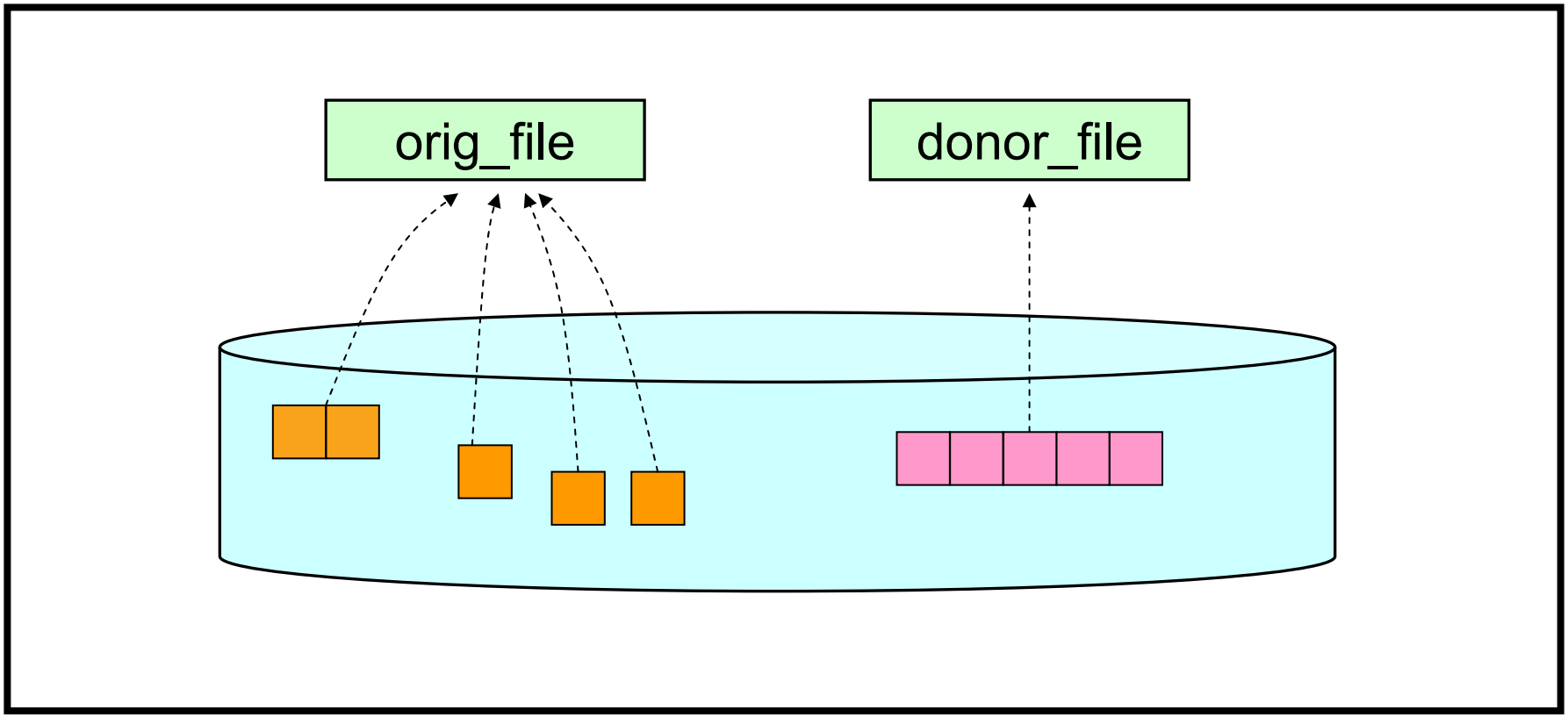
```
# e4defrag -v /mnt/mp1/file1
ext4 defragmentation for /mnt/mp1/file1
[1/1]/mnt/mp1/file1:            100%  extents: 6 -> 2        [ OK ]
 Success:                                  [1/1]
```

ext4オンラインデフラグの処理手順①



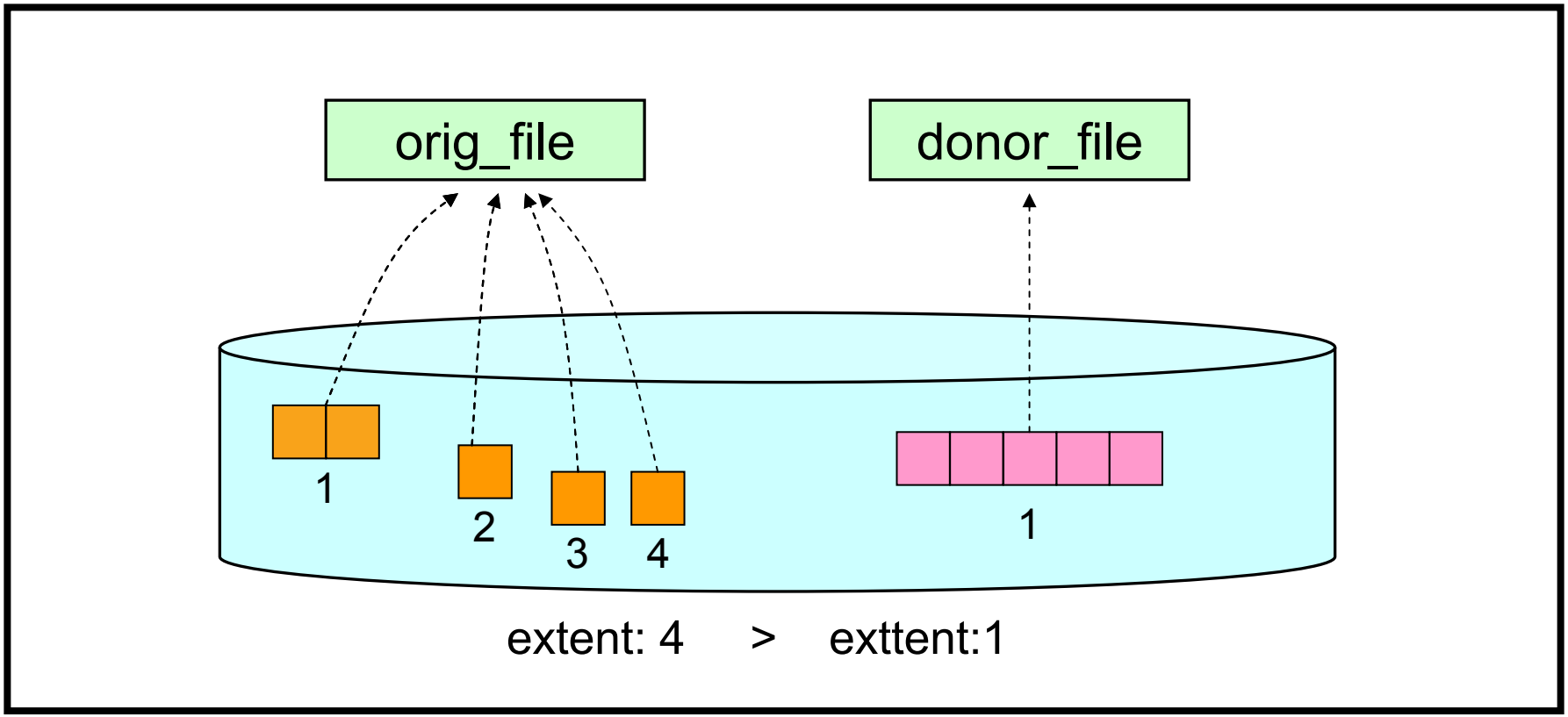
データブロック交換用のdonor_fileを作成

ext4オンラインデフラグの処理手順②



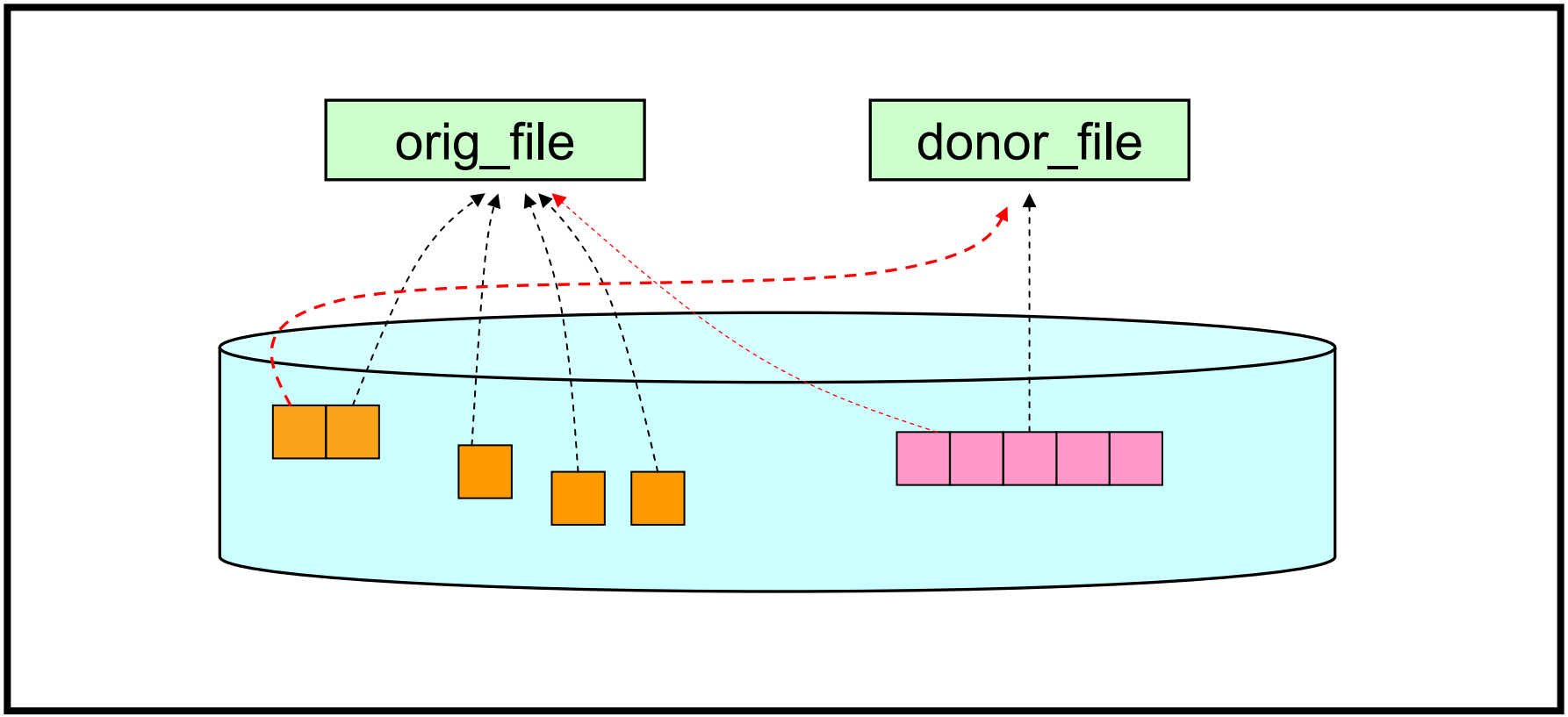
fallocateシステムコールでブロックを割り当てる

ext4オンラインデフラグの処理手順③



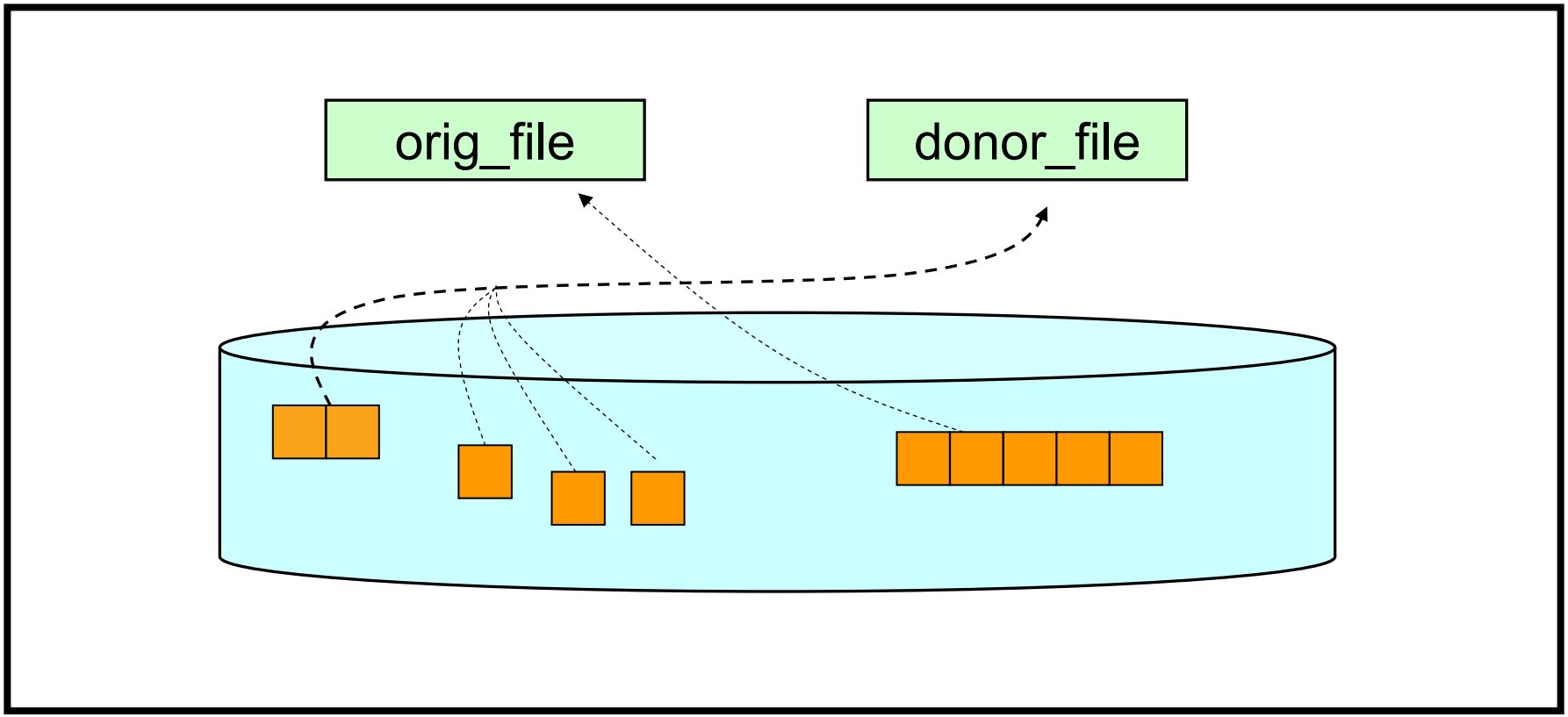
fiemapでextent数を比較

ext4オンラインデフラグの処理手順④



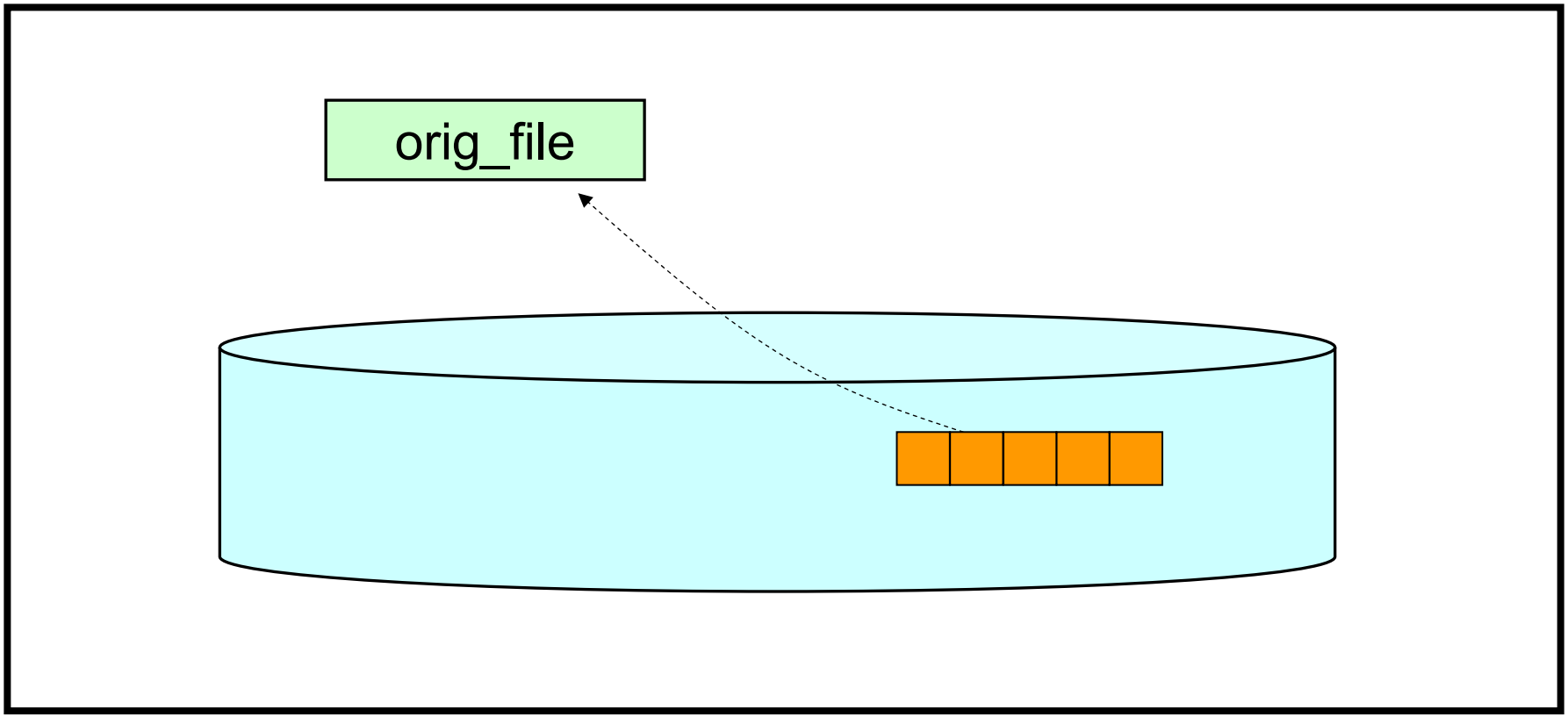
orig_file, donor_file**間でブロックを交換、**
orig_file**のデータを**donor_file**へ書き込む**

ext4オンラインデフラグの処理手順⑤



すべてのブロックが交換され、orig_file**の**
フラグメント数は減少

ext4オンラインデフラグの処理手順⑥



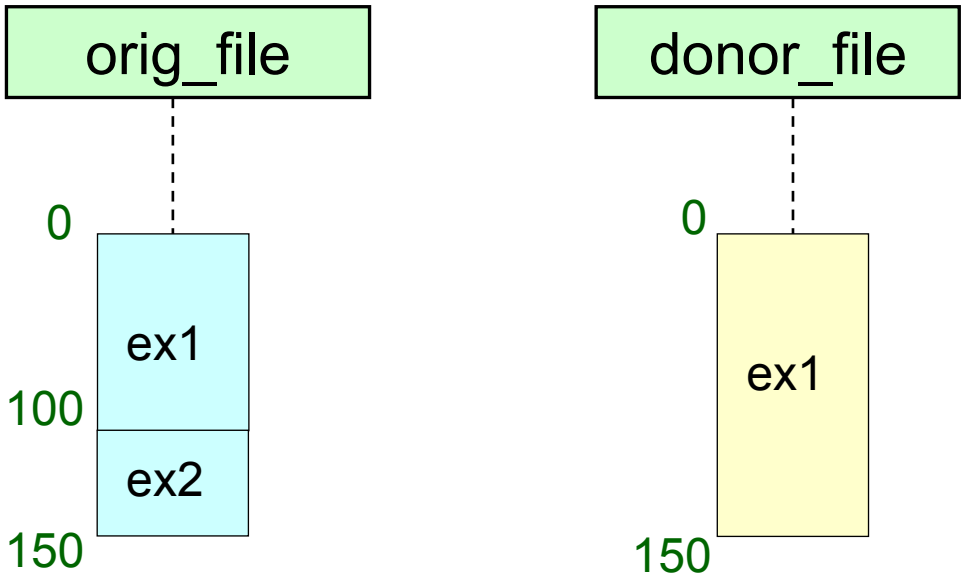
不要になったdonor_fileを削除

EXT4_IOC_MOVE_EXT ioctl

linux/fs/ext4/ext4.h

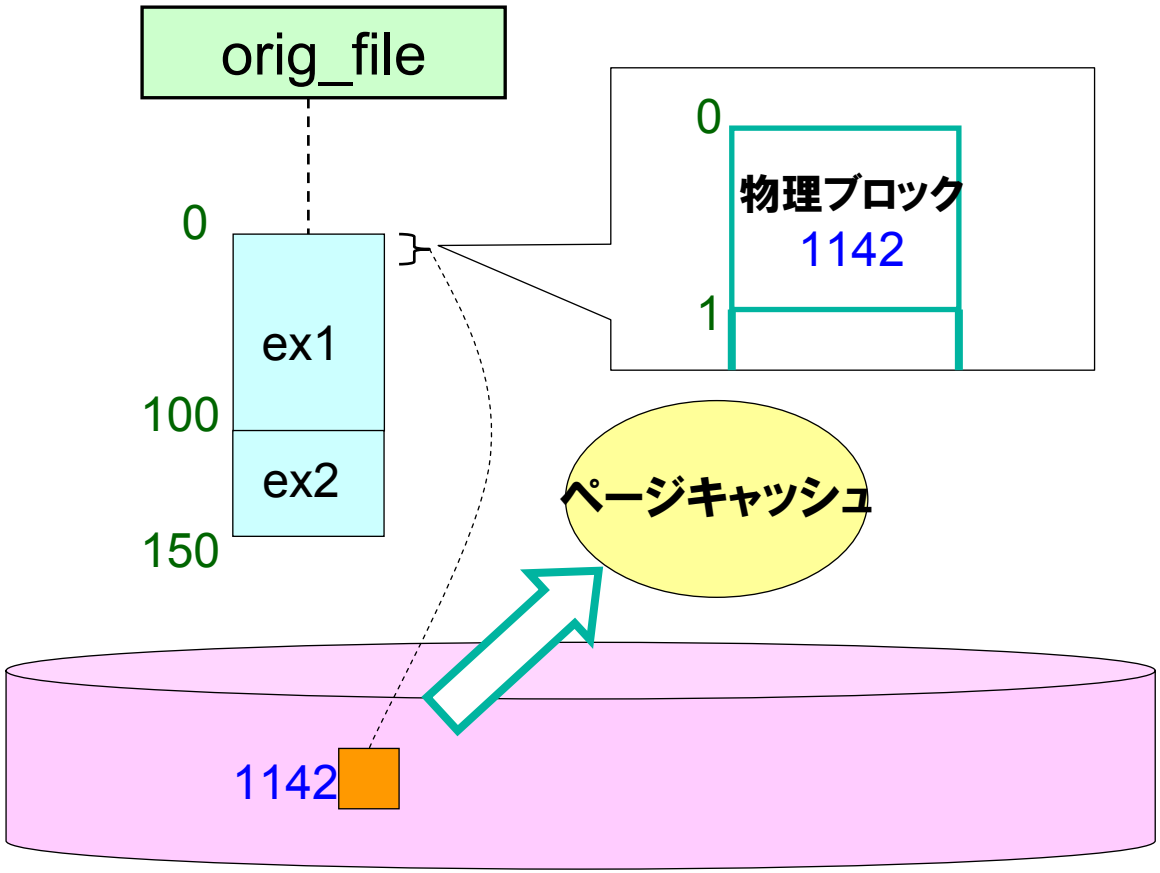
```
#define EXT4_IOC_MOVE_EXT  _IOWR('f', 15, struct move_extent)

struct move_extent {
        __u32 reserved;     未使用
        __u32 donor_fd;     donor file
        __u64 orig_start;   交換を開始するoriginal fileの論理オフセット
        __u64 donor_start;  交換を開始するdonor fileの論理オフセット
        __u64 len;          交換要求のブロック数
        __u64 moved_len;    交換に成功したブロック数
};
```

ブロック交換処理①



	論理オフセット(loff)	物理オフセット(poff)	ブロック長(len)
orig_file:ex1	0	1142	101
donor_file:ex1	0	6231	151

ブロック交換処理②

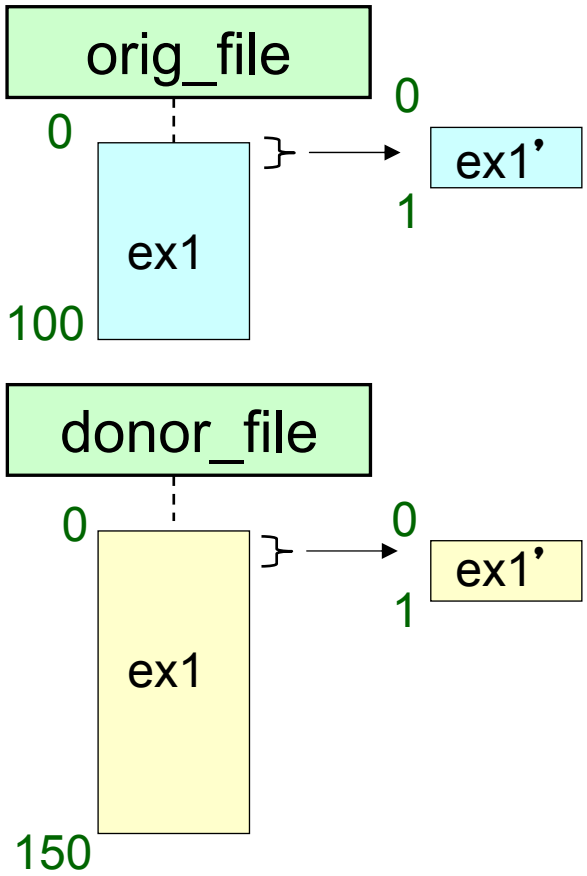


orig_file: ex 1	
loff	0
poff	1142
len	101

donor_file: ex 1	
loff	0
poff	6231
len	151

該当するブロックのデータをページサイズ分
ページキャッシュへ読み込む

ブロック交換処理③

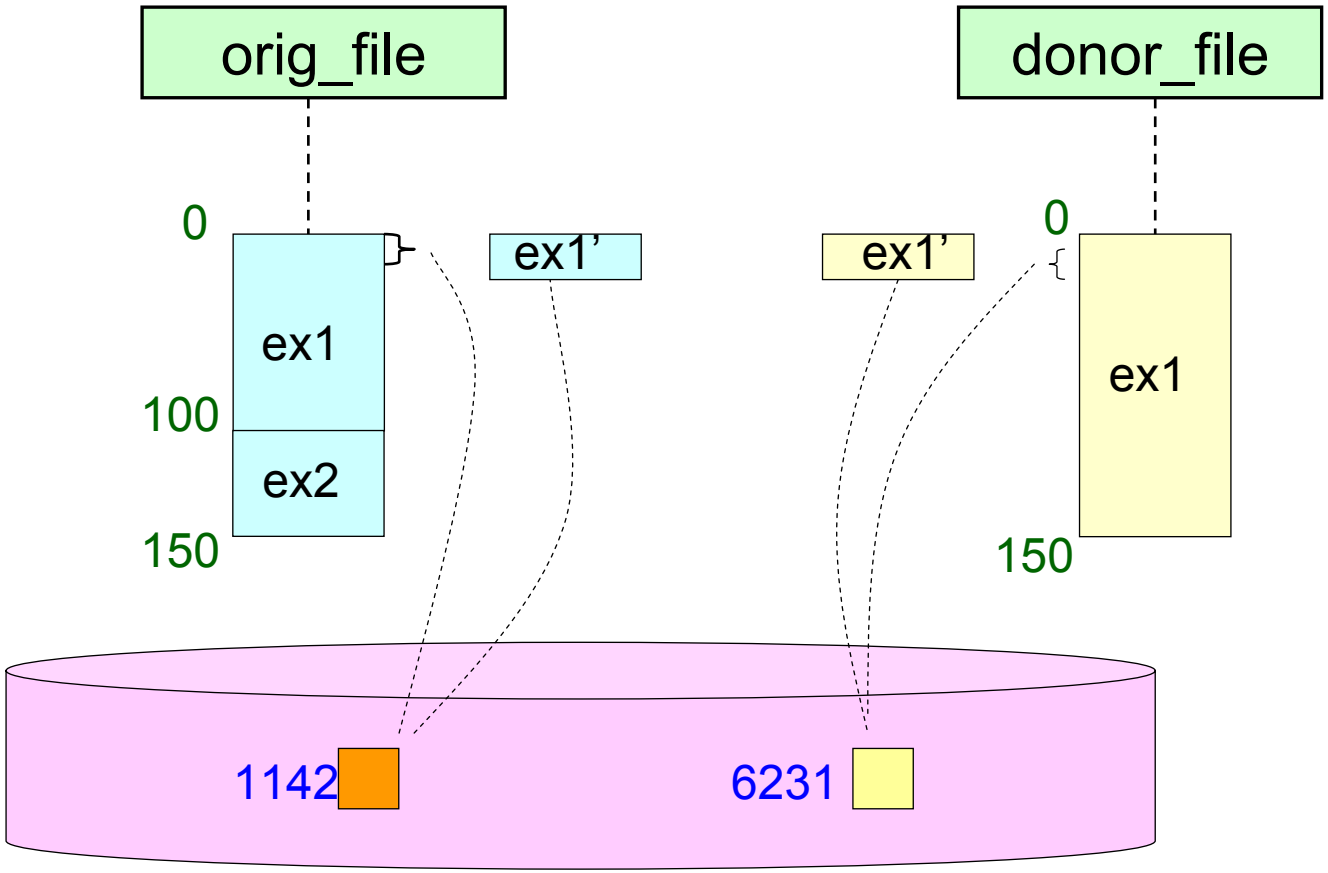


orig_file: ex 1		ex1'
loff	0	0
poff	1142	1142
len	101	1

donor_file: ex 1		ex1'
loff	0	0
poff	6231	6231
len	151	1

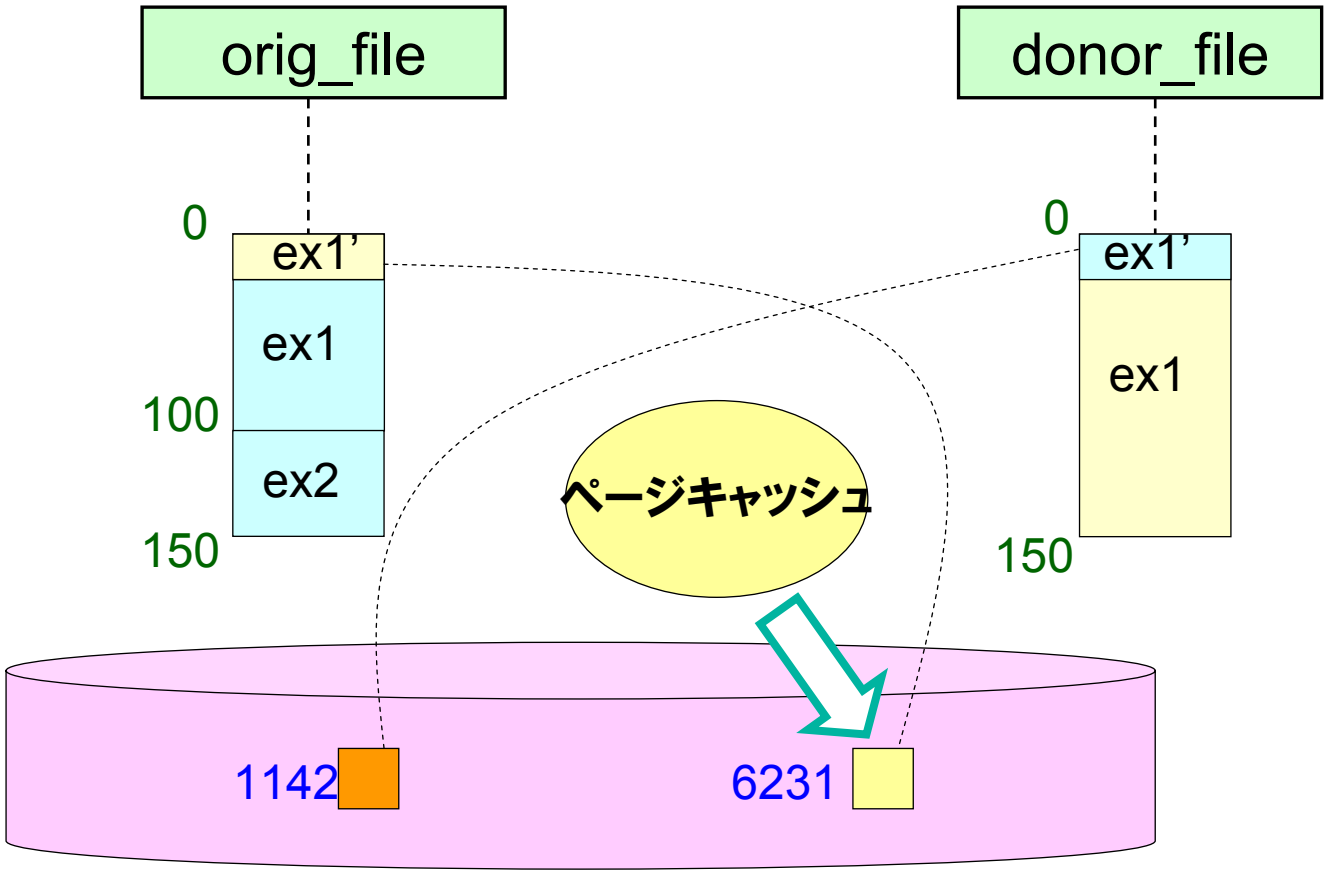
データ交換用の1ページサイズ分のエクステントを複製する

ブロック交換処理④



複製したエクステントを交換する

ブロック交換処理⑤



読み込んだページキャッシュを書き出す

(1142ブロック内容が6231ブロックへ書き出される)

ext4オンラインデフラグでの改善

性能測定の結果

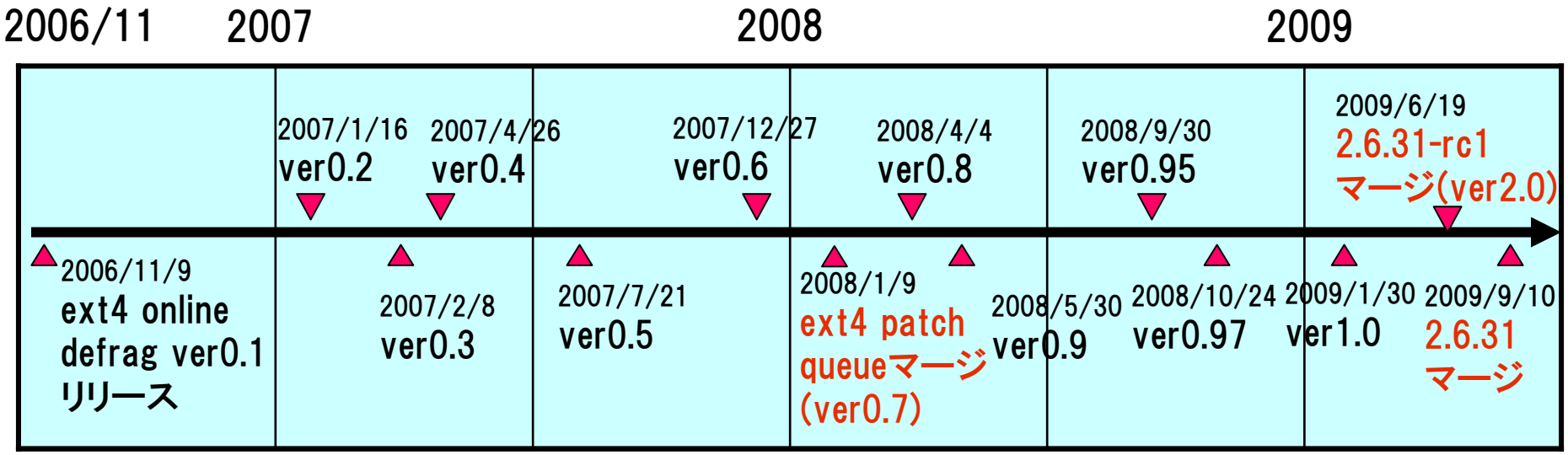
デフラグ	フラグメント数	読み込み(秒)	改善
前	257	39.893	15.6%
後	24	33.653	

評価環境　カーネル：2.6.31-rc3,アーキテクチャ：i386　CPU：Xeon 3.0GHz, メモリ：1GB
測定方法：dd コマンドを 10 並列で実行し作成した 2GB のファイルに対し、
ext4 オンラインデフラグ実行前後での cat コマンドの読み込み時間を測定。

-> フラグメントが解消されたことによりファイルの読み込み性能の改善

ext4オンラインデフラグがマージされるまで

開発開始からおよそ3年の歳月をかけ計13回(0.1->2.0)のバージョンアップを実施



発表等の対外的な活動

- OSDL Japan Symposium (2007), BoF (2006),
- USENIX Filesystem Workshop (2007),
- Ottawa Linux Symposium (2007),
- Linux Foundation Japan Symposium BoF (2008)

開発で大変だったこと

- 開発当初はext4自体開発中であり、ext4の仕様変更が多々あった
  -> 他の機能の変更の影響を受け、動作が不安定
  -> 退行評価TPなどを充実させ、仕様変更の影響を確認しながら開発を実施

- デフラグが高機能化し、コードが複雑になったため。コミュニティからレビューコメントがなかなかもらえなかった
  -> 基本機能が出来上がってから、付加機能の開発をすべきだった
  -> メーリングリストだけでなく、メンテナを含む関連する開発者にも依頼し、
  レビューしやすいようにパッチを機能毎に分割したほうが見てもらいやすい

- 開発のスピードが速く、英語で進められるため議論に参加するのが大変
  -> すべてを追えない場合は、開発に関連するものを取捨選択する
  -> 英文はシンプルを意識して作成し、パッチを元に議論する

今後の開発の予定

デフラグの更なる高機能化

1. 指定したディレクトリ配下のファイルを物理的に近い領域に配置する機能

   -> 特定のアプリケーションの性能向上

2. FSの空き領域が少ない状態でも指定したファイルのフラグメントを優先的に解消する機能

   -> アクセス頻度の高いファイルのフラグメントを解消

->新機能の追加のためにext4ブロックアロケート機能の強化提案中。

他プロジェクト(OHSM)の開発者達と連携し、Linuxへのマージを促進する

参考

- The Linux Kernel Archives
  http://kernel.org/
- Mailing list ARChives (linux-ext4)
  http://marc.info/?l=linux-ext4&r=1&w=2
- ext4 patch queue
  http://repo.or.cz/w/ext4-patch-queue.git
- e2fsprogs
  http://git.kernel.org/?p=fs/ext2/e2fsprogs.git;a=summary
- Takashi Sato. ext4 online defragmentation. In Ottawa Linux Symposium (2007)
  https://ols2006.108.redhat.com/2007/Reprints/sato-Reprint.pdf
- Ext4 (and Ext2/Ext3) Wiki
  http://ext4.wiki.kernel.org/index.php/Main_Page

Empowered by Innovation