UNICUS

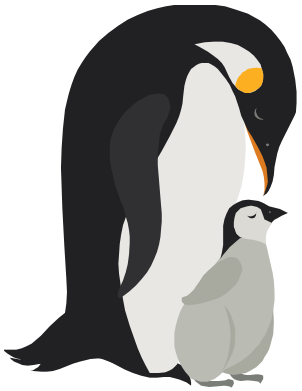
NILFS2ソフトウェアツールの開発

関場 治朗
株式会社ウニークス
jir@unicus.jp

Agenda

- 背景
- NILFS2紹介
- 個々のツールの対応
  - GRUB2
  - blkid
  - Parted
  - Palimpsest
- まとめと今後の課題

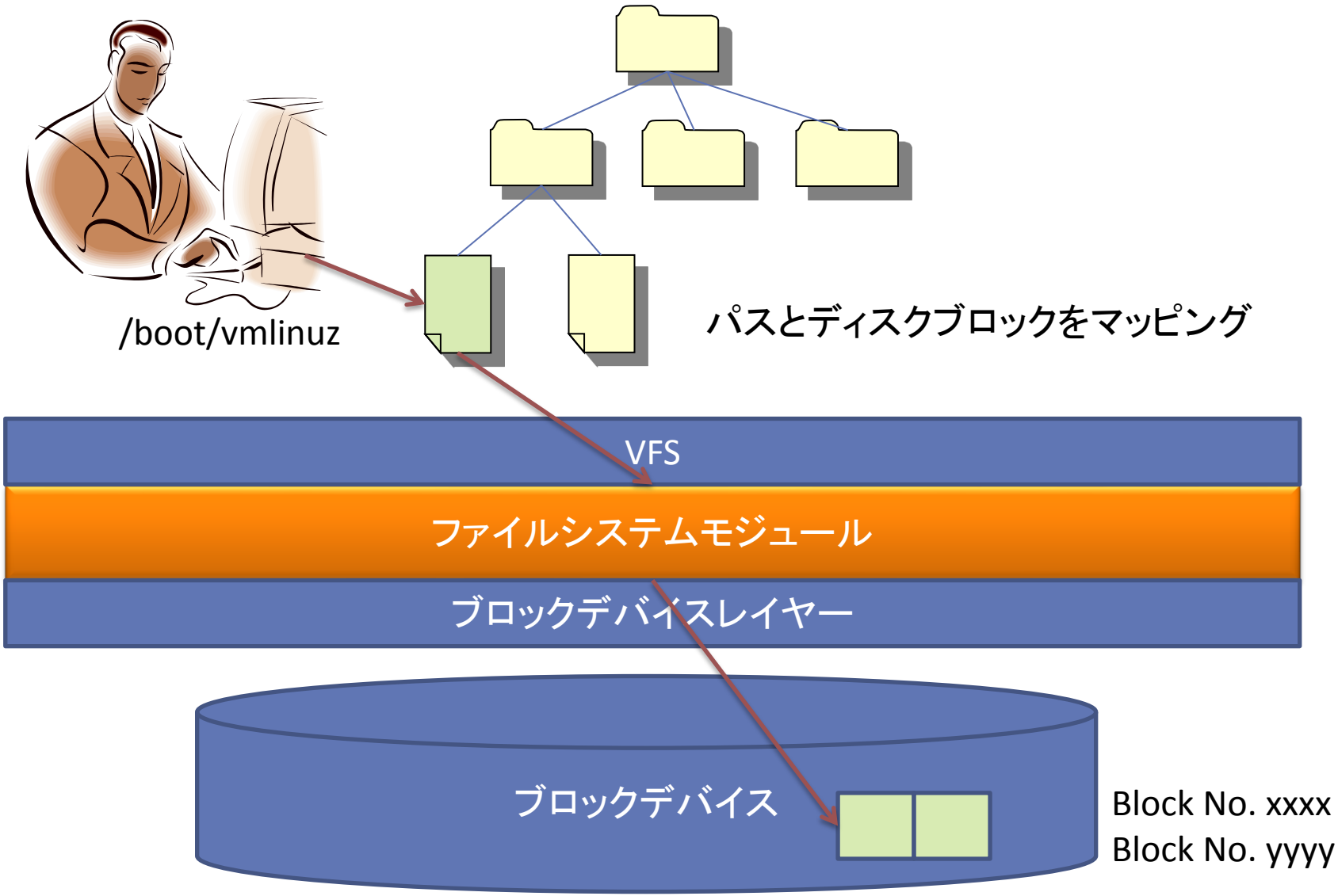
BACKGROUND



NILFS2とは？

- NTTサイバースペース研究所が開発した filesystem
  - New implementation of a log-structured file system version 2
- Linux vanilla kernel に 2.6.30 から同梱
- http://nilfs.org/

Filesystem


/boot/vmlinuz
パスとディスクブロックをマッピング
VFS
ファイルシステムモジュール
ブロックデバイスレイヤー
ブロックデバイス
Block No. xxxx
Block No. yyyy

Linuxで利用できるfilesystem

ディスクユーティリティで使用可能 - 8種

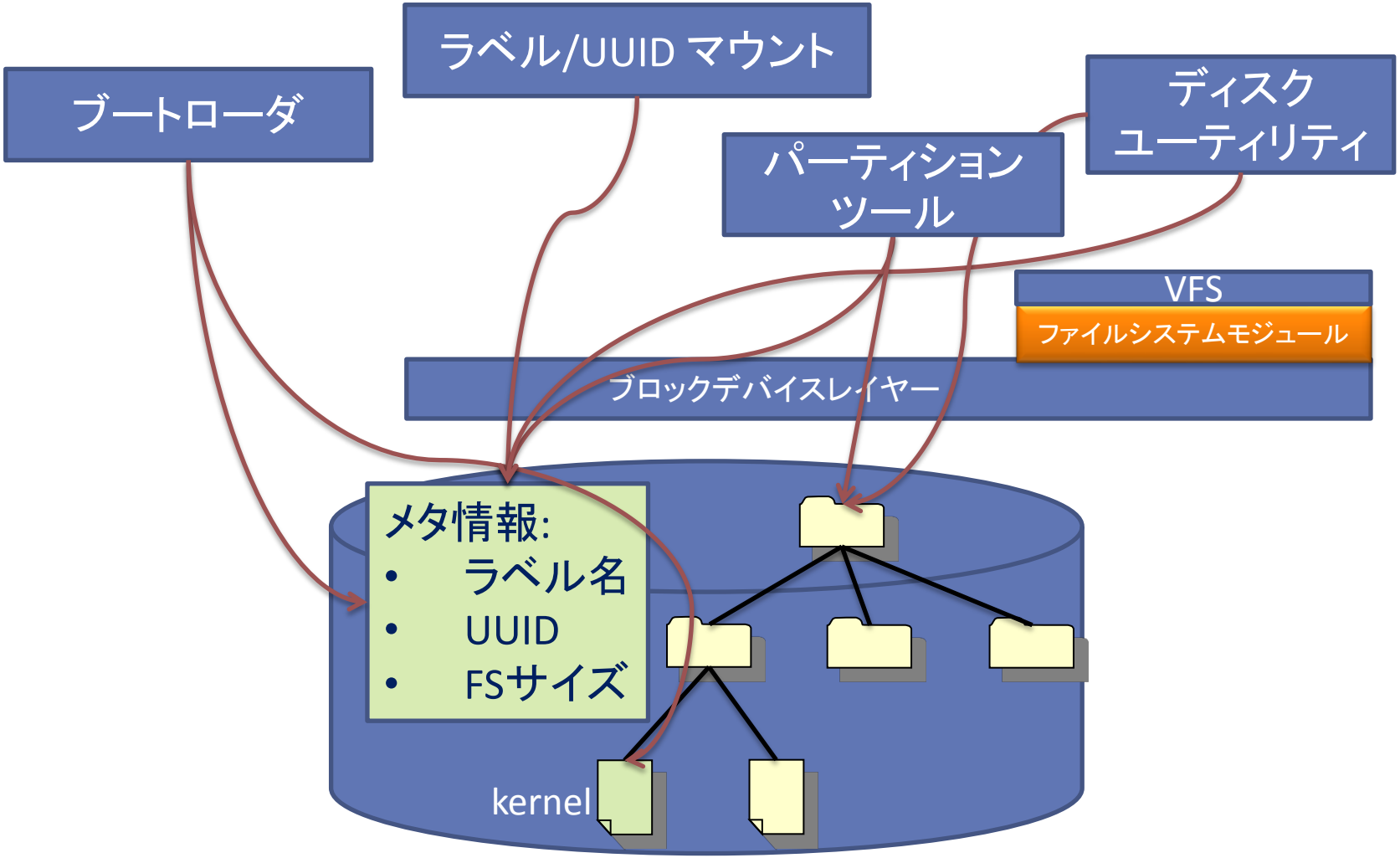
このギャップは？

Kernel モジュール 約30種

ext3 ext4 xfs nilfs btrfs logfs exofs

UNICUS

ブロックデバイスに直接アクセスするユーティリティ



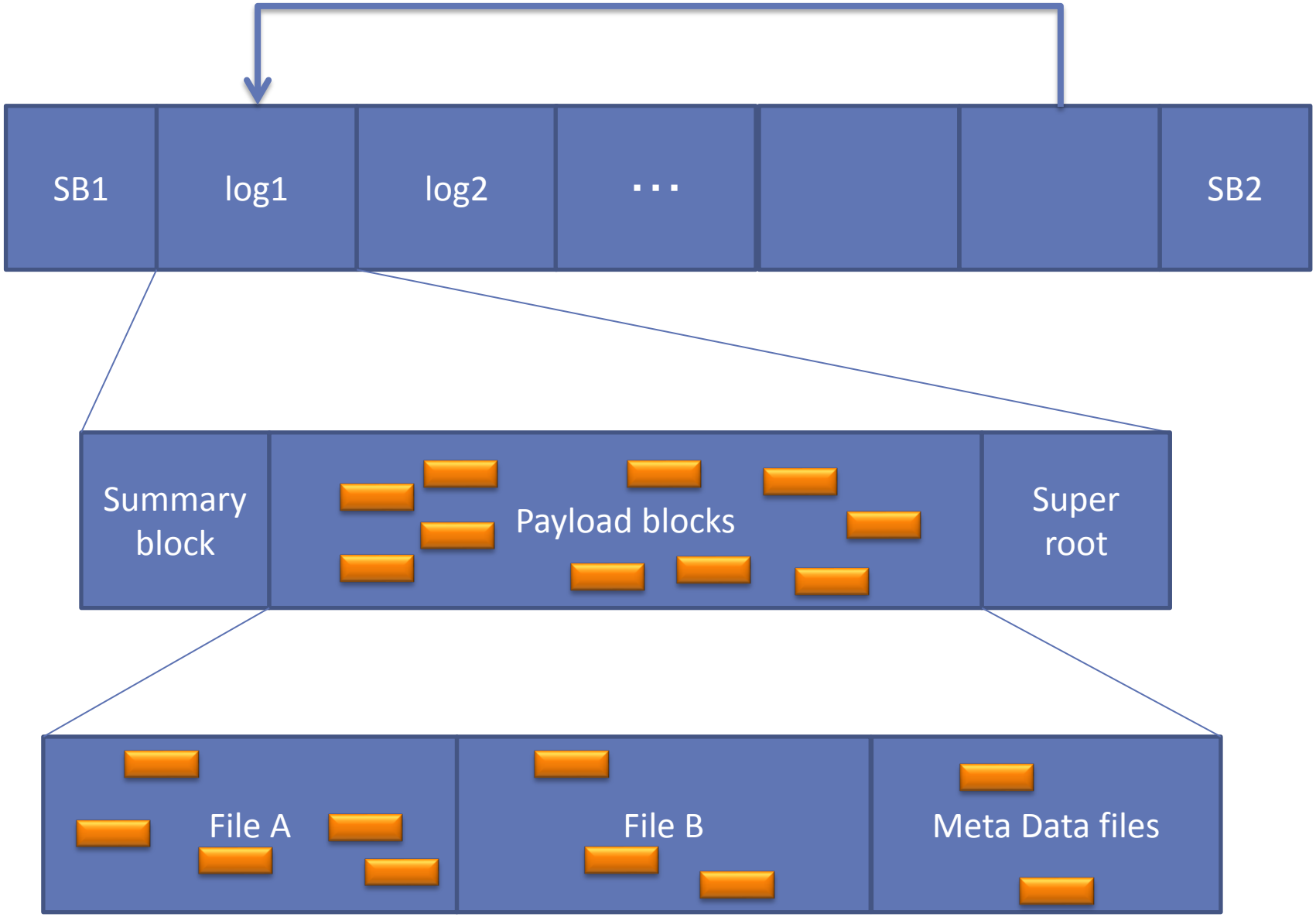
NILFS2

NILFS2の特徴

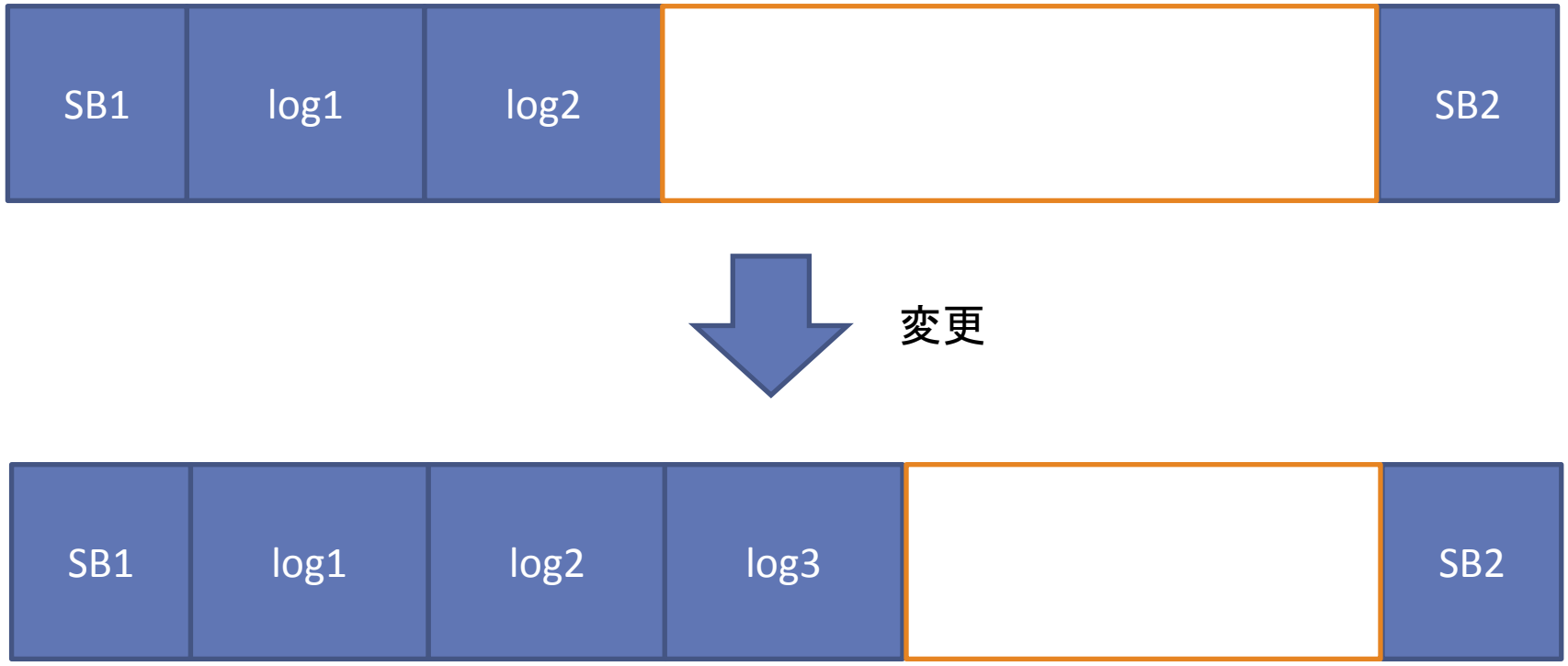
- 追記型ファイルシステム
- ファイルシステムの履歴を保存する連続的checkpoint
- 空き領域作成の為のガベージコレクタ
- ファイルシステムの状態を保存するスナップショット

UNICUS

NILFS2 ディスクフォーマット

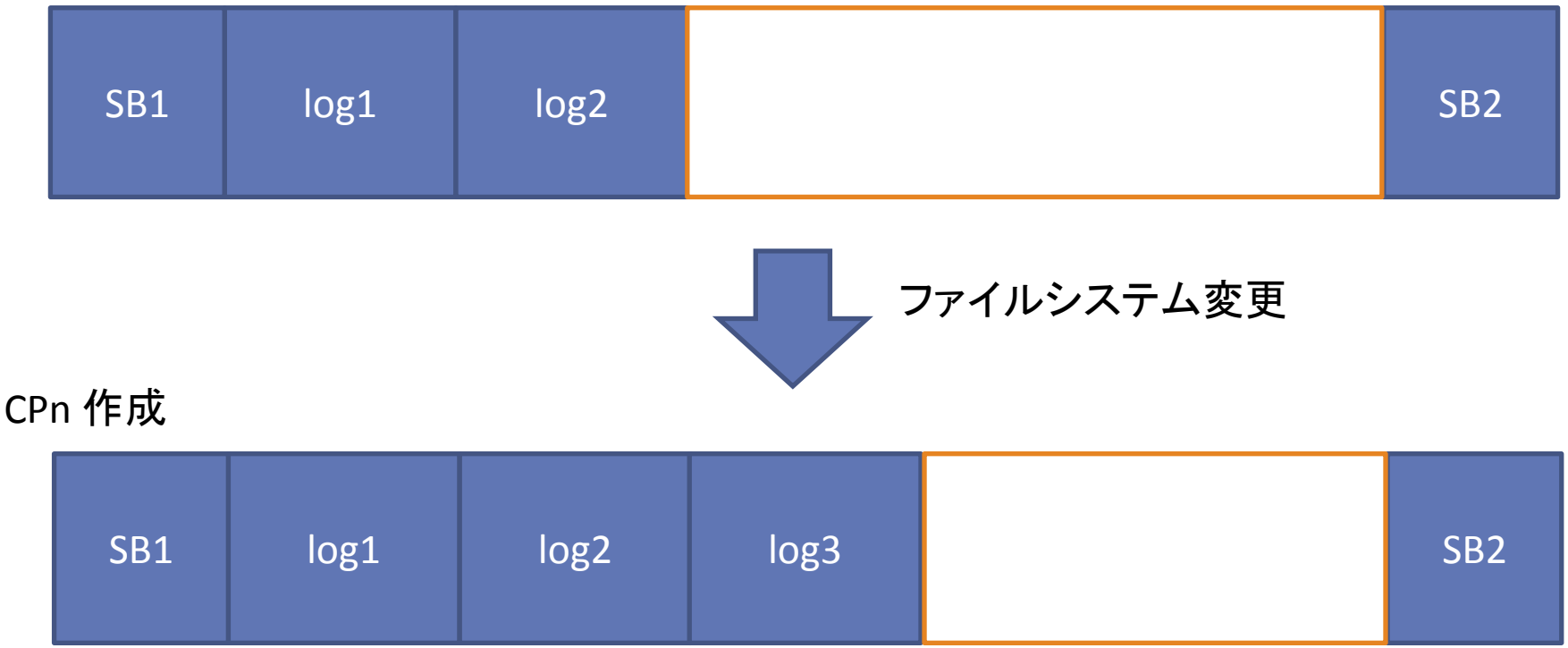


追記型ファイルシステム



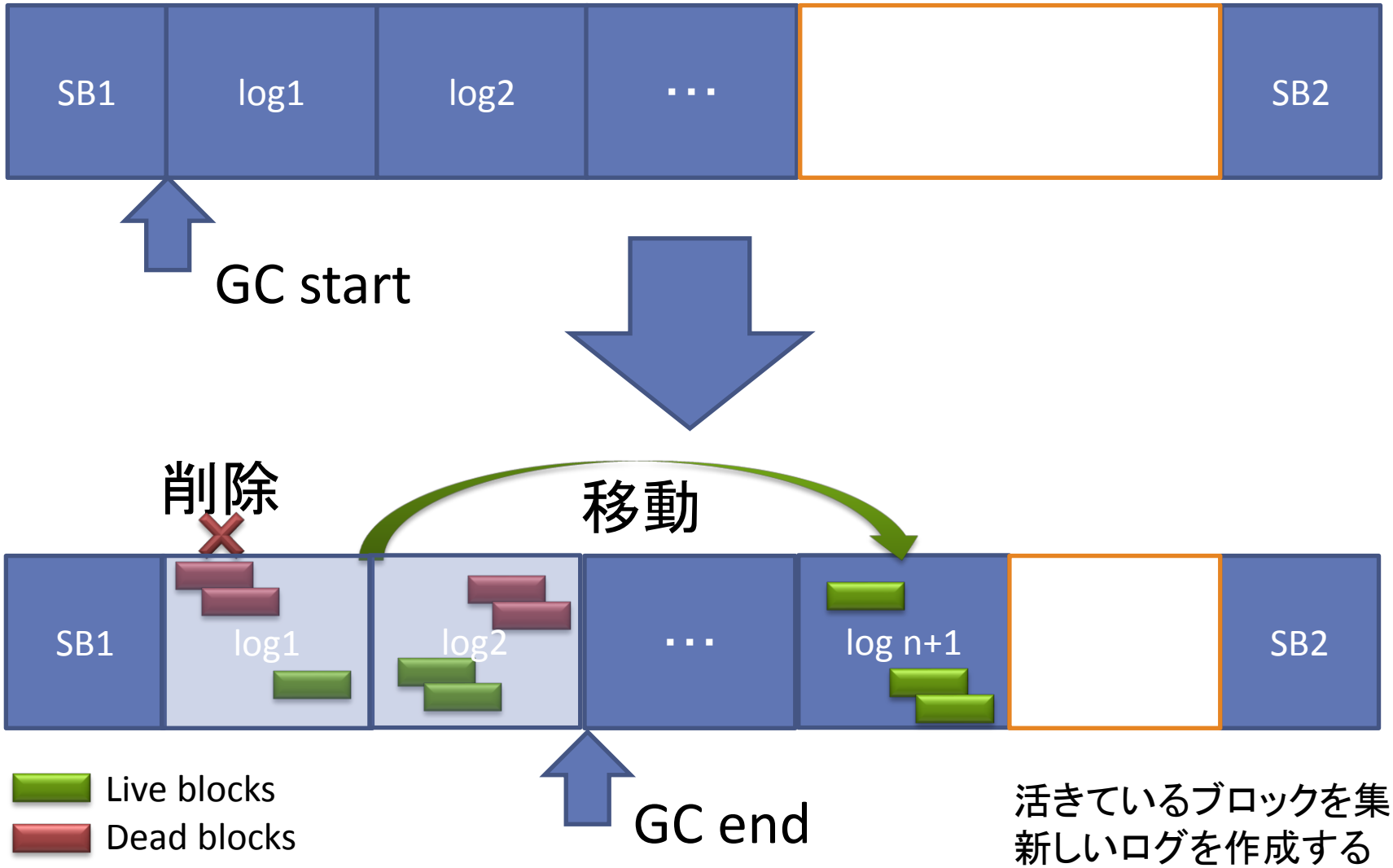
全ての変更は追記で行われる
スーパーブロック以外の書き換えは行われない

連続的checkpoint



明示的な操作は必要ない

ガベージコレクタ



活きているブロックを集め
新しいログを作成する

状態を保存するスナップショット

CPn

ブロックを活きている状態に固定化

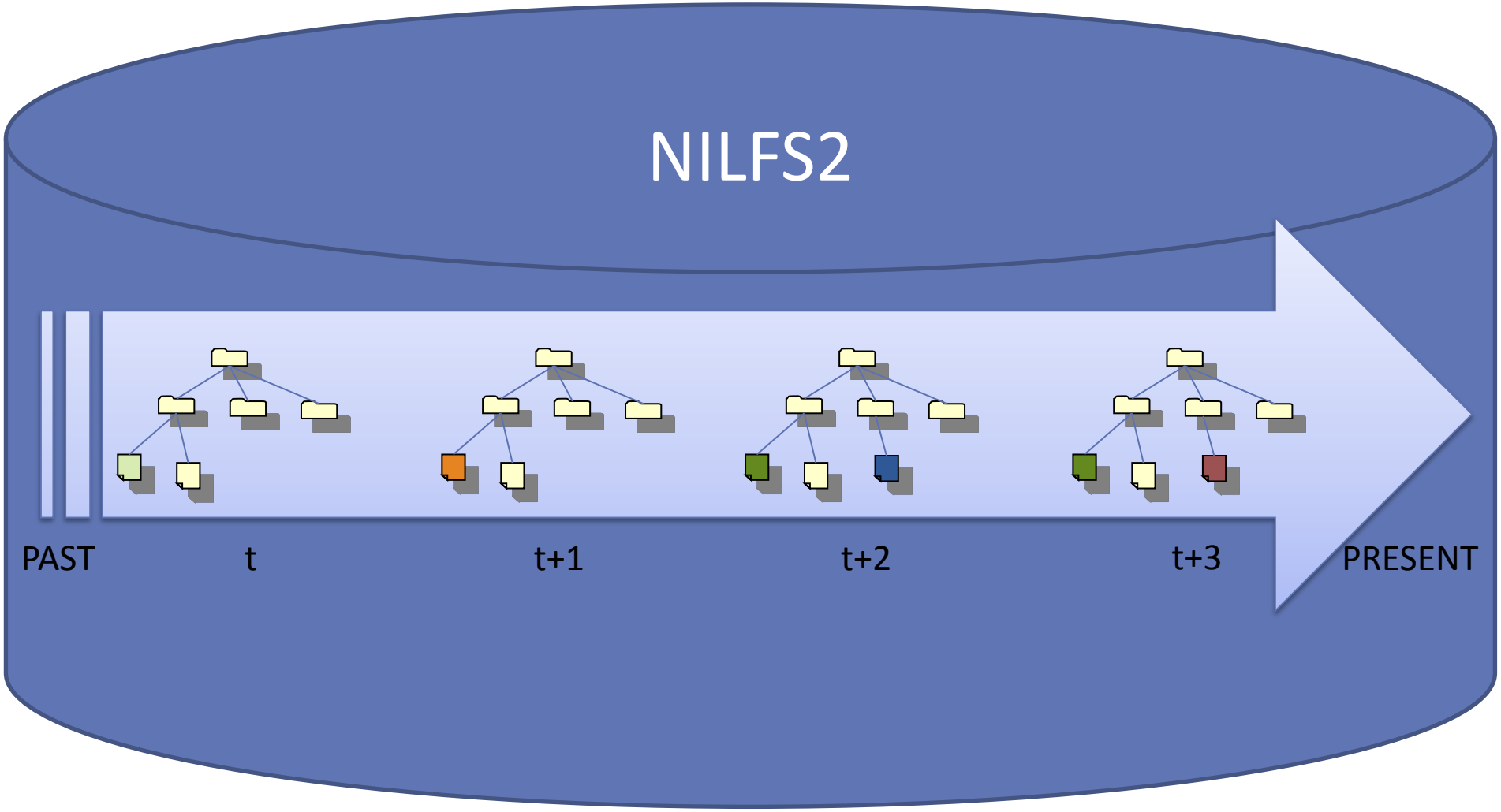
CPn-1のスナップショット化

Live blocks
Dead blocks

スナップショットで状態が永続化
どの時点のcheckpointもスナップショット化可能

UNICUS

NILFS2の特徴まとめ



過去の履歴が常に保存

BOOTLOADER – GNU GRUB2

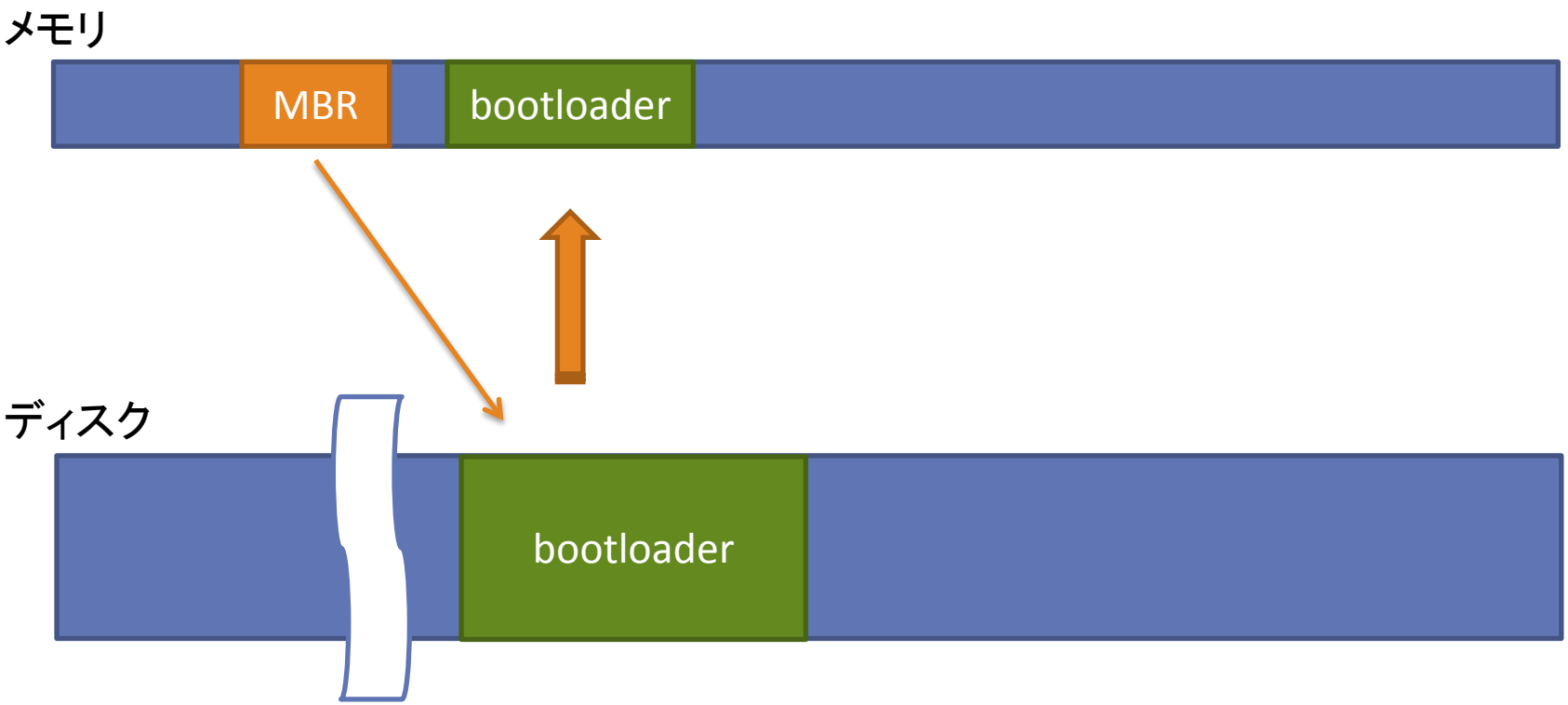
ブートローダ

- 最初に動くプログラム
- Kernelをメモリにロード
- Linux で利用されているもの
  - PC では lilo/grub/grub2 など
  - 他では yaboot/grub2 or それ以外のboot loader
- 今回は PC で grub2 を説明

ブートストラップ

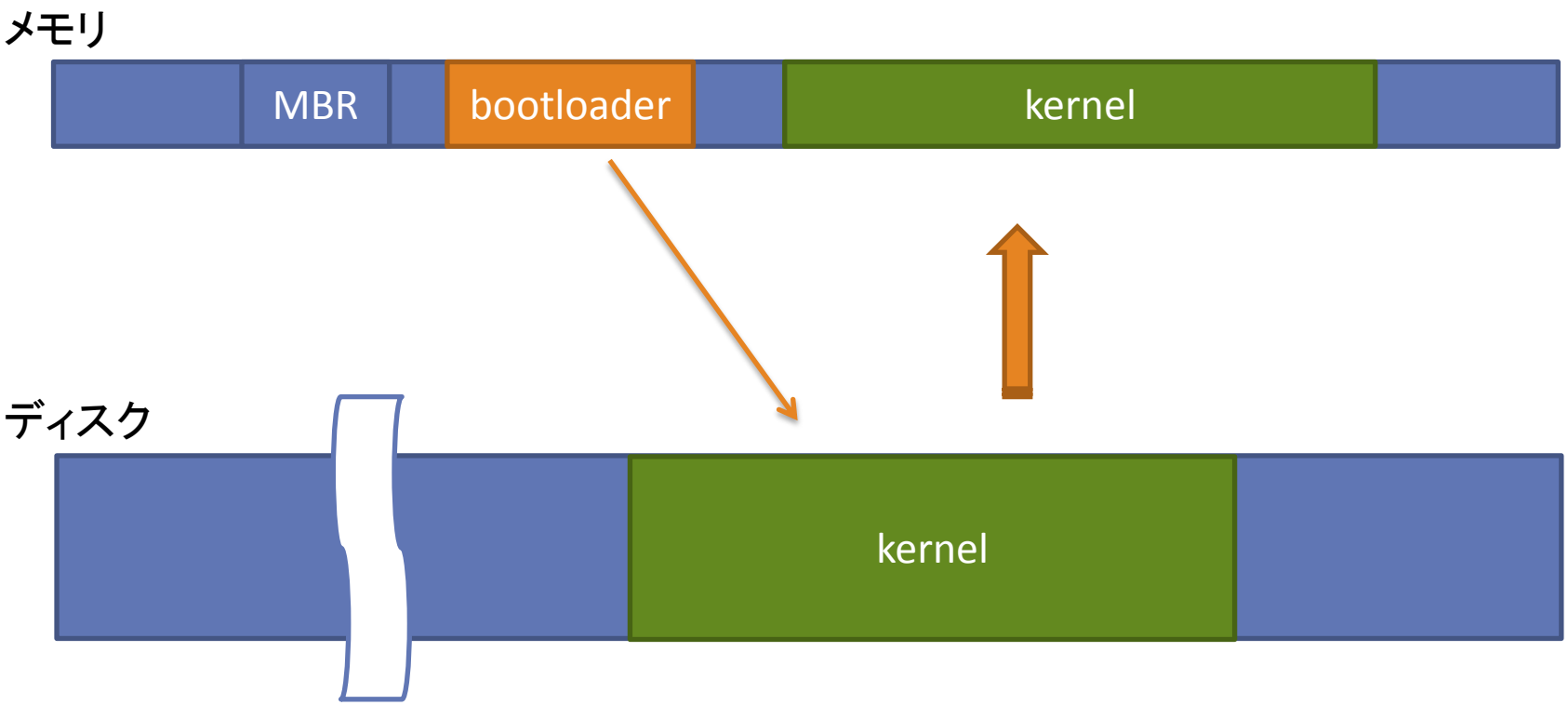
BISOがMBRをメモリにロード

ブートストラップ



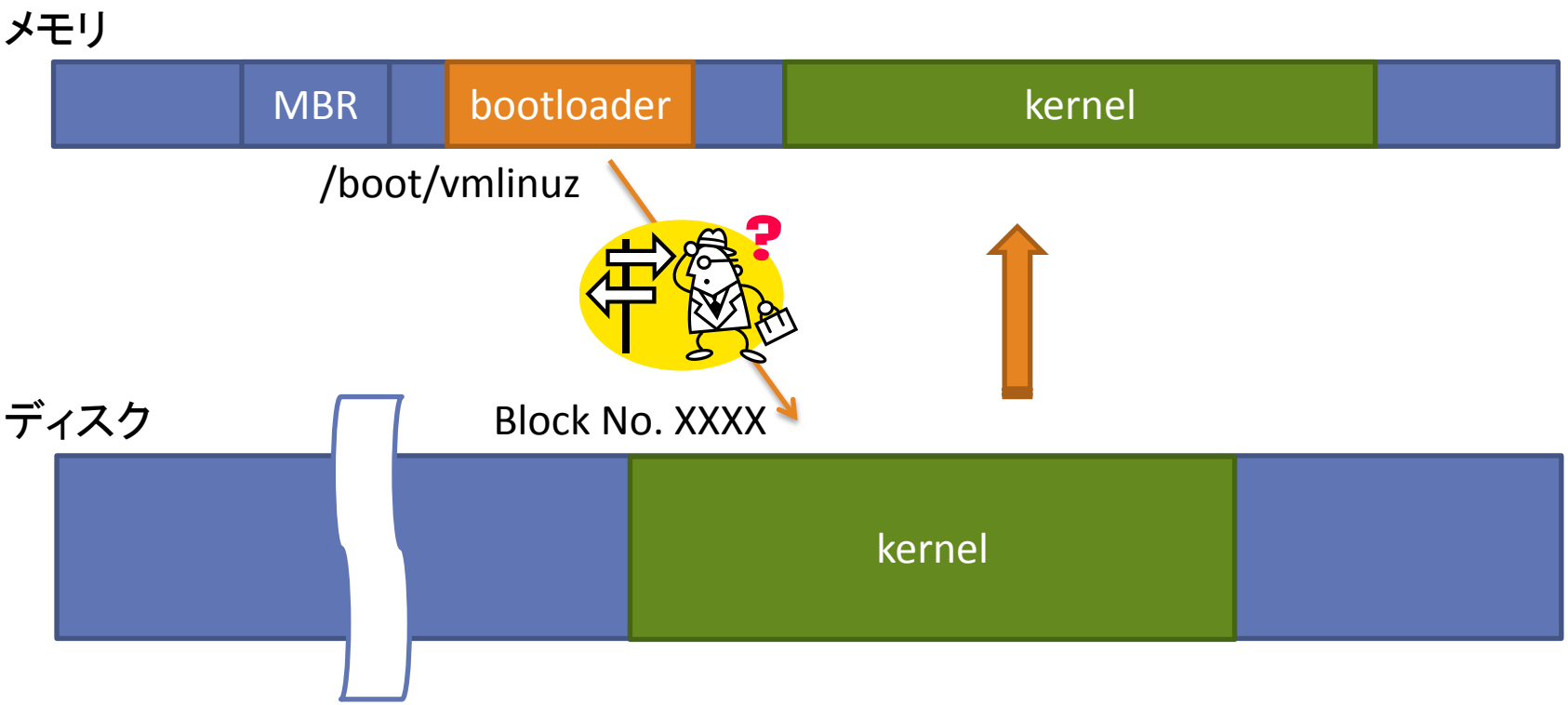
MBRがbootloaderをメモリにロード

ブートストラップ



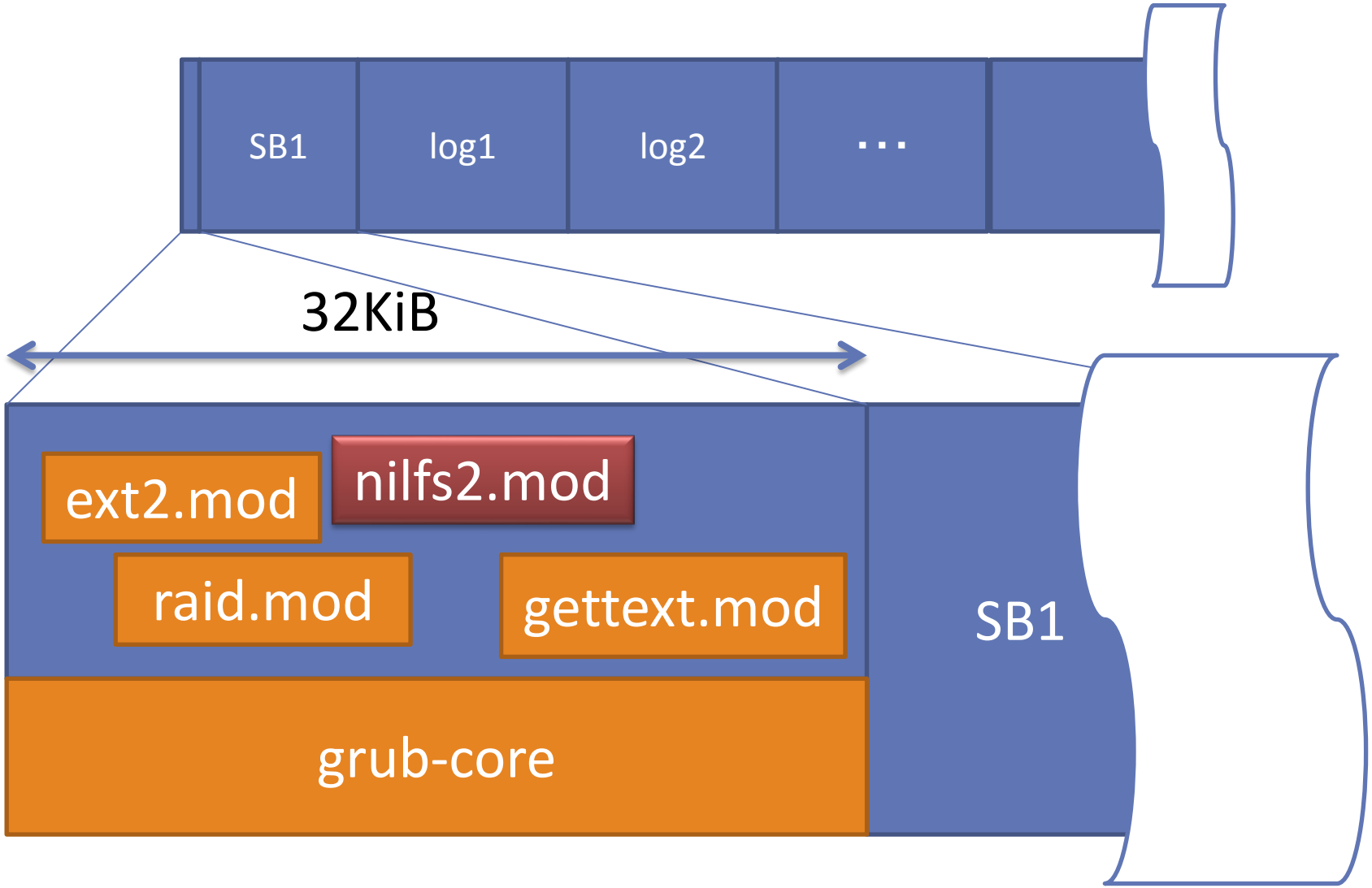
Bootloaderがkernelをメモリにロード

GRUB2 - bootloader

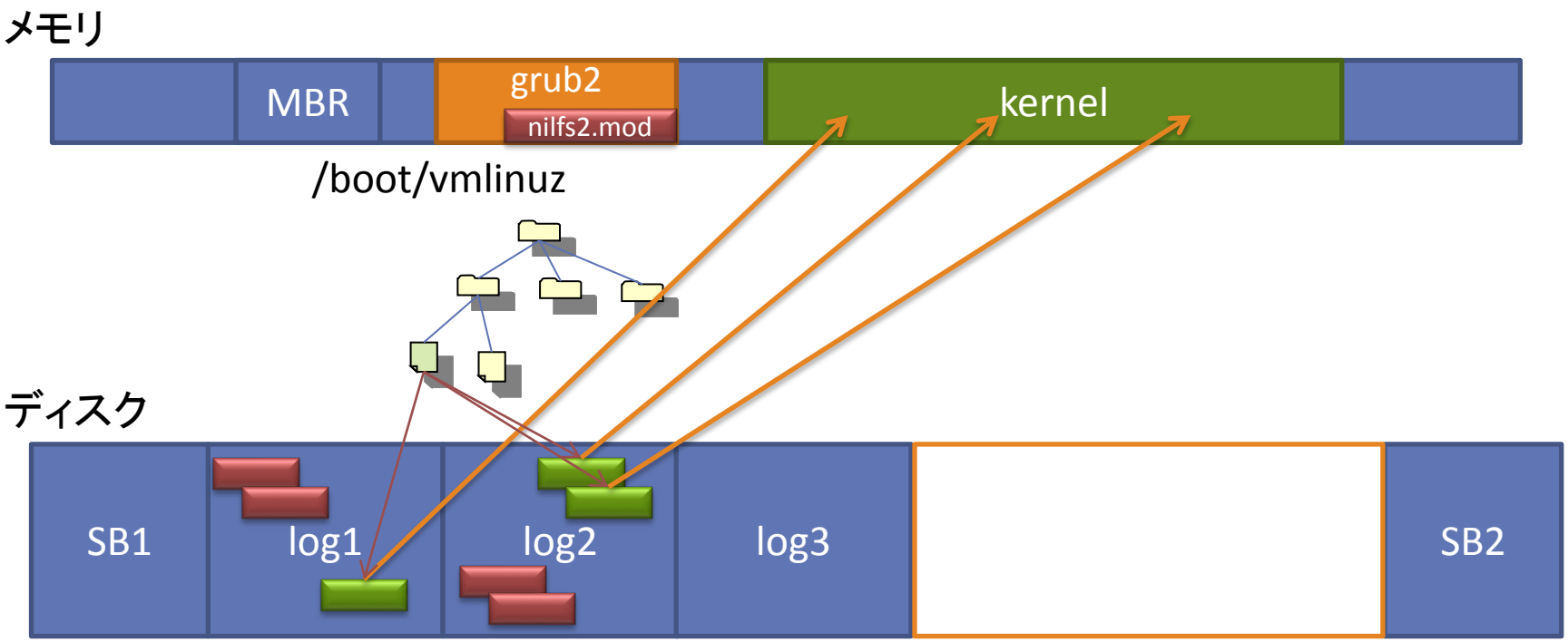


Bootloaderがkernelをメモリにロード

GRUB2 – 概要


SB1
log1
log2
...
32KiB
ext2.mod
nilfs2.mod
raid.mod
gettext.mod
grub-core
SB1

NILFS2 module



GRUB2 実装

- ファイルの読み出しまで
- SBやlogのvalidationは最小限
  - NILFS2では SB や log は crc32 を使って checksum を取っているが、現状ではそこまで厳密な validation は行っていない
- Checkpoint 番号指定の読み出しは不可能
  - 最新のファイルシステムの状態のみ

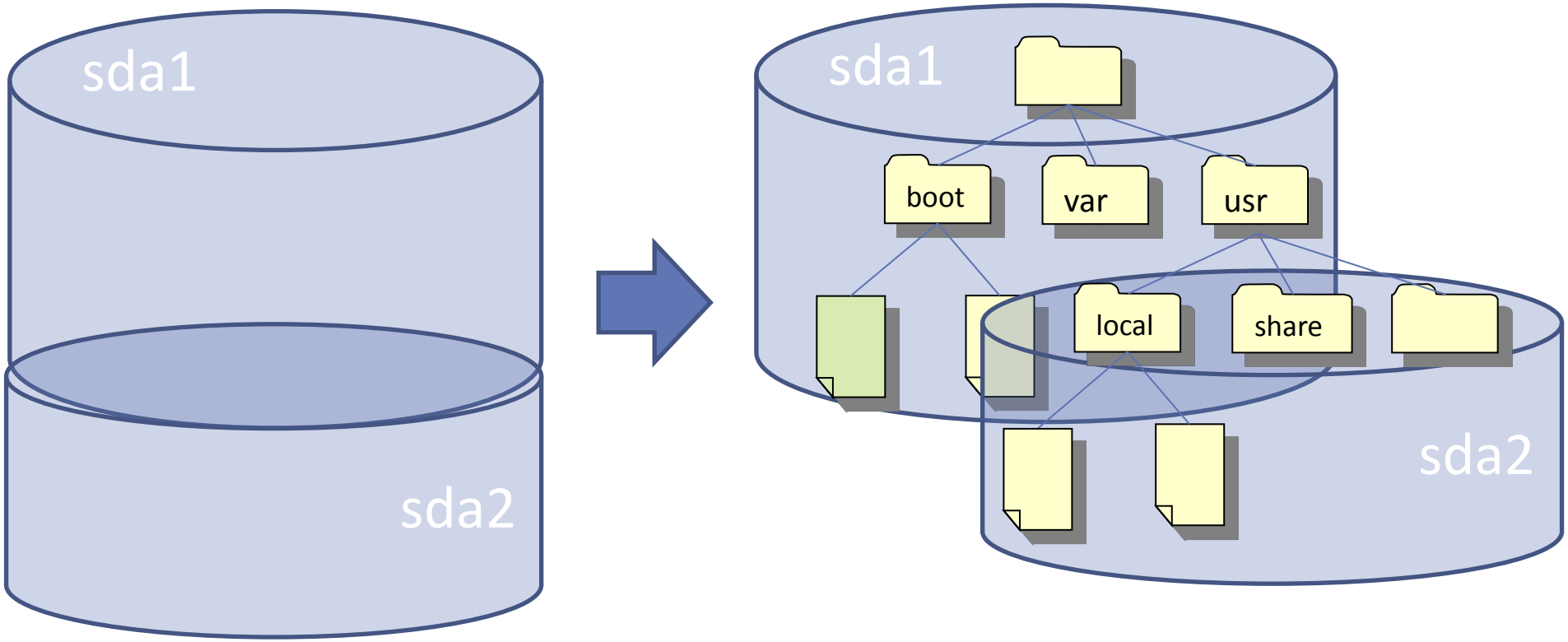
GRUB2 開発状況

- 全ての patch は accepted
- Patch が入ってからのリリースはない
  - Debian sid では既に使える

LABEL/UUID MOUNT

mount

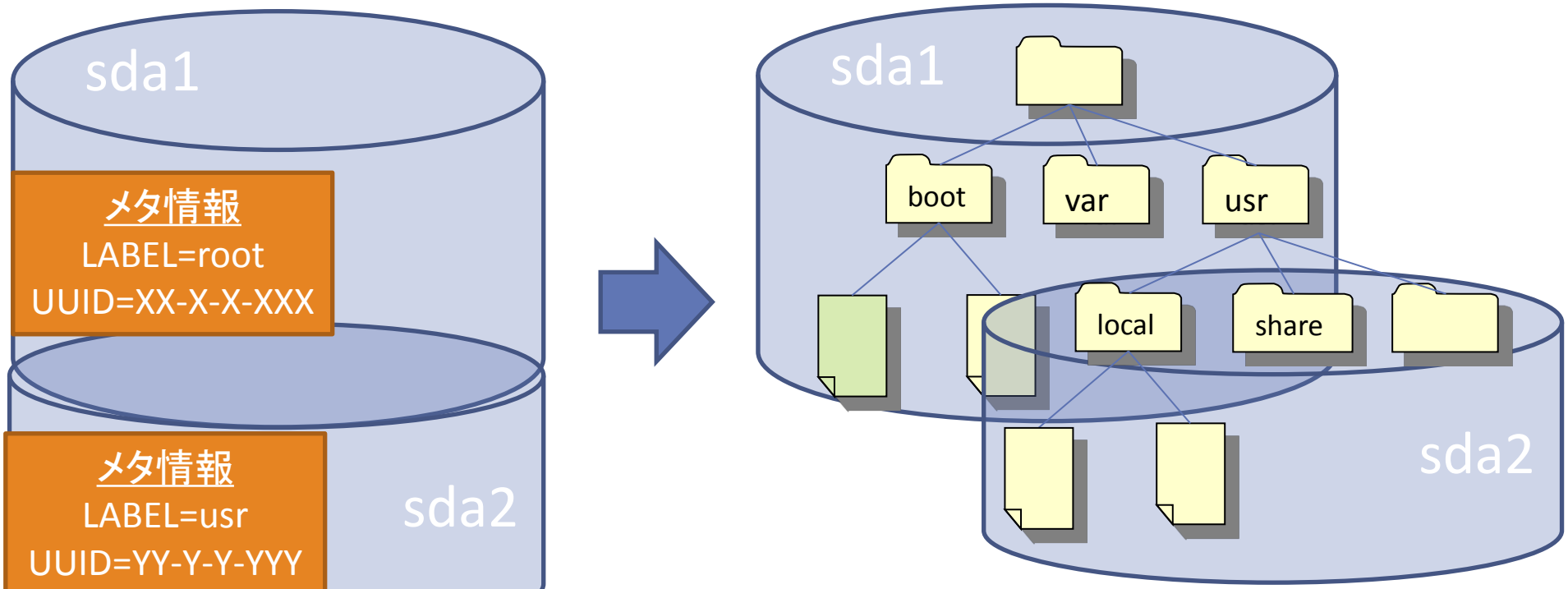
```
# mount /dev/sda1 /
# mount /dev/sda2 /usr
```



ブロックデバイスをツリーに接続

label/uuid mount

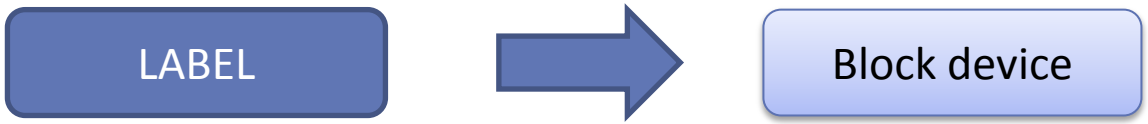
```
# mount LABEL=root /
# mount LABEL=usr /usr
```



パーティション内のファイルシステムに付いている
ラベル指定でツリーに接続

blkid

- mount(2)システムコールにはブロックデバイスが必要
  - LABEL/UUIDとブロックデバイスのマッピングが必要

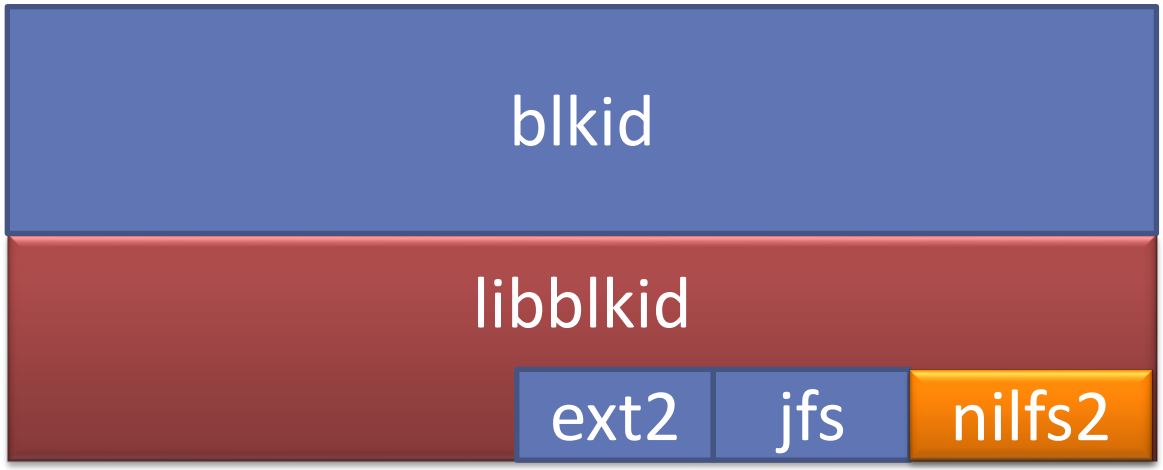


- blkid: そのマッピングを行う

```
$ /sbin/blkid
/dev/sda6: UUID="c918bf4c-74b0-4675-b4f7-9ea5bfd8f1c2" TYPE="swap"
/dev/sda1: UUID="e3e052dd-3d76-463b-a8b1-17b76fe0d898" TYPE="ext3"
/dev/sda5: UUID="b81e8cdf-9dce-40cf-b5b3-e9797f6d660b" TYPE="ext3"
/dev/sda7: UUID="24c3cd7b-8380-4b5d-9955-0deaca22d8fa" TYPE="ext3"
/dev/sdb1: LABEL="au" UUID="f8ebd010-50ba-4c29-b6e7-8ae99122082f" TYPE="ext4"
$
```

blkid – アーキテクチャ

- マッピング動作は libblkid
- libblkid はファイルシステム毎に分割されたパーティション検知モジュールが存在
- nilfs2モジュールを新たに作成



blkid – 実装

- LABEL/UUID ともに認識可能

```
$ blkid
/dev/sda1: LABEL="hoge" UUID="c2bed900-02cc-45f7-a8ab-36526a3f58b4" TYPE="ext4"
/dev/sda5: UUID="c2c7c8e6-8868-448f-b788-d7a9fffe5801" TYPE="swap"
/dev/sdb1: LABEL="xfs" UUID="0dee0fe3-daf7-47d8-a243-8fb0ac3beffd" TYPE="xfs"
/dev/sdb2: LABEL="NILFS2 is great!" UUID="b6c8260f-3f58-44ee-b417-133d5f63093f"
TYPE="nilfs2"
/dev/sdb3: UUID="4c96f277-e04f-42ca-aa46-ff7fc7b6566d" TYPE="ocfs2"
$
```

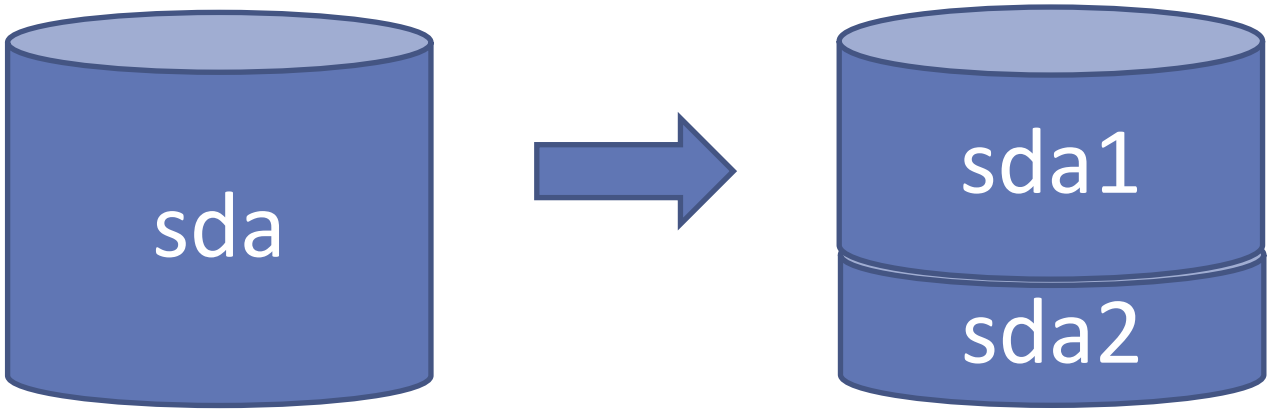
Blkid(util-linux-ng) 開発状況

- Patch は accepted
- その後のリリースはない
  - 次のリリースは不明

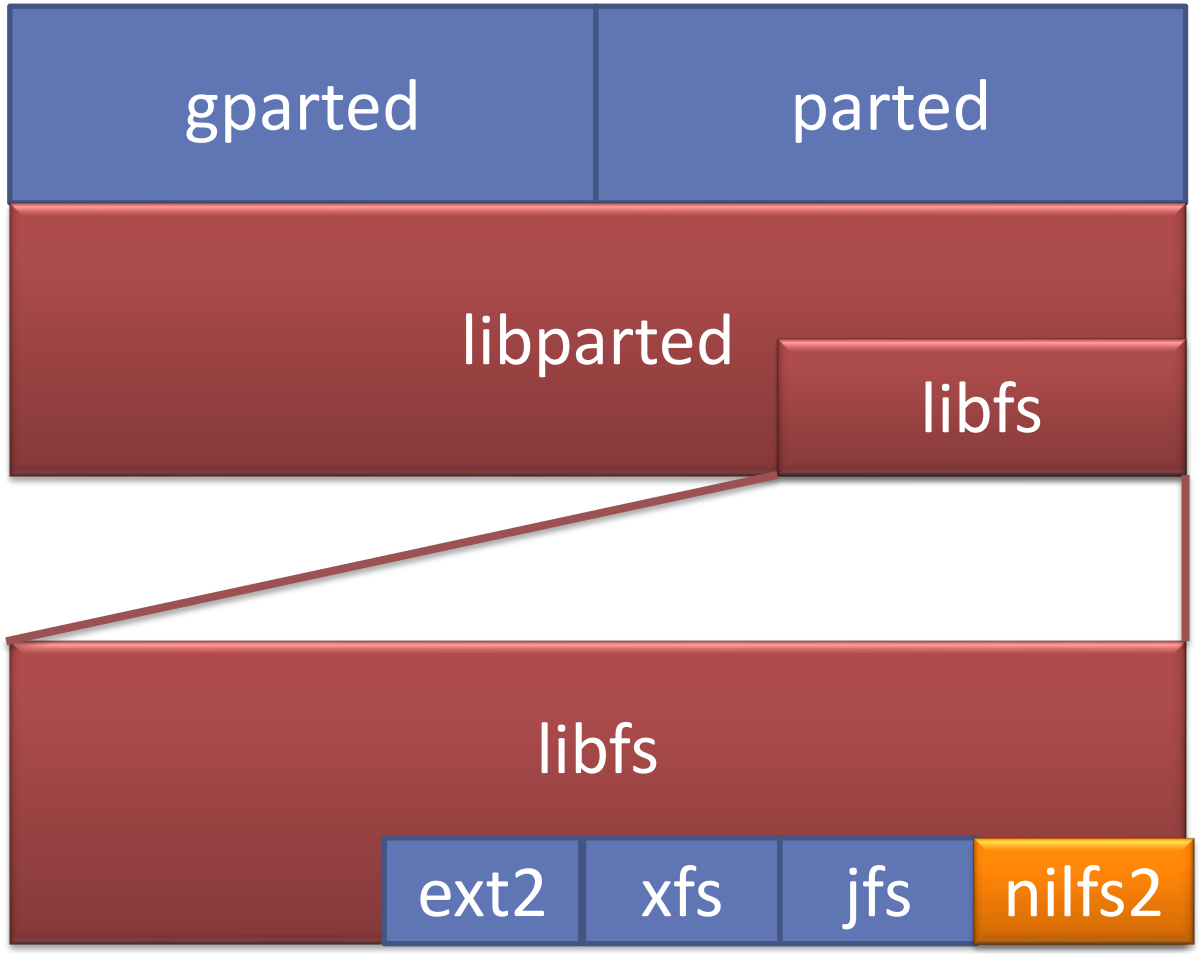
PARTITION TOOL – GNU PARTED

Partition tool

- パーティション作成
- Linux では　fdisk、cfdisk 、parted など
- Partedでは既存パーティションの拡大/縮小
  - ファイルシステムが対応していればファイルシステムの拡大/縮小



GNU Parted – アーキテクチャ



Parted – 実装

- 今回の実装
  - NILFS2 Partition 認識
  - NILFS2 Partition の破壊

```
$ sudo ./parted/parted /dev/hdb
GNU Parted 2.3.10-5896
Using /dev/hdb
Welcome to GNU Parted! Type 'help' to view a list of commands.
(parted) print
Model: QEMU HARDDISK (ide)
Disk /dev/hdb: 8590MB
Sector size (logical/physical): 512B/512B
Partition Table: msdos

Number  Start   End     Size    Type     File system  Flags
 1      32.3kB  165MB   164MB   primary  nilfs2
 2      165MB   411MB   247MB   primary  ext3

(parted)
```

Parted 開発状況

- Parted は FSFが主導している
- FSFは10行以上のpatchは譲渡契約を結ぶ必要がある
  - その契約がまだ終わらず patch を受け取ってもらえない状況

DISK UTILITY - PALIMPSEST

Disk utility – palimpsest

8.6 GB ハードディスク (ATA VBOX HARDDISK) [/dev/sda] — ディスク・ユーティリティ

ファイル(F) ヘルプ(H)

ストレージ・デバイス(S)

- ローカルストレージ
  jir@localhost
  - PATA ホストアダプタ
    82371AB/EB/MB PIIX4 IDE
    - 8.6 GB ハードディスク
      ATA VBOX HARDDISK
    - 2.1 GB ハードディスク
      ATA VBOX HARDDISK
  - CD/DVD ドライブ
    VBOX CD-ROM
- 周辺機器
  USB、FireWire、他の周辺機器
- フロッピー ドライブ
  Floppy Drive

ドライブ

モデル名:	ATA VBOX HARDDISK	シリアルナンバー:	VB328060dd-42936b7c
ファームウェアバージョン:	1.0	World Wide Name:	-
位置:	PATA ホストアダプタ の ポート 1	デバイス:	/dev/sda
書き込みキャッシュ:	有効	回転速度:	-
容量:	8.6 GB (8,589,934,592 バイト)	接続:	ATA
パーティション:	マスターブートレコード	SMART 状態:	サポートされていません

ドライブをフォーマット(D)
ドライブの消去またはパーティション分け

ベンチマーク(B)
ドライブのパフォーマンスを計測

ボリューム(V)

hoge
8.2 GB ext4

拡張領域
417 MB

スワップ領域 417 MB
417 MB

利用方法:	ファイルシステム	デバイス:	/dev/sda1
パーティション・タイプ:	Linux (0x83)	パーティション・ラベル:	-
パーティション・フラグ:	ブート可能	容量:	8.2 GB (8,169,455,616 バイト)
タイプ:	Ext4 (バージョン 1.0)	利用可能:	-
ラベル:	hoge	マウントポイント:	/ でマウント

ボリュームをアンマウント(M)
ボリュームをアンマウント

ボリュームをフォーマット(R)
ボリュームの消去またはフォーマット

ファイルシステムをチェック(C)
ファイルシステムのチェックと修復

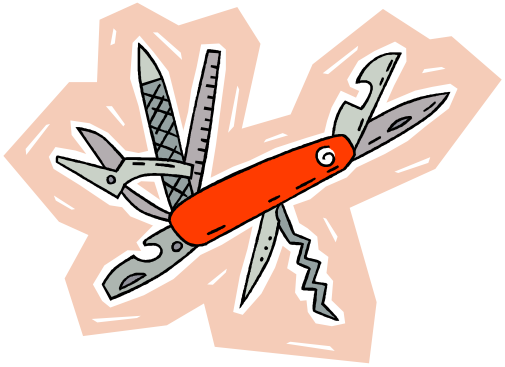
ファイルシステムのラベルを編集(L)
ファイルシステムのラベルを変更

パーティションを編集(I)
パーティション・タイプやラベルやフラグを変更

パーティションを削除(E)
パーティションを削除

Palimpsest – 機能

- ファイルシステム認識
- パーティションニング
- mount/umount
- フォーマット
- ラベル変更
- fsck



UNICUS

Palimpsest - architecture

dbus client process

palimpsest

D-bus

udisks

libparted libblkid

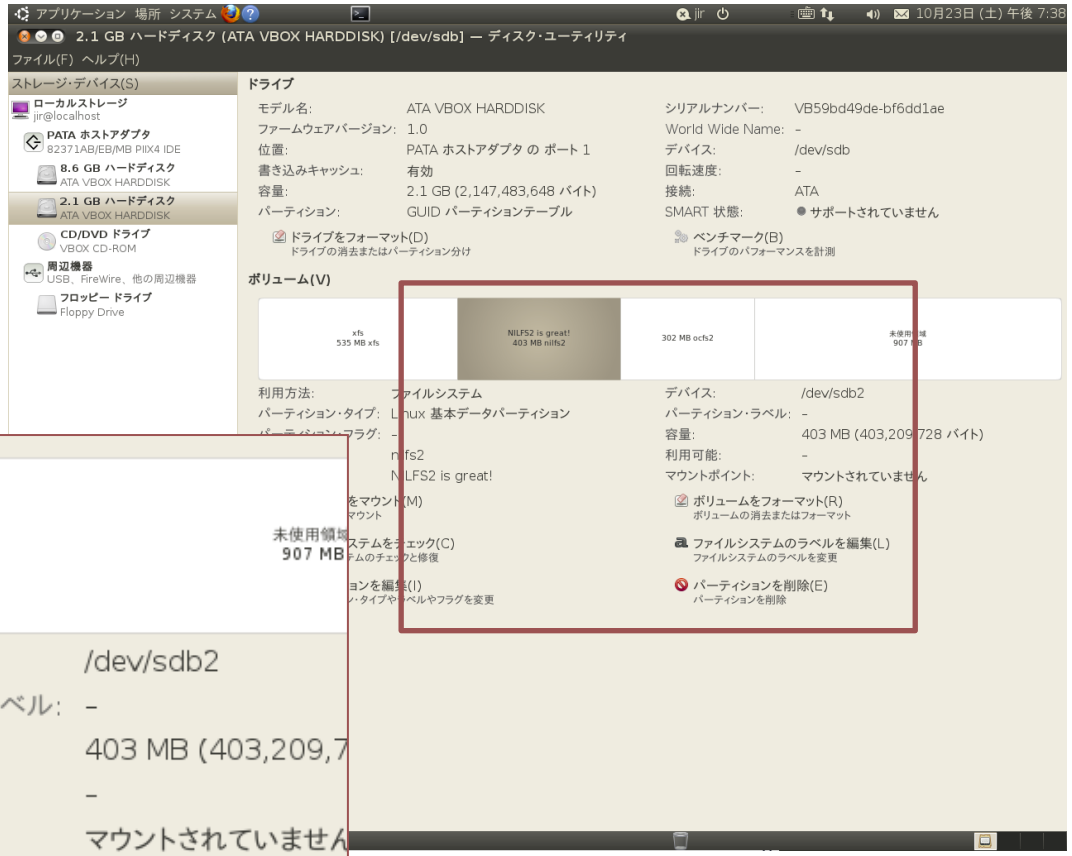
dbus server process

Udisks – 実装

- 今回実装したもの
  - ファイルシステムの検知
  - ファイルシステムの作成
  - ファイルシステムラベルの変更

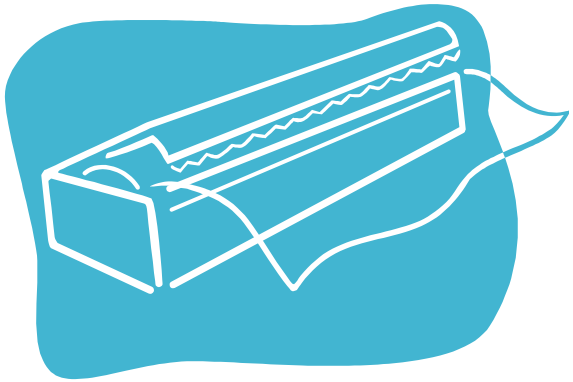
UNICUS

Palimpsest – 実装結果



Udisks 開発状況

- 全てのpatchは accepted
- Checkpoint指定のmountなどは出来ない
  - Udisks に Interface が存在しない



WRAP UP

まとめ

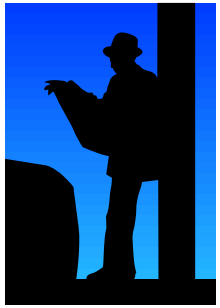
- ファイルシステムのサポートとは？
  - Kernel モジュールの数
  - Disk utility でのサポート数
- NILFS2 の紹介
- NILFS2 対応したユーティリティ
  - GRUB2
  - blkid
  - Parted
  - Palimpsest

今後の課題

- Parted の対応
  - NILFS2 のフォーマット
  - パーティションの拡張
- NILFS2独自の機能の追加
  - checkpoint を指定した kernel ロード
  - Unclean な状態からの log の roll forwarding

最後に

- NILFS2 git tree は日本発のプロジェクトで唯一 linus が直接 pull する git tree
- linux-nilfs2 ML の活動は kernel community としてはとても敷居が低い
  - ユーザレベルの質問もあります
- 最新の技術的な質問も日本語で可能
  - ML等はさすがに英語ですが・・・
- NILFS2 はあなたの参加を**お待ち**しております



UNICUS