VFS and Block I/O

VA Linux Systems Japan

Hirotaka.Sasaki

sasaki@valinux.co.jp

VA LINUX SYSTEMS JAPAN

目次

- ファイルシステム概要
- VFS（仮想ファイルシステム）
- ページキャッシュとバッファキャッシュ
- ブロックI/O共通レイヤ

ファイルシステム概要

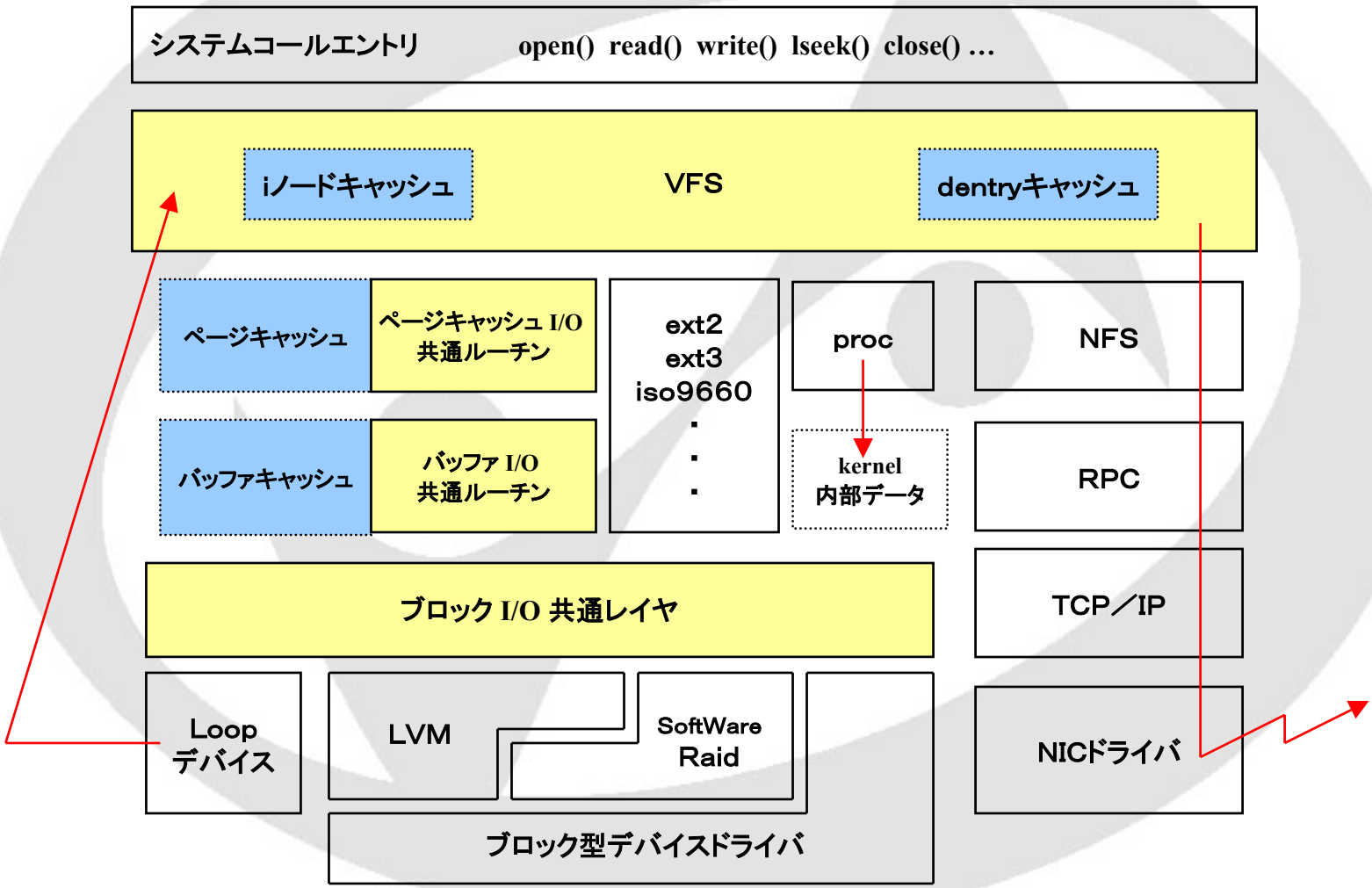
VA LINUX SYSTEMS JAPAN

ファイルシステムとは？

記憶デバイス上のデータをファイルという論理的な構造で抽象化し利用するための機構

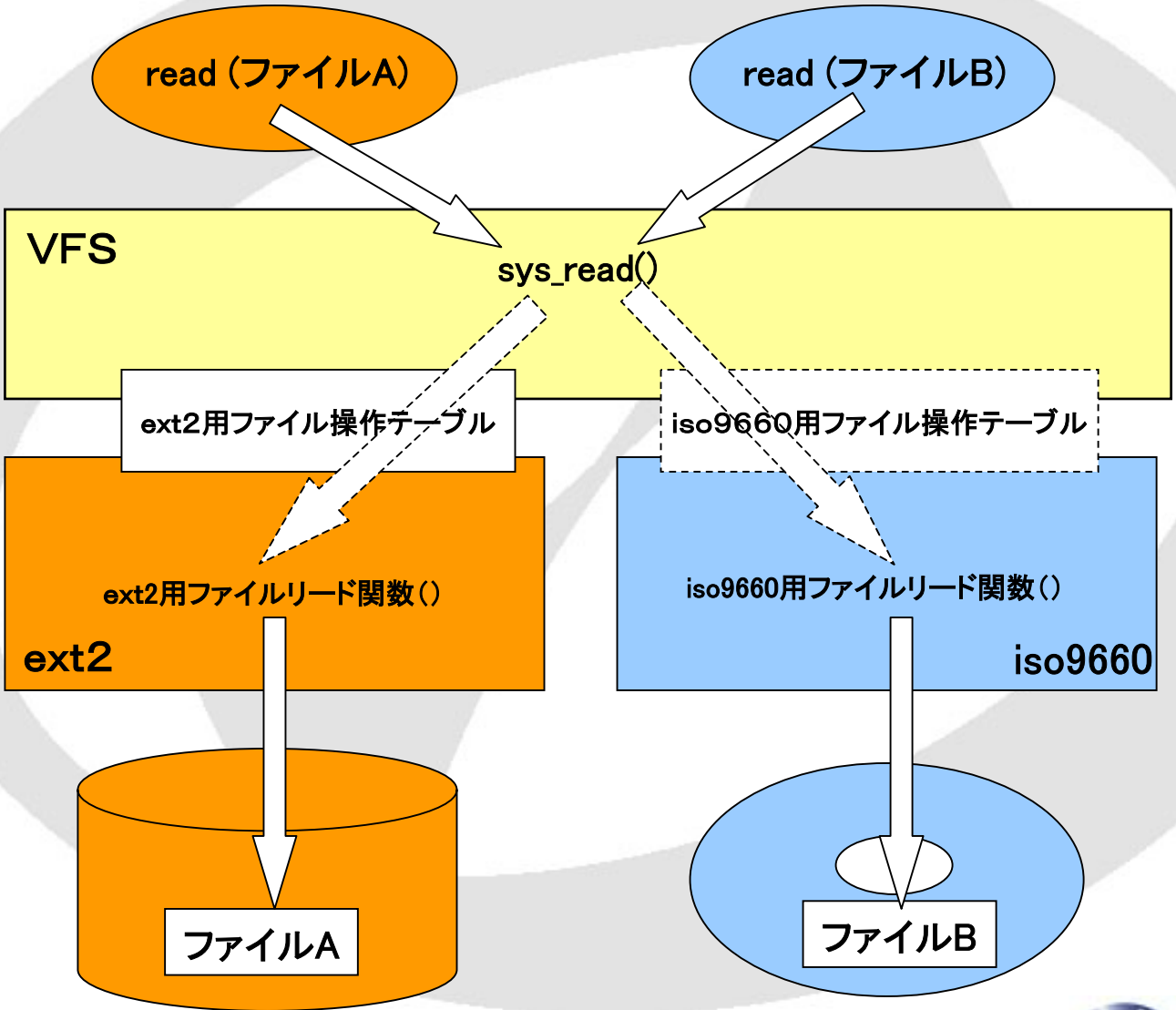
- 論理的なファイルと物理的なディスクブロックの対応を管理
- ファイルに対するアクセス要求を、ディスクブロックへのアクセス要求へ変換
- ファイルの生成、拡張、削除に応じた、ディスクブロックの割り当て

ファイルシステムのレイヤ構造



VA LINUX SYSTEMS JAPAN

VFSの概要



VA LINUX SYSTEMS JAPAN

さまざまなファイルシステム

- ローカルファイルシステム
  - ローカルディスク上のファイルシステム
  - Linuxの標準はext2ファイルシステム
  - 他、FFS（BSD）、VFAT（Win98）、iso9660など多数
  - ext3、XFSなどのジャーナリングFSもサポート
- ネットワークファイルシステム
  - ネットワークの先にあるファイルシステムを共有
  - 代表的（Unix系）なものはNFS
  - 他、CODA、SMB、NCP、Intermezzoなど
- 特殊なファイルシステム
  - procファイルシステム
  - kernel内部データをファイルとして抽象化

ブロックI／O共通レイヤ

- ディスクブロックのアクセス要求をもとにブロック型デバイスドライバを起動
- ファイルシステム構造には関与しない
- バッファI／O要求のクラスタリング

さまざまなキャッシュ

キャッシュによる処理の高速化

- ページキャッシュ
  （ファイルデータのキャッシュ）
- バッファキャッシュ
  （ディスクブロックのキャッシュ）
- iノードキャッシュ
- dentryキャッシュ

VFS

VA LINUX SYSTEMS JAPAN

VFSのデータオブジェクト

- file_system_typeオブジェクト
- super_blockオブジェクト
- inodeオブジェクト
- dentryオブジェクト
- fileオブジェクト
- address_spaceオブジェクト

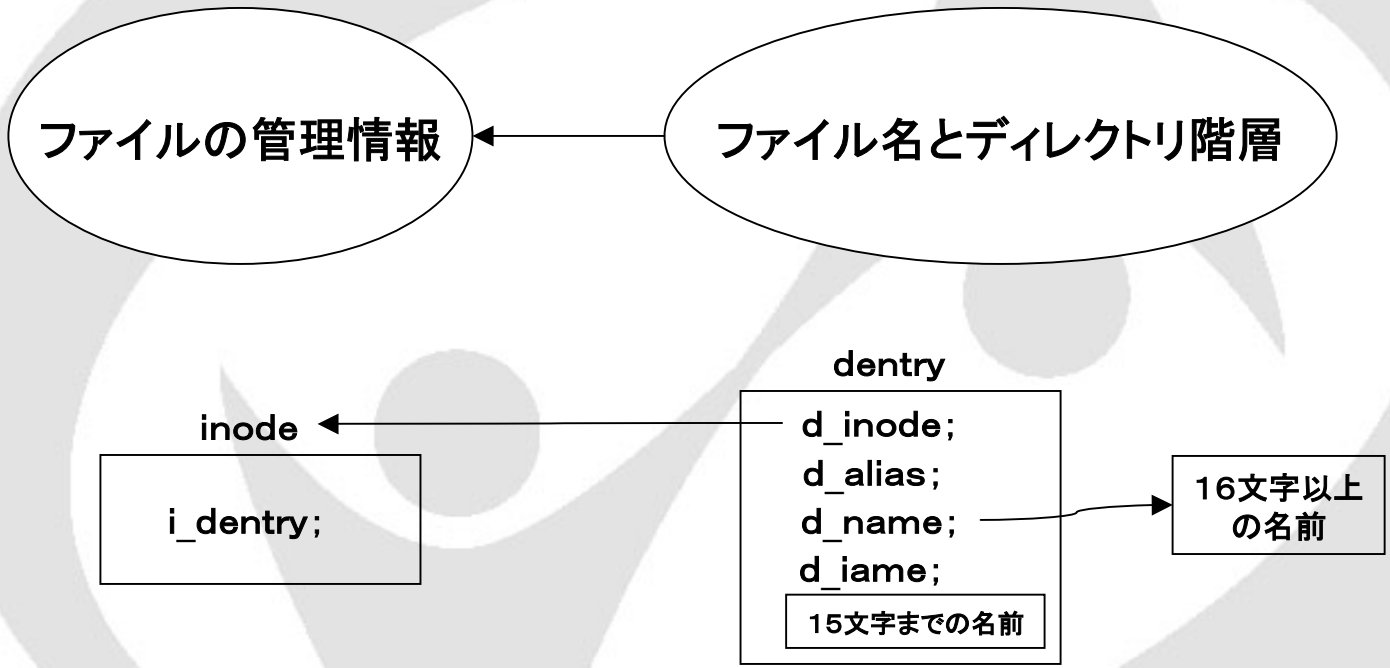
iノードオブジェクト

```
struct inode {
  iノード番号;
  利用カウンタ;
  デバイス番号;
  ハードリンク数;
      :
  union {
      minix固有のiノード情報;
      ext2固有のiノード情報;
          :
  }
}
```

- minix固有のiノード情報
- ext2固有のiノード情報
- ext3固有のiノード情報
- nfs固有のiノード情報
- :

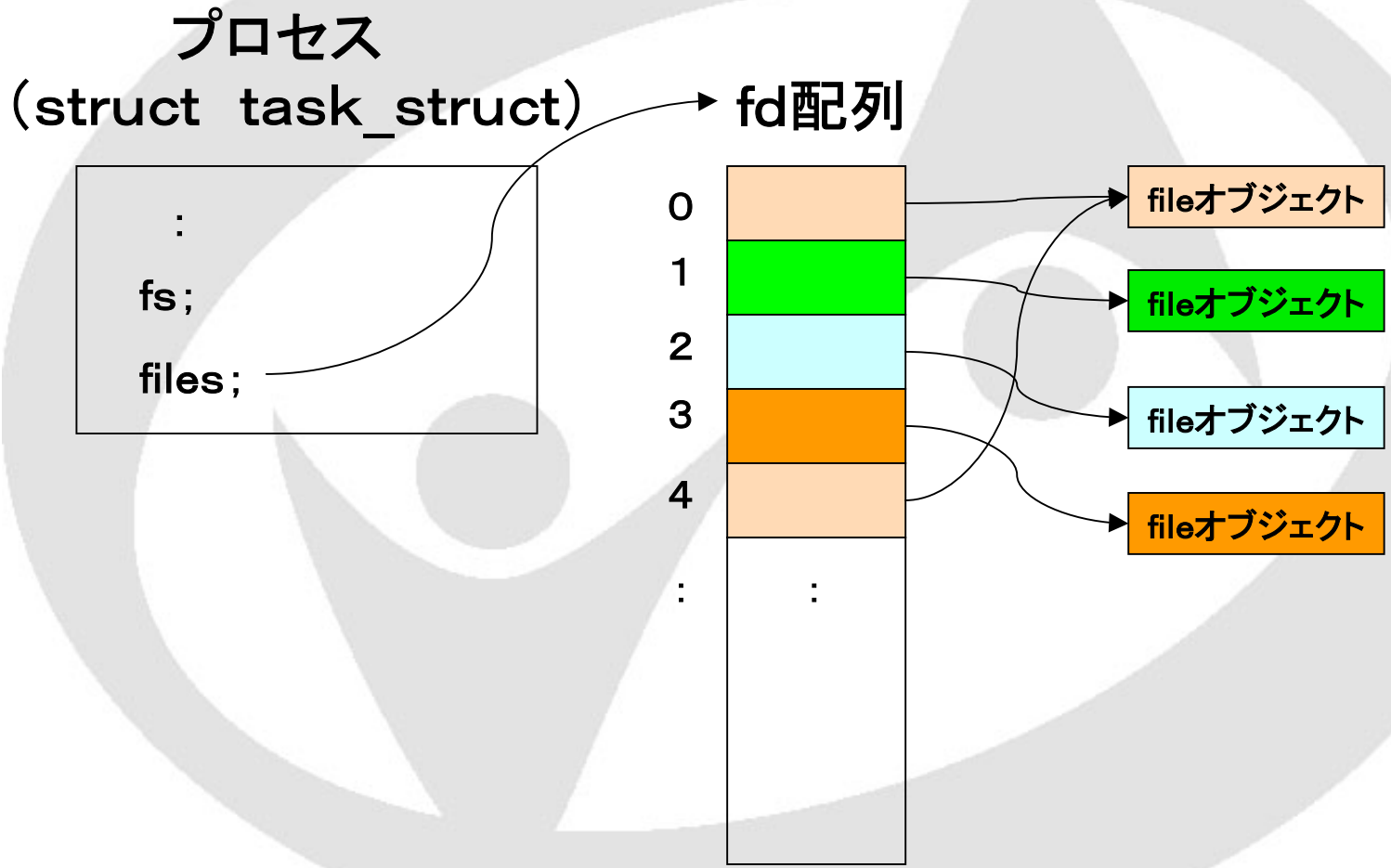
VA LINUX SYSTEMS JAPAN

iノードオブジェクトとdentryオブジェクト



VA LINUX SYSTEMS JAPAN

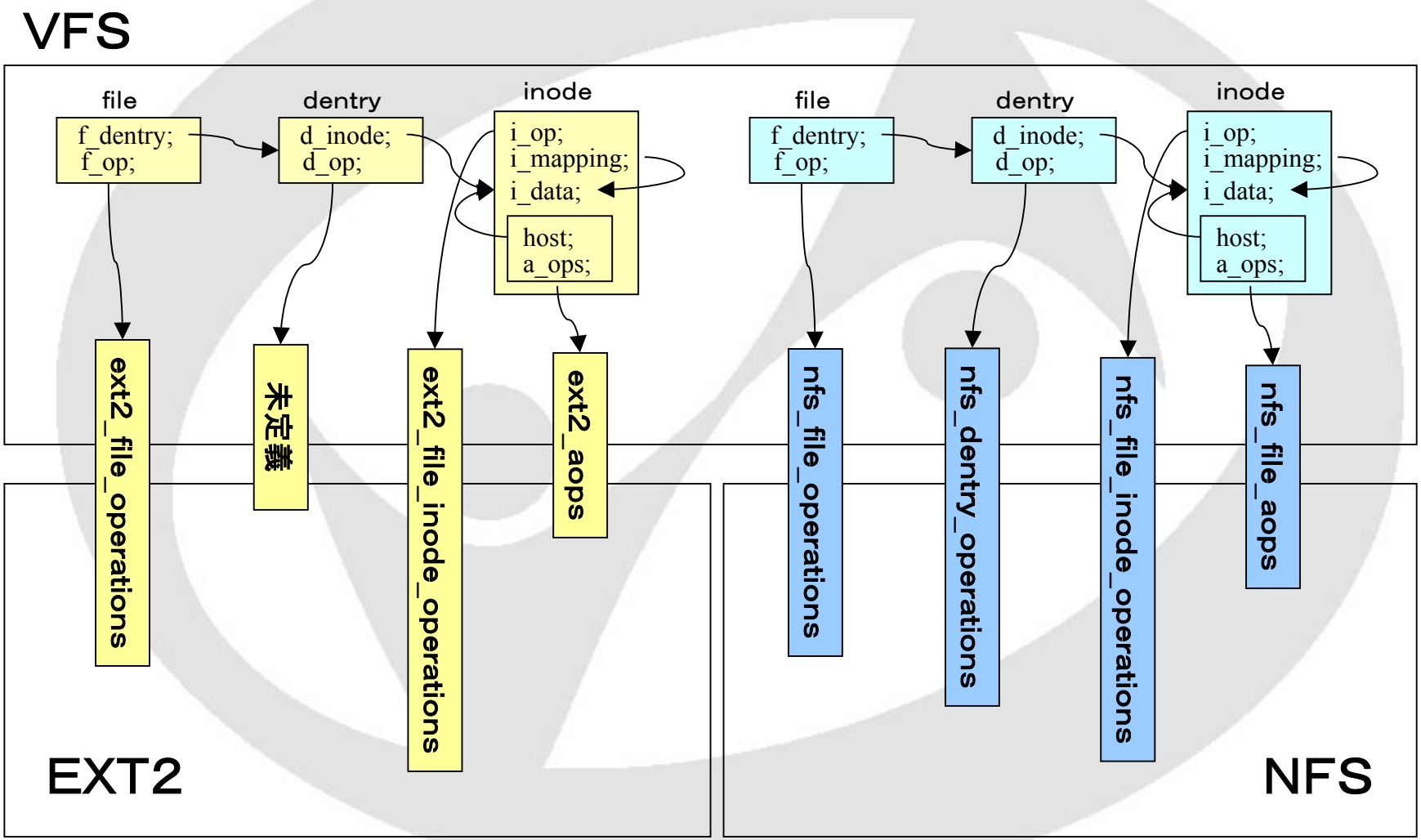
プロセスとファイル



VFSとファイルシステムのI／F

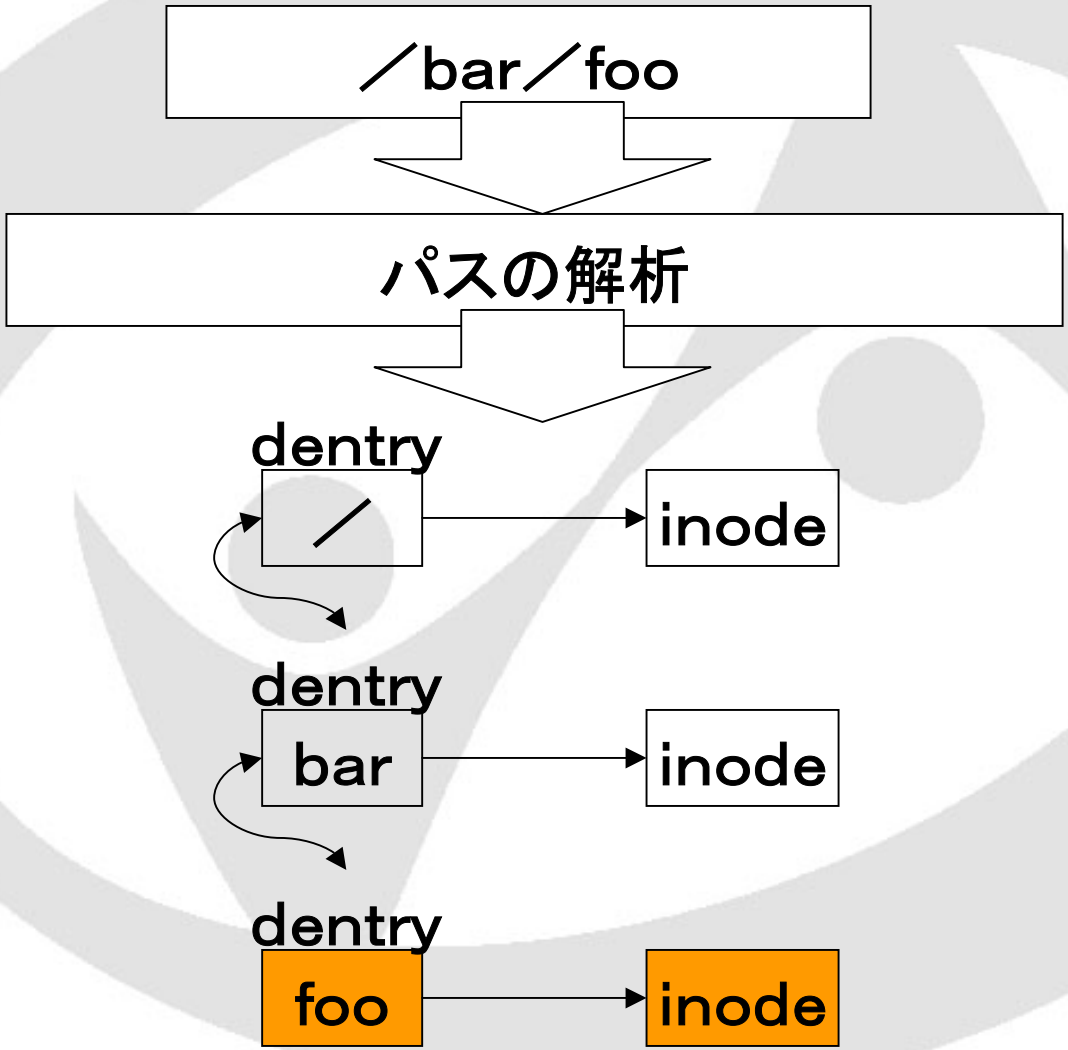
- super_block読み込みメソッド
  （file_system_typeオブジェクトのread_superメンバ）
- super_operationsテーブル
  （super_blockオブジェクトのs_opメンバ）
- inode_operationsテーブル
  （inodeオブジェクトのi_opメンバ）
- dentry_operationsテーブル
  （dentryオブジェクトのd_opメンバ）
- file_operationsテーブル
  （fileオブジェクトのf_opメンバ）
- address_space_operationsテーブル
  （address_spaceオブジェクトのa_opsメンバ）

VFSとファイルシステムのI／F（通常ファイル）

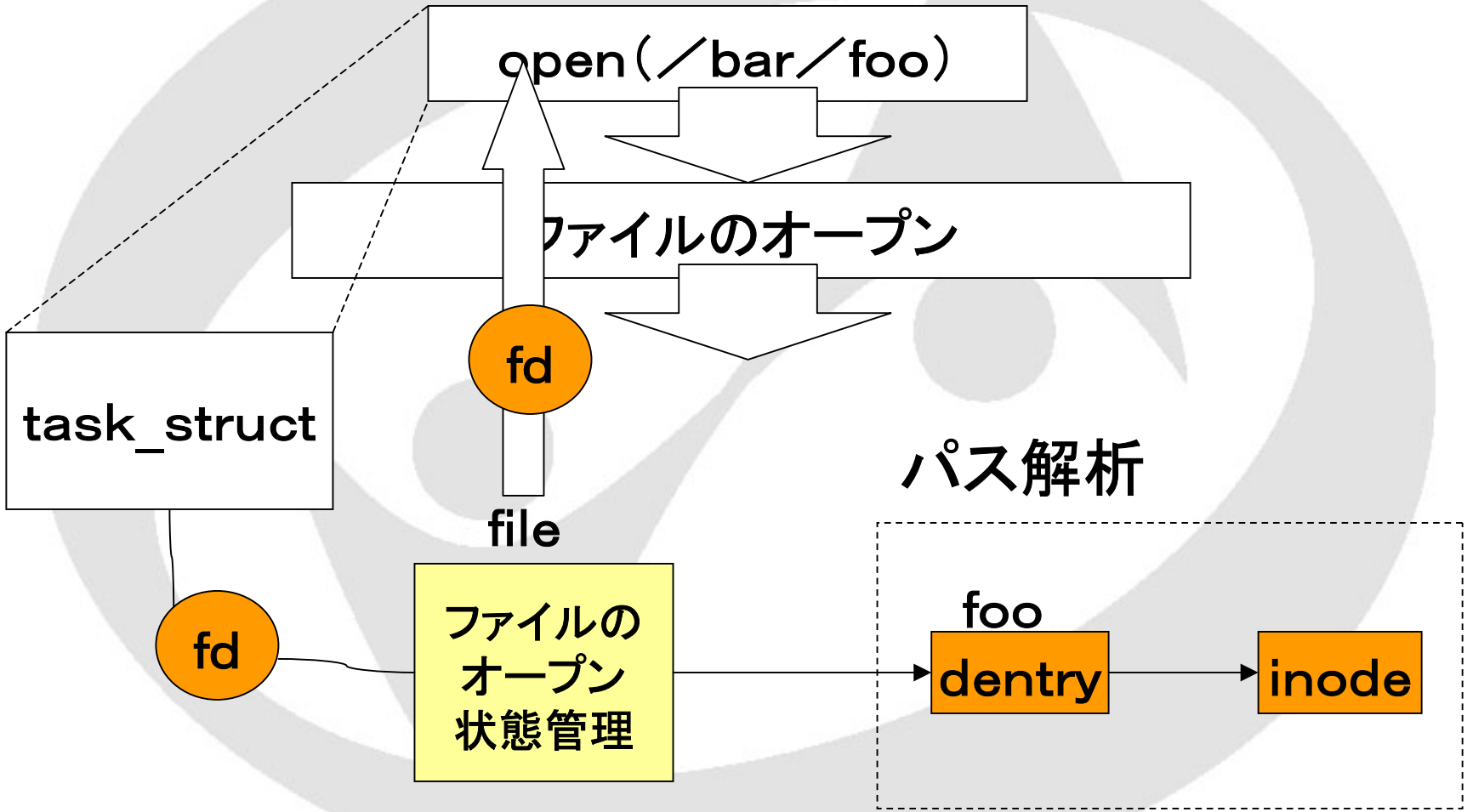


VA LINUX SYSTEMS JAPAN

パスの解析

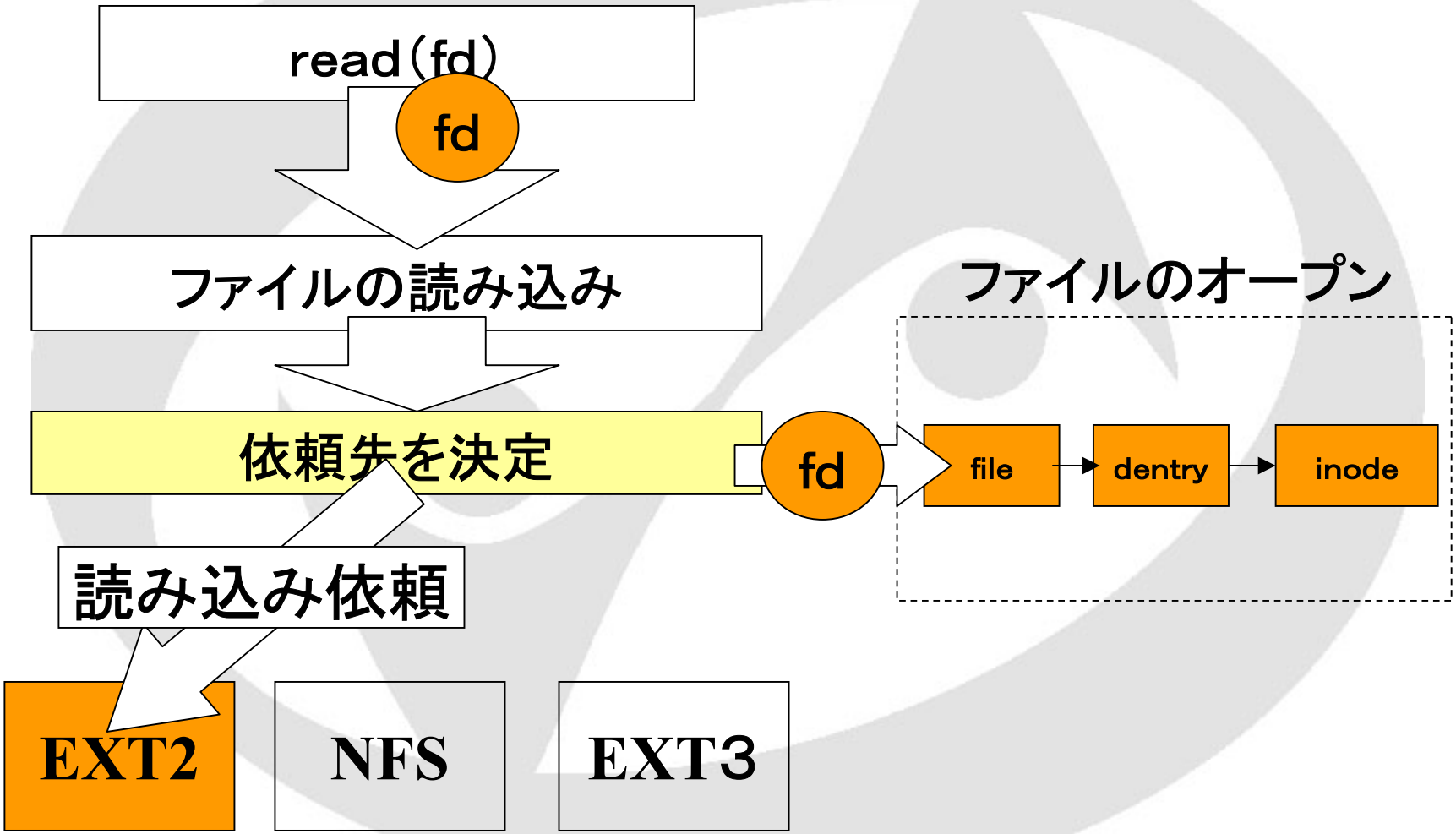


ファイルのオープン



VA LINUX SYSTEMS JAPAN

ファイルの読み込み

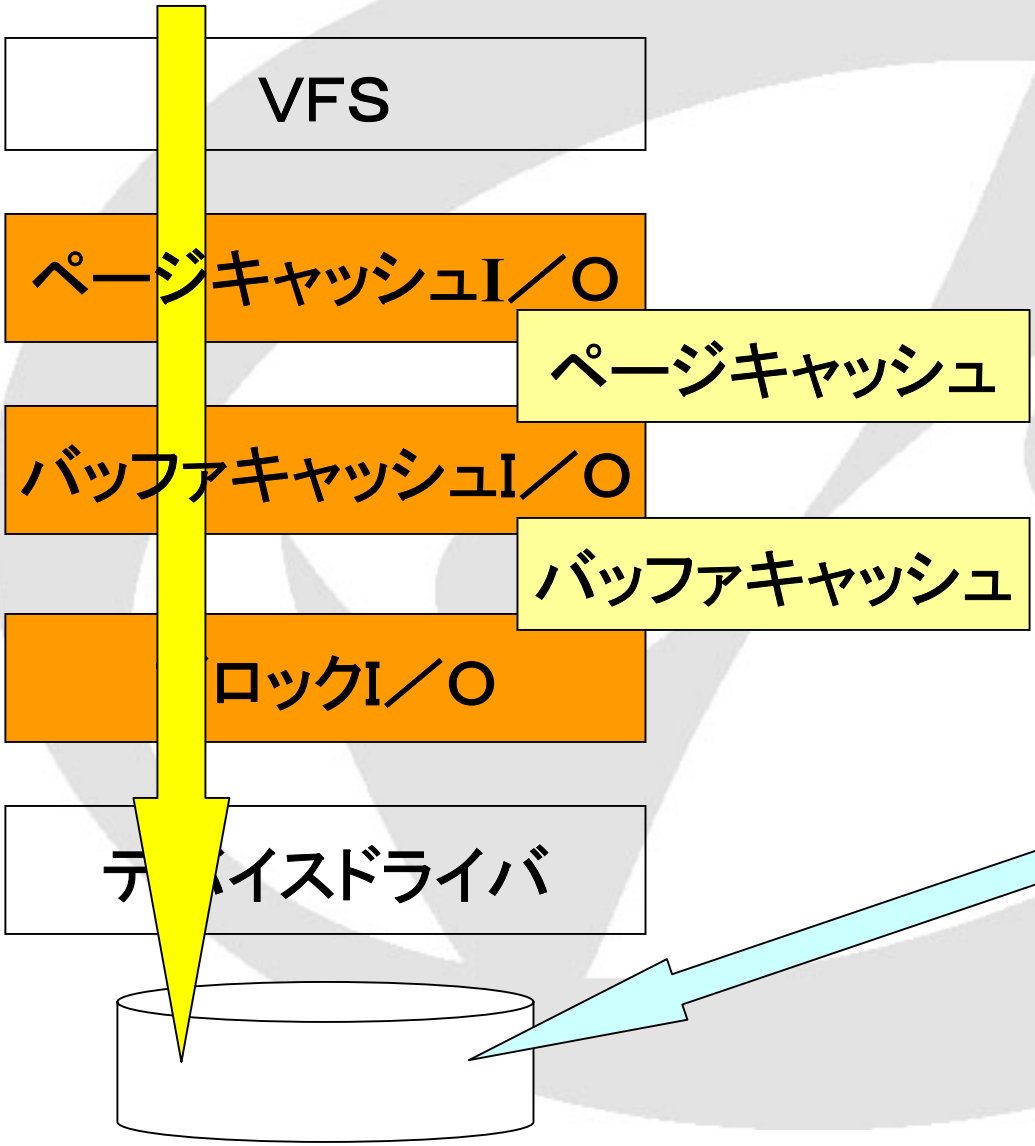


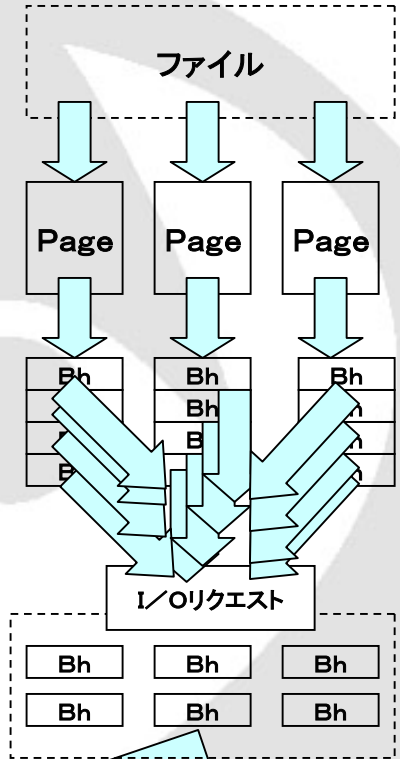
VA LINUX SYSTEMS JAPAN

ページキャッシュ
と
バッファキャッシュ

VA LINUX SYSTEMS JAPAN

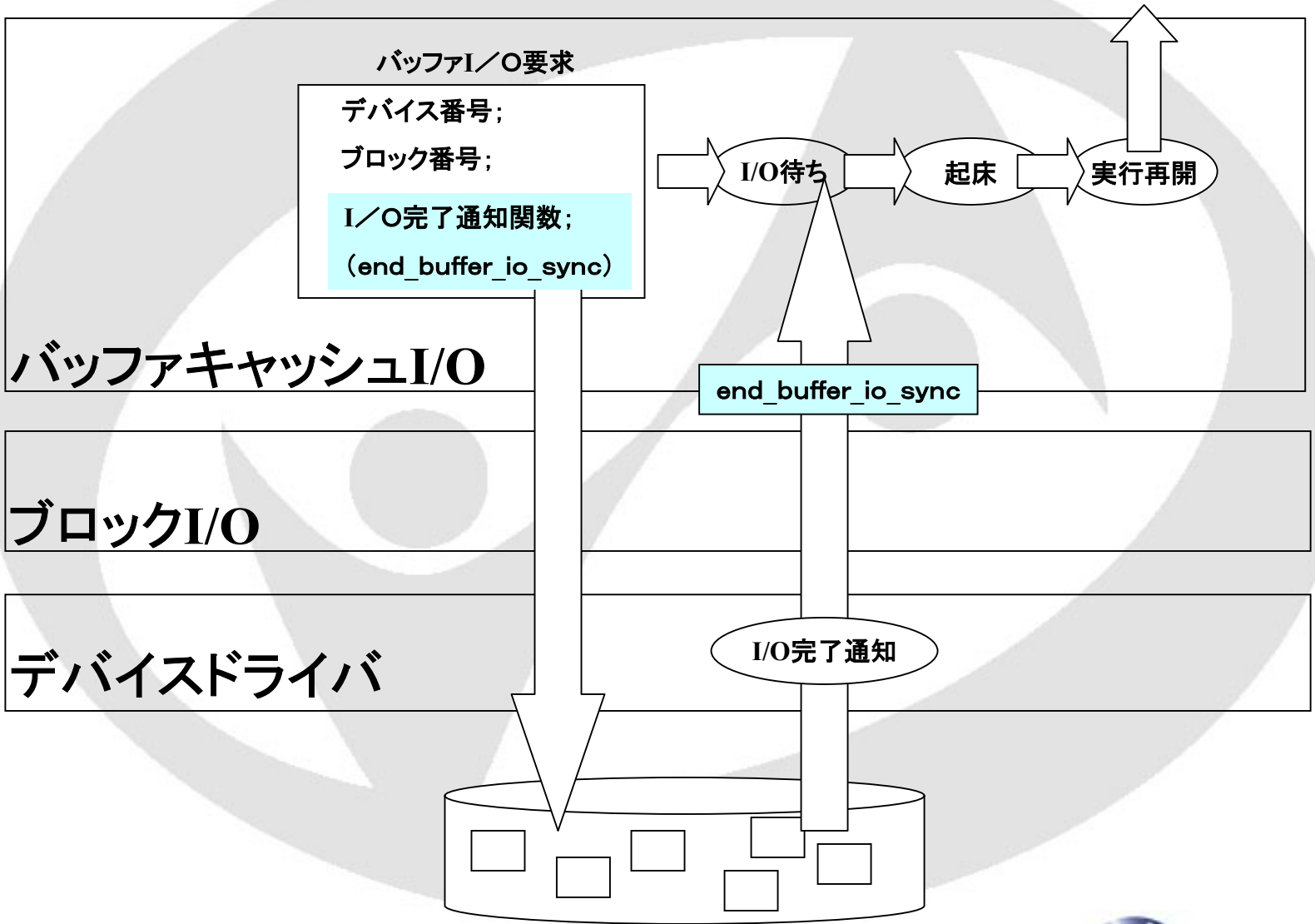
ファイルI／Oの処理階層





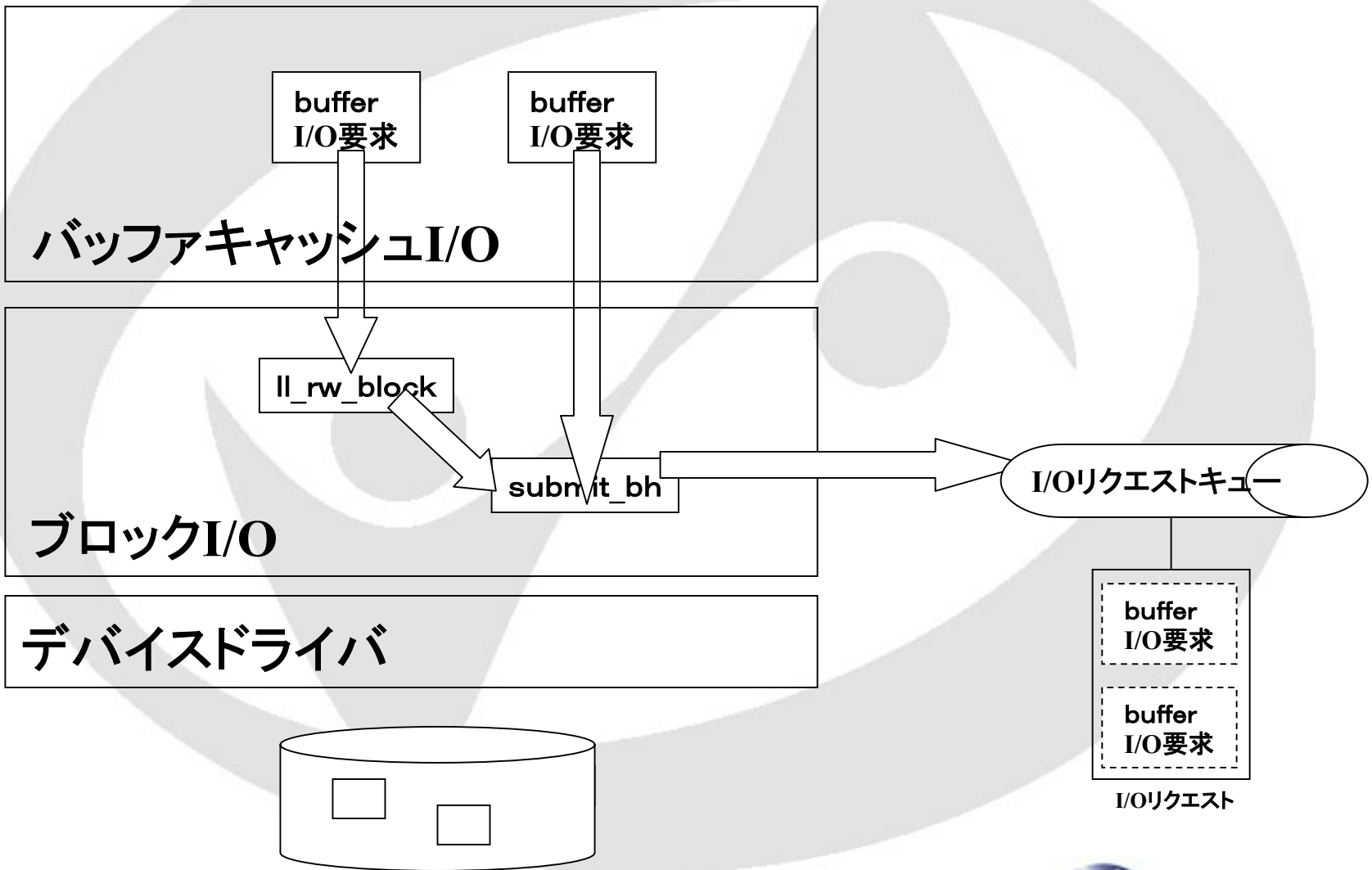
VA LINUX SYSTEMS JAPAN

バッファキャッシュI／Oの概要



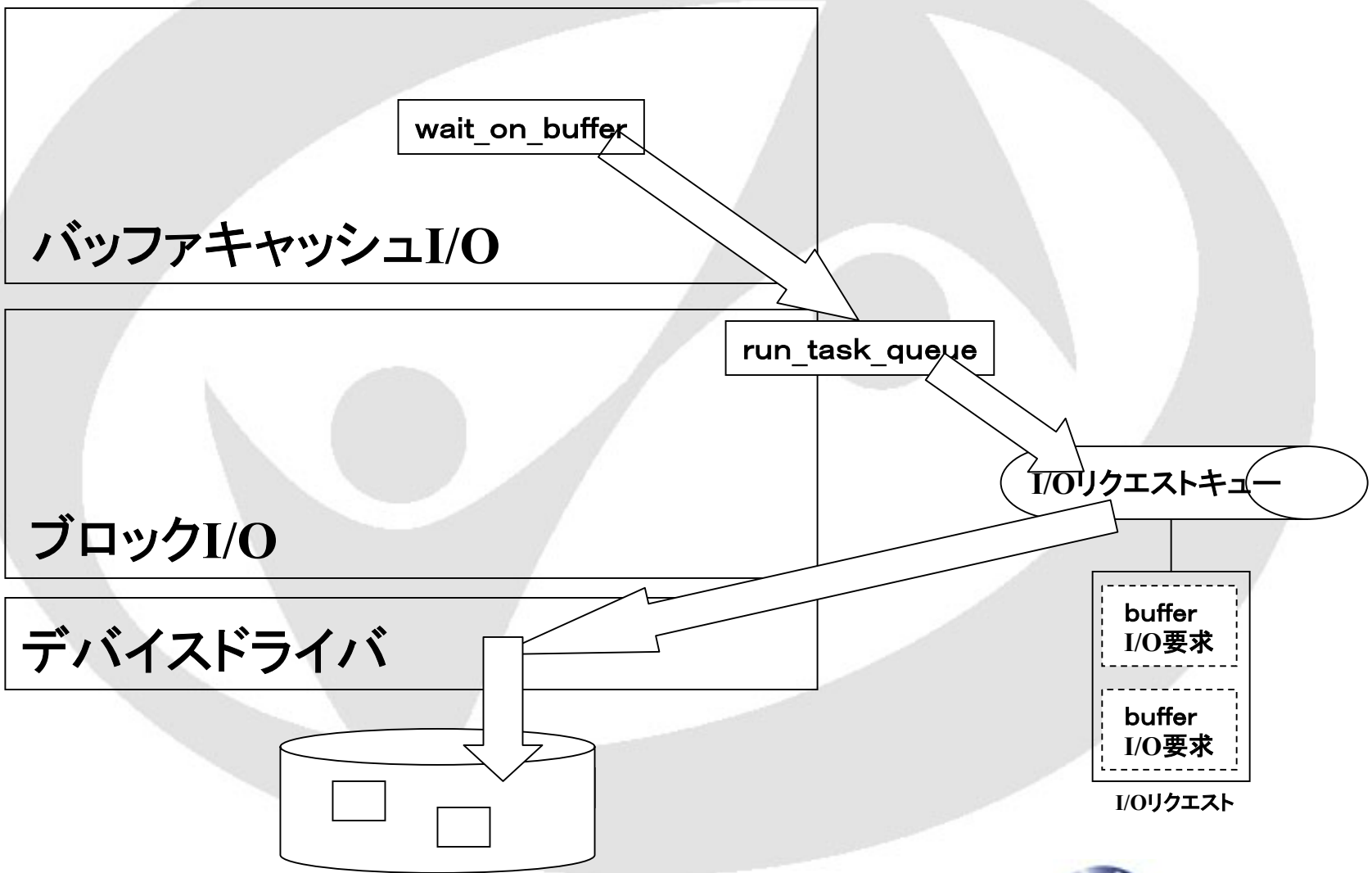
VA LINUX SYSTEMS JAPAN

バッファのI／O要求



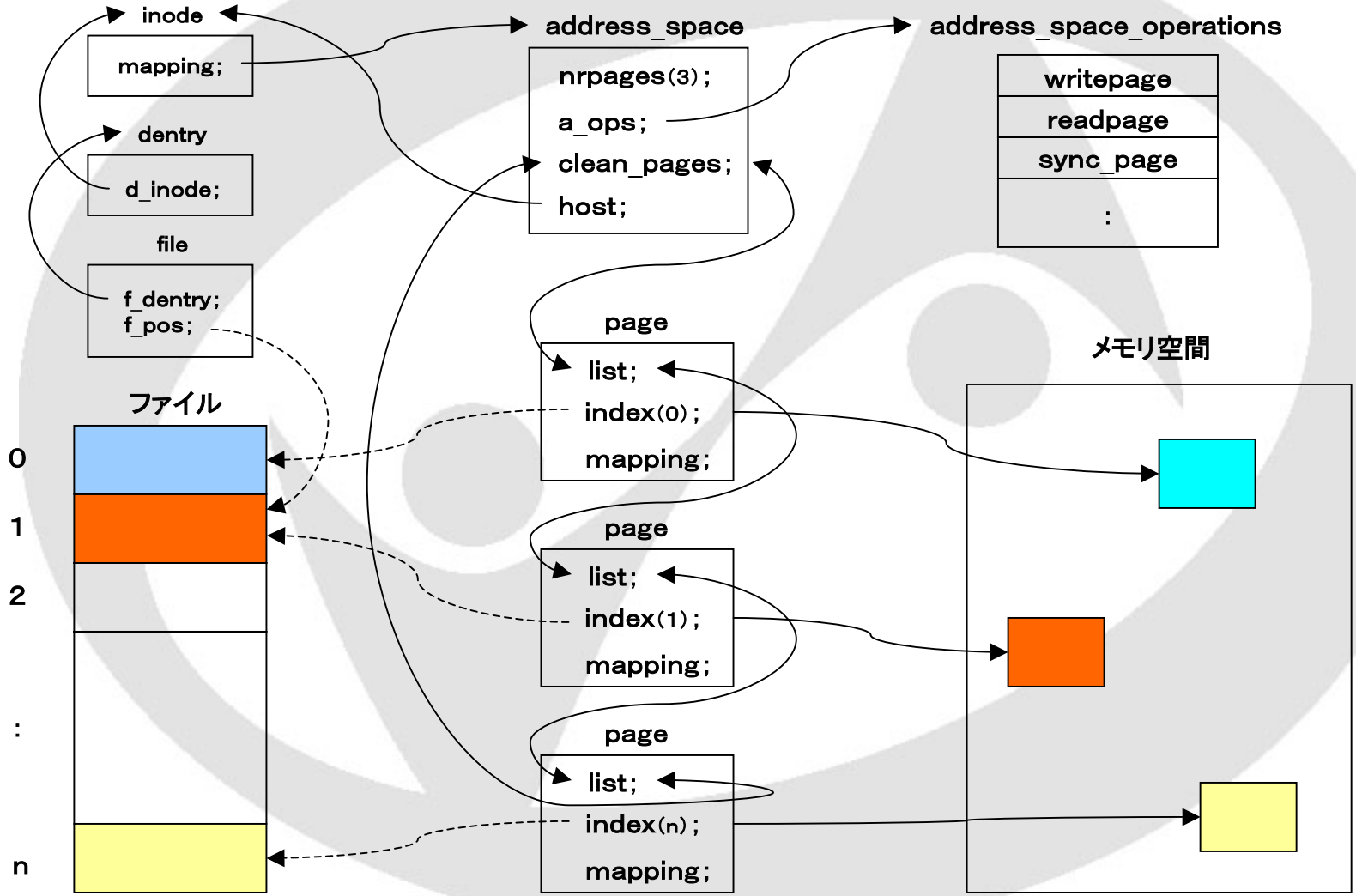
VA LINUX SYSTEMS JAPAN

バッファのI／O実行



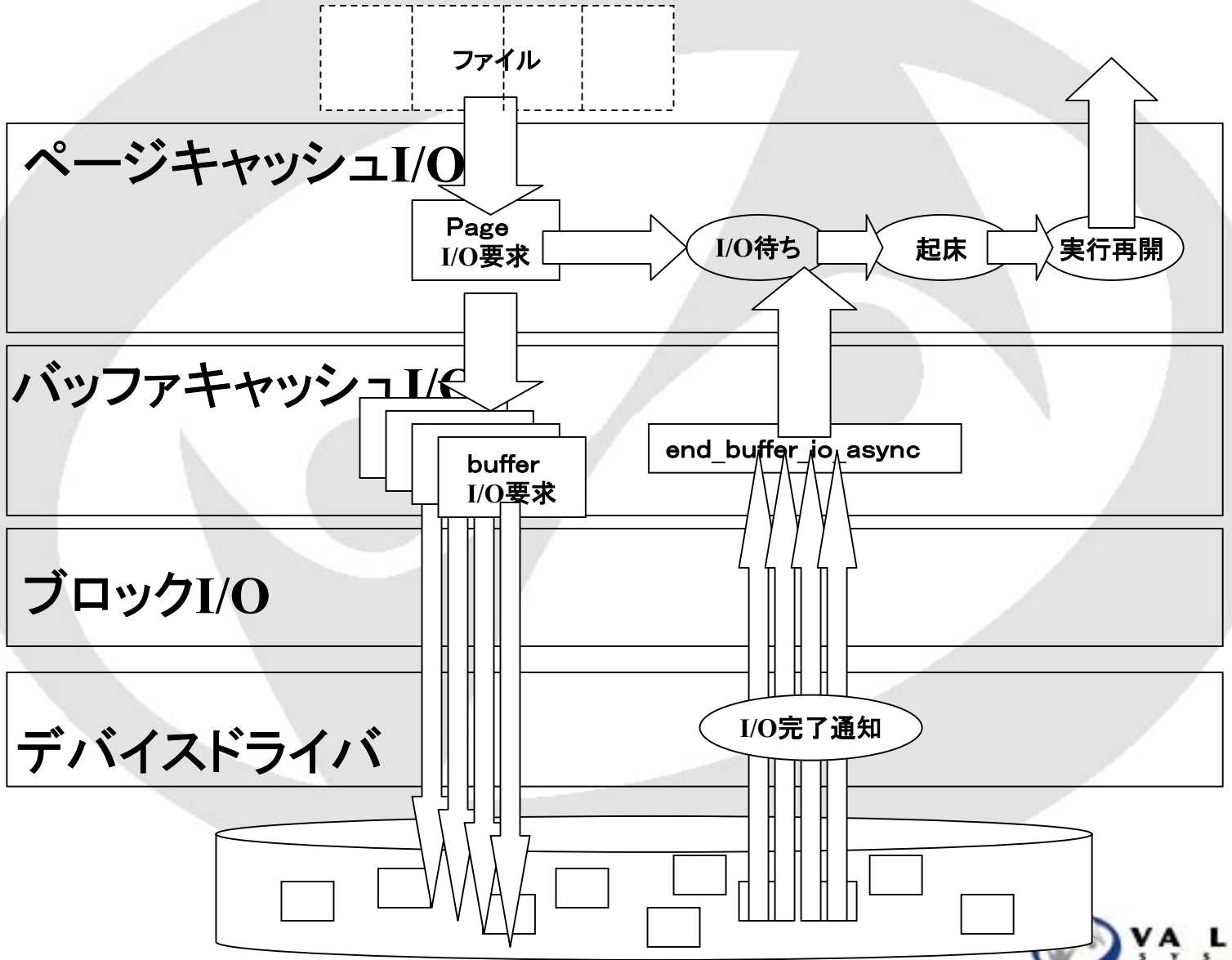
VA LINUX SYSTEMS JAPAN

ファイルとページキャッシュ



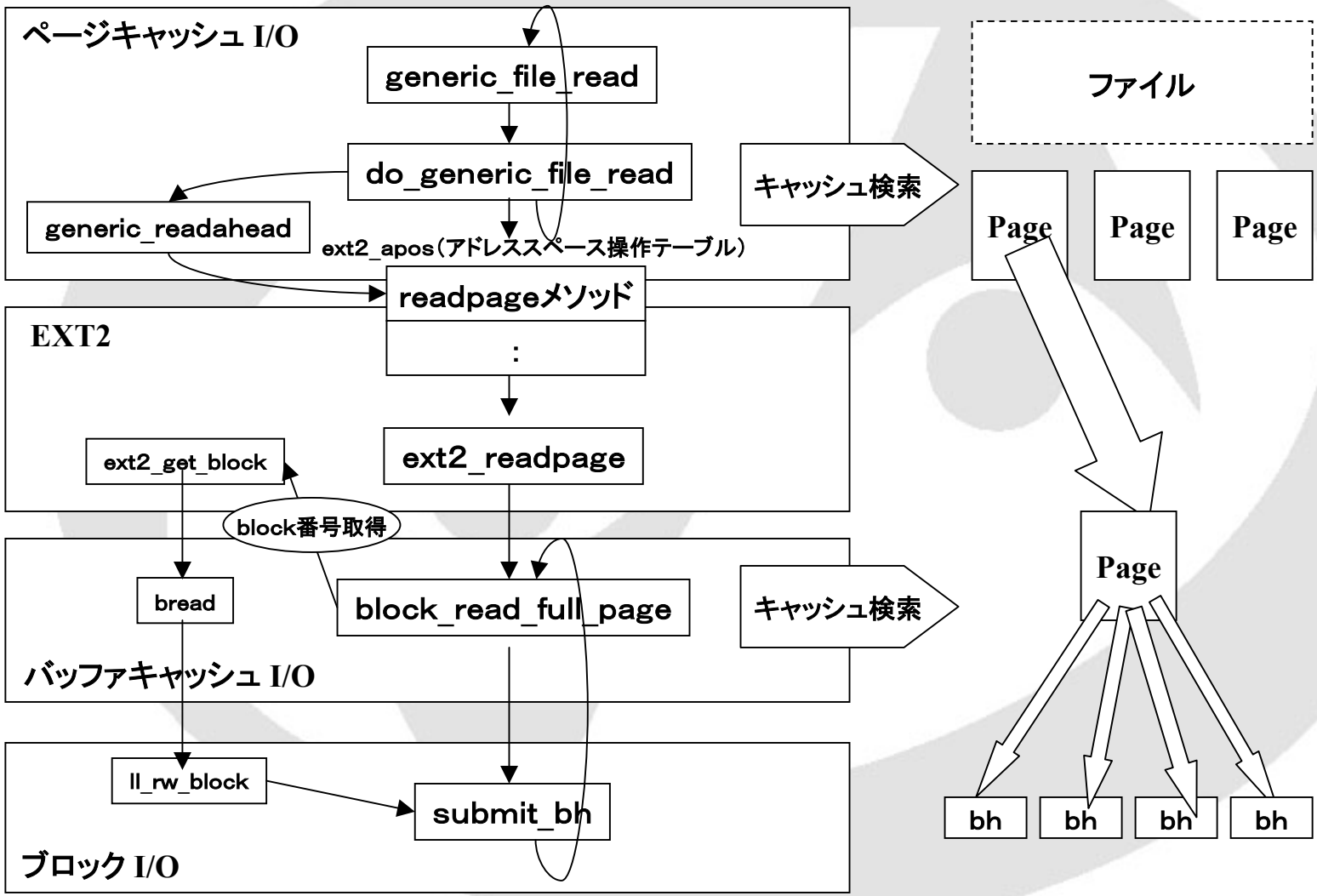
VA LINUX SYSTEMS JAPAN

ページキャッシュI／Oの概要



VA LINUX SYSTEMS JAPAN

通常ファイル読み込みの流れ

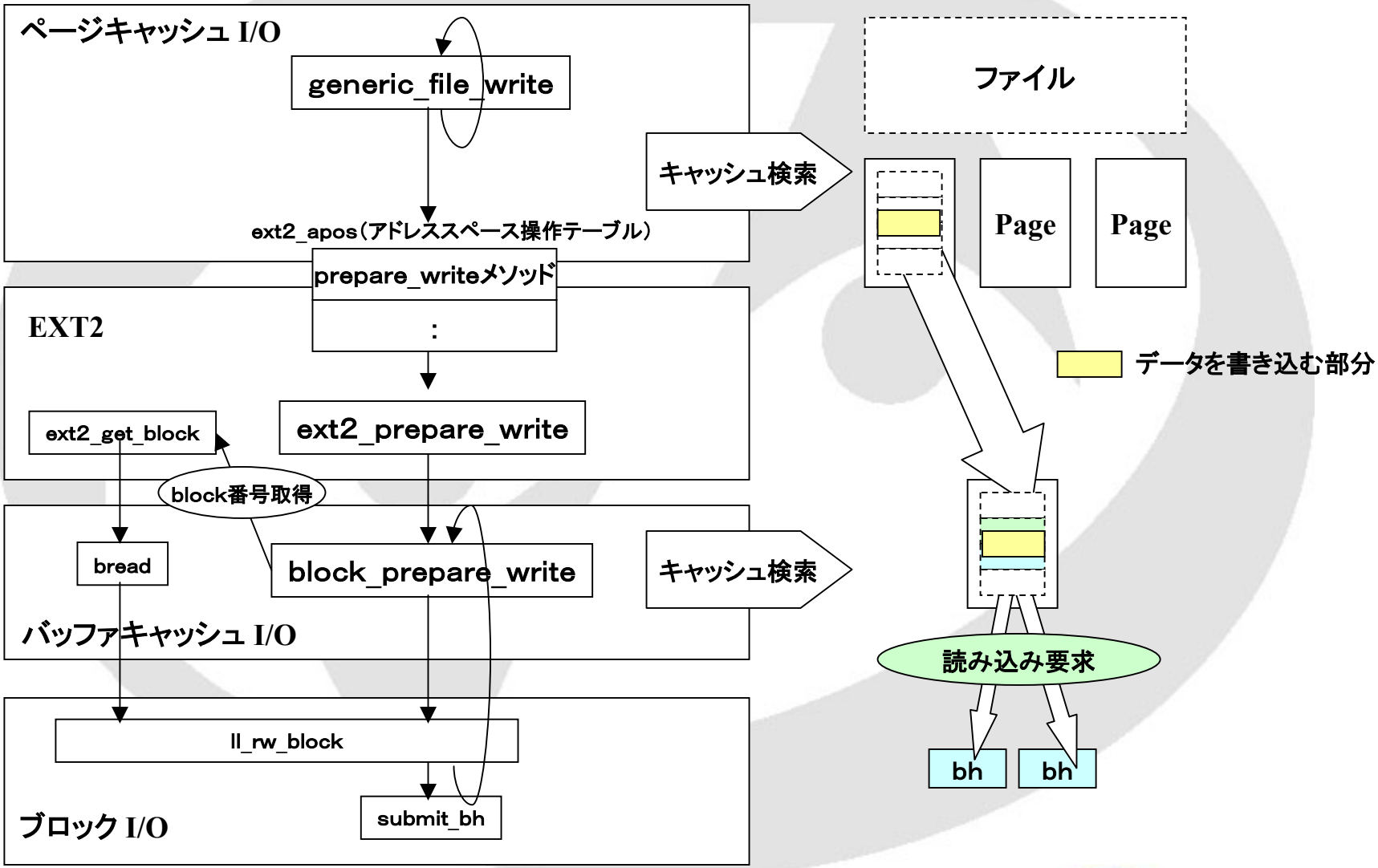


VA LINUX SYSTEMS JAPAN

通常ファイルの書き込み

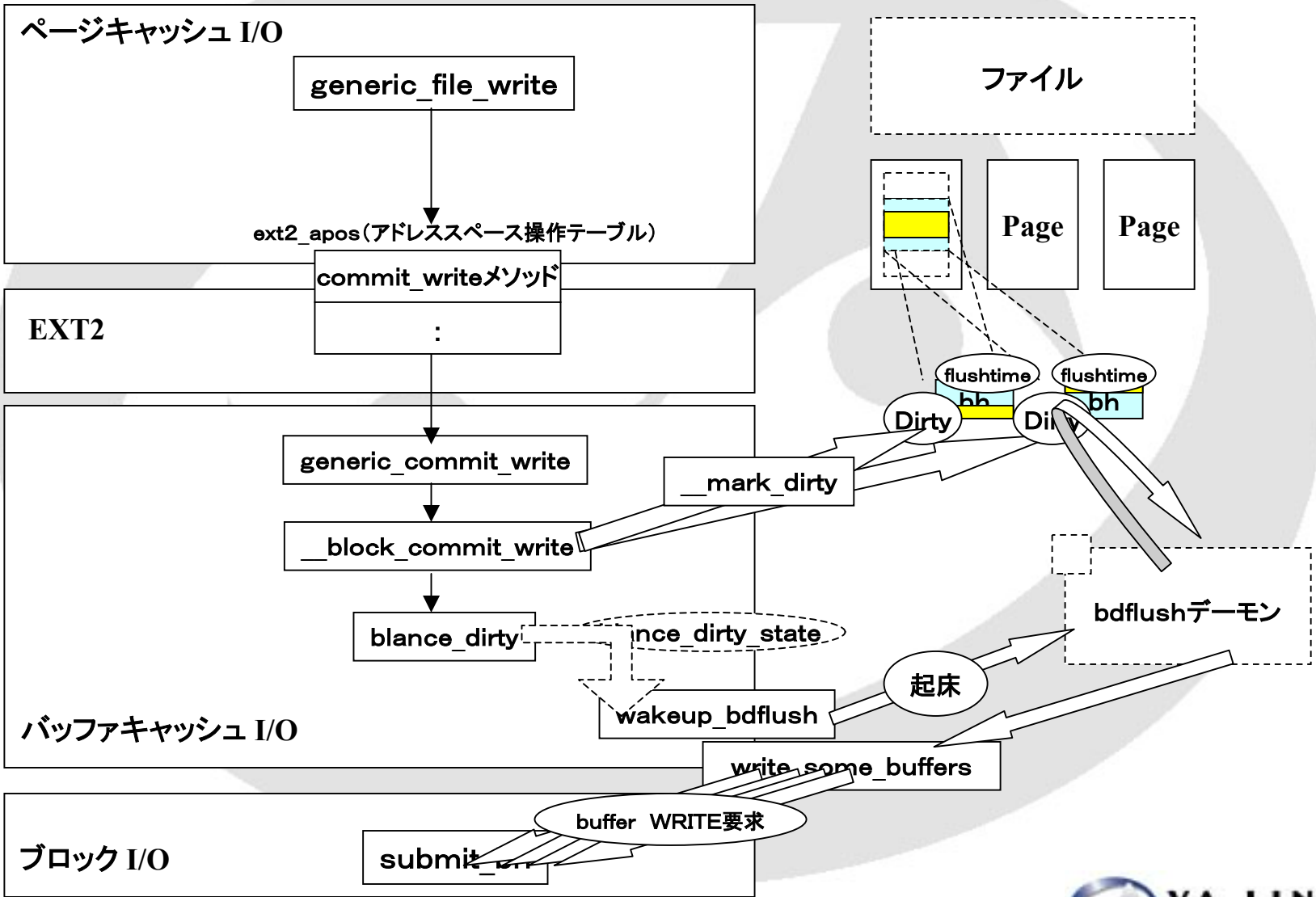
- prepare_write
- 書き込みデータコピー
- commit_write

通常ファイル書き込みの流れ（prepare_write）



VA LINUX SYSTEMS JAPAN

通常ファイル書き込みの流れ（commit_write）



VA LINUX SYSTEMS JAPAN

bdflushとkupdateデーモン

- bdflushデーモン
  - Dirty属性（遅延書き込み）のバッファが一定数以上になったらディスクに書き戻す
- kupdateデーモン
  - Dirty属性になって一定時間を経過したバッファをディスクに書き戻す
- その他のディスクへの書き戻し処理
  - updateコマンド（bdflushシステムコール）
  - syncコマンド（syncシステムコール）
  - 空きメモリの枯渇

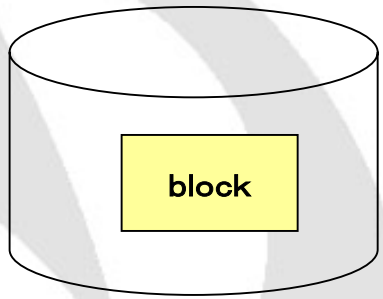
ブロックI／O共通レイヤ

VA LINUX SYSTEMS JAPAN

バッファI/Oとブロック I/O

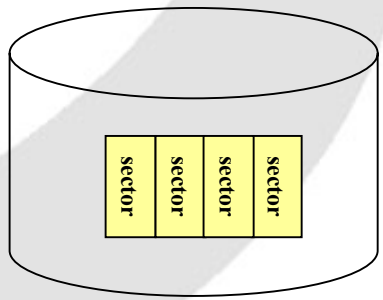
バッファI/O

- ブロック単位で処理
- ブロックサイズはファイルシステム依存
  （ページサイズは超えられない）



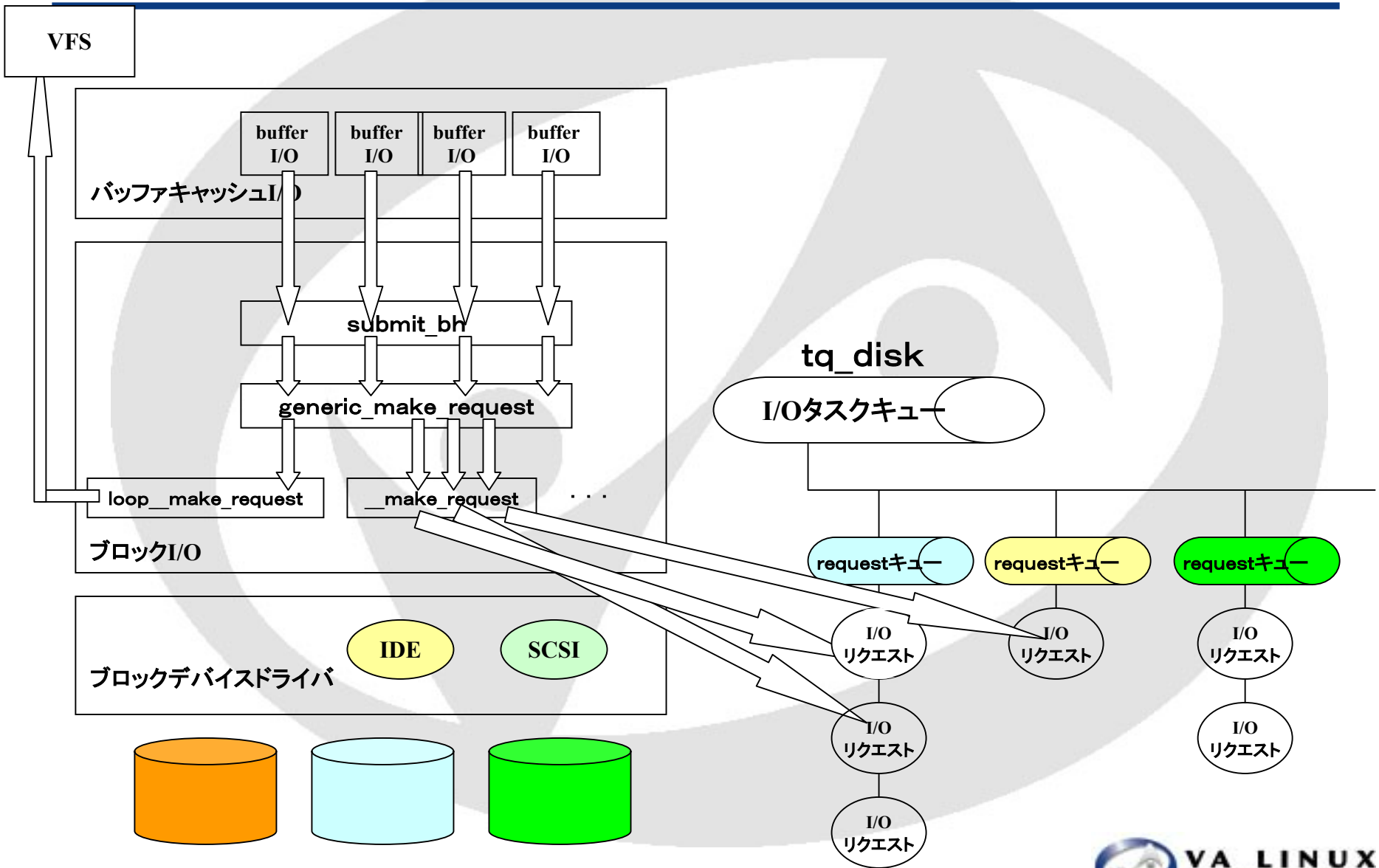
ブロックI/O共通レイヤ

- セクタ単位で処理
- セクタサイズは512バイト
  （ファイルシステムサイズの制限）



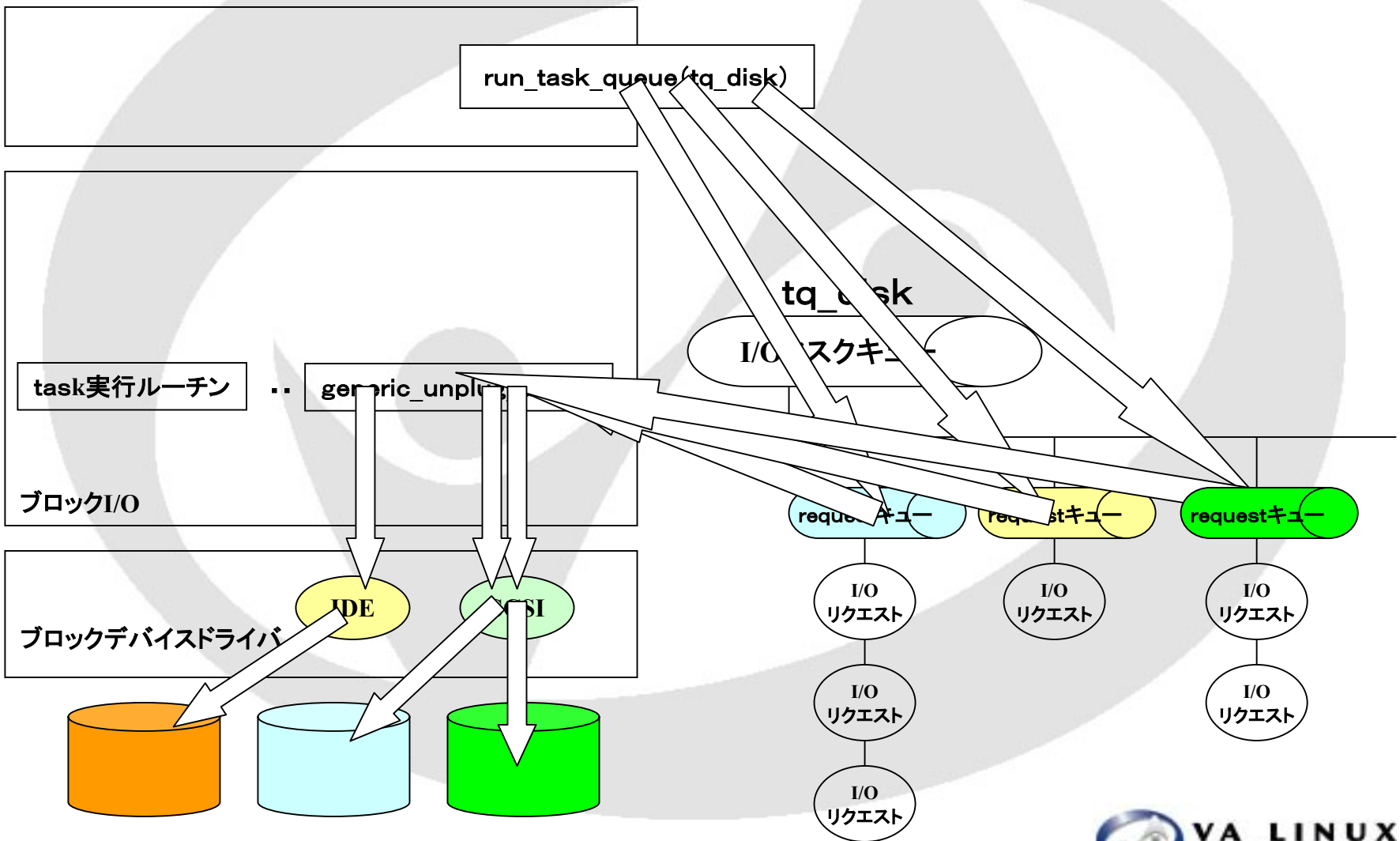
VA LINUX SYSTEMS JAPAN

I/Oリクエストの作成



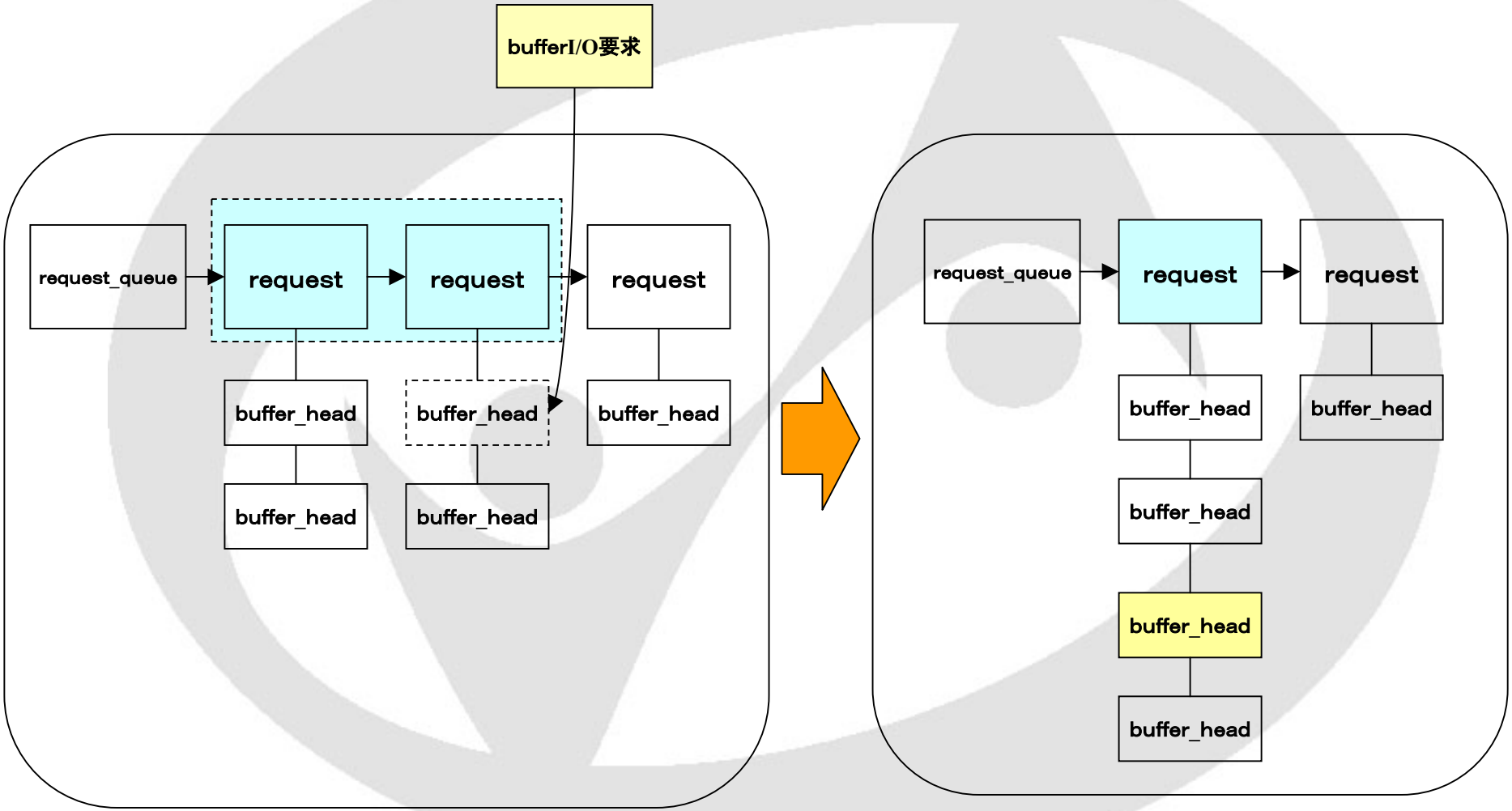
VA LINUX SYSTEMS JAPAN

I/Oタスクの実行



VA LINUX SYSTEMS JAPAN

ブロックI／Oのクラスタリング



VA LINUX SYSTEMS JAPAN

おわり

http:/valinux.co.jp/

VA LINUX SYSTEMS JAPAN