HITACHI
Inspire the Next

HiRDB技術解説

HiRDB

HiRDBアーキテクチャと運用管理解説

2011/01

株式会社 日立製作所 情報・通信システム社
ITプラットフォーム事業本部 開発統括本部 DB設計部

Human Dreams.
Make IT Real.

HITACHI
Inspire the Next

Opening

HiRDB

クラウド時代を支える「ワンランク上の」高性能・高信頼データベース HiRDB Version 9をリリースしました。
本資料では、HiRDBの基本的なアーキテクチャを説明し、そのアーキテクチャを前提としたHiRDBの運用管理の基本と、運用管理について説明します。

HITACHI
Inspire the Next

Contents

HiRDB

HITACHI
Inspire the Next

HiRDB

1. HiRDBとは？

HITACHI
Inspire the Next

「止めない」設計思想を貫く
高信頼ノンストップデータベース

社会基盤を支えるために
日立が自社開発にこだわり続ける純国産RDBMS

ハイアールディービー
HiRDB Version 9
Highly Scalable Relational DataBase

今まで培った信頼性をベースに
クラウド時代を支える「ワンランク上の」
データベースを目指します。

HITACHI
Inspire the Next

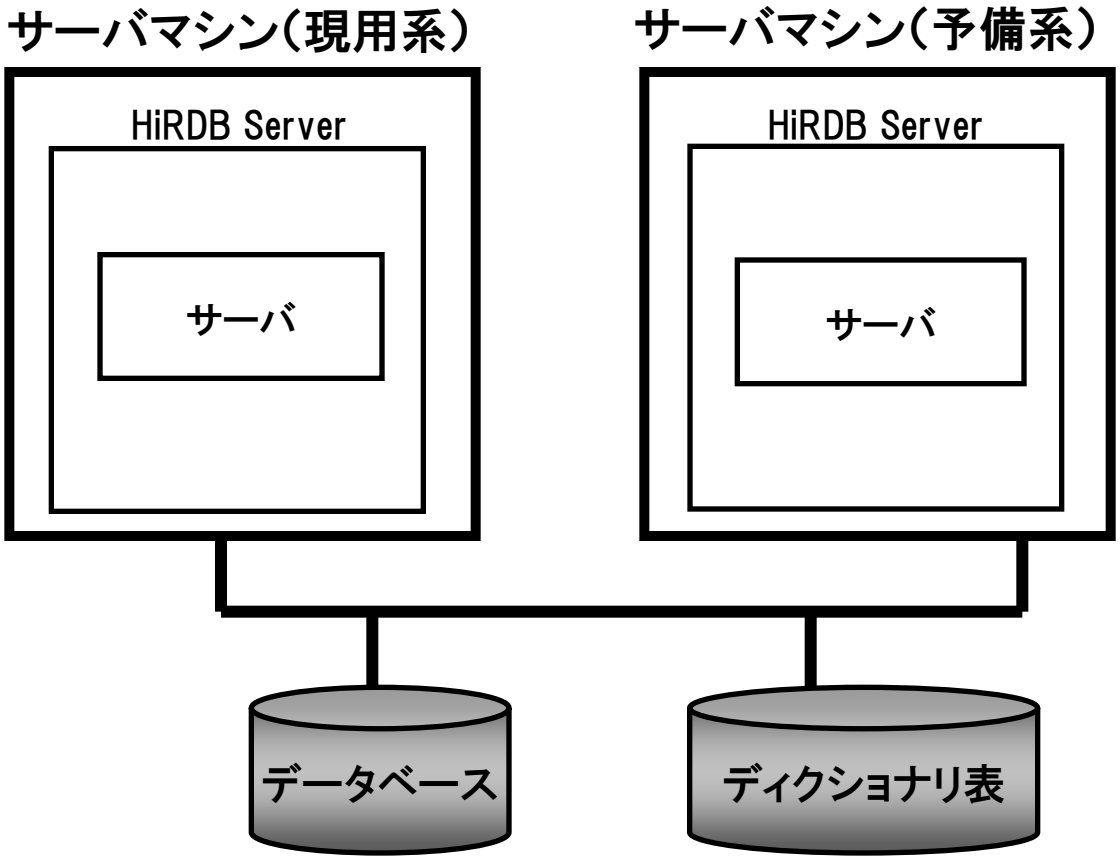
HiRDB

2. HiRDBのアーキテクチャ

2-1 サーバの構成の概要

HITACHI
Inspire the Next

解説 HiRDB Server Version 9の基本的な構成を示します。

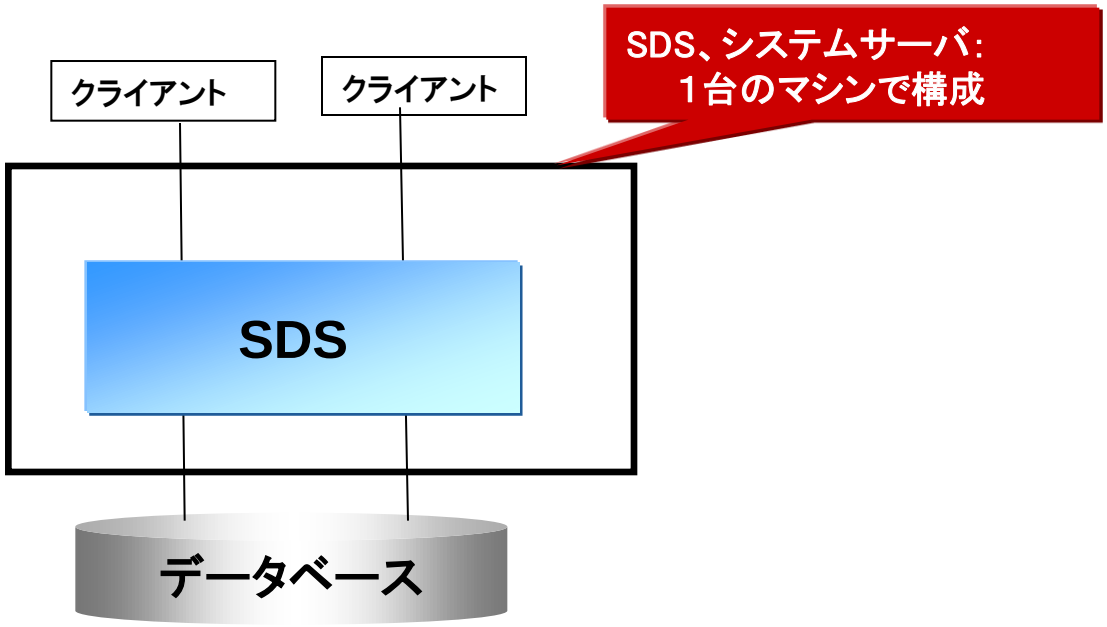


HiRDB Server Version 9は、Single ServerとParallel Serverの2つを含んでおり、構築時にどちらかを選択します。

2-1 サーバの構成の詳細

HITACHI
Inspire the Next

解説 Single Server構成のサーバ構成について説明します。



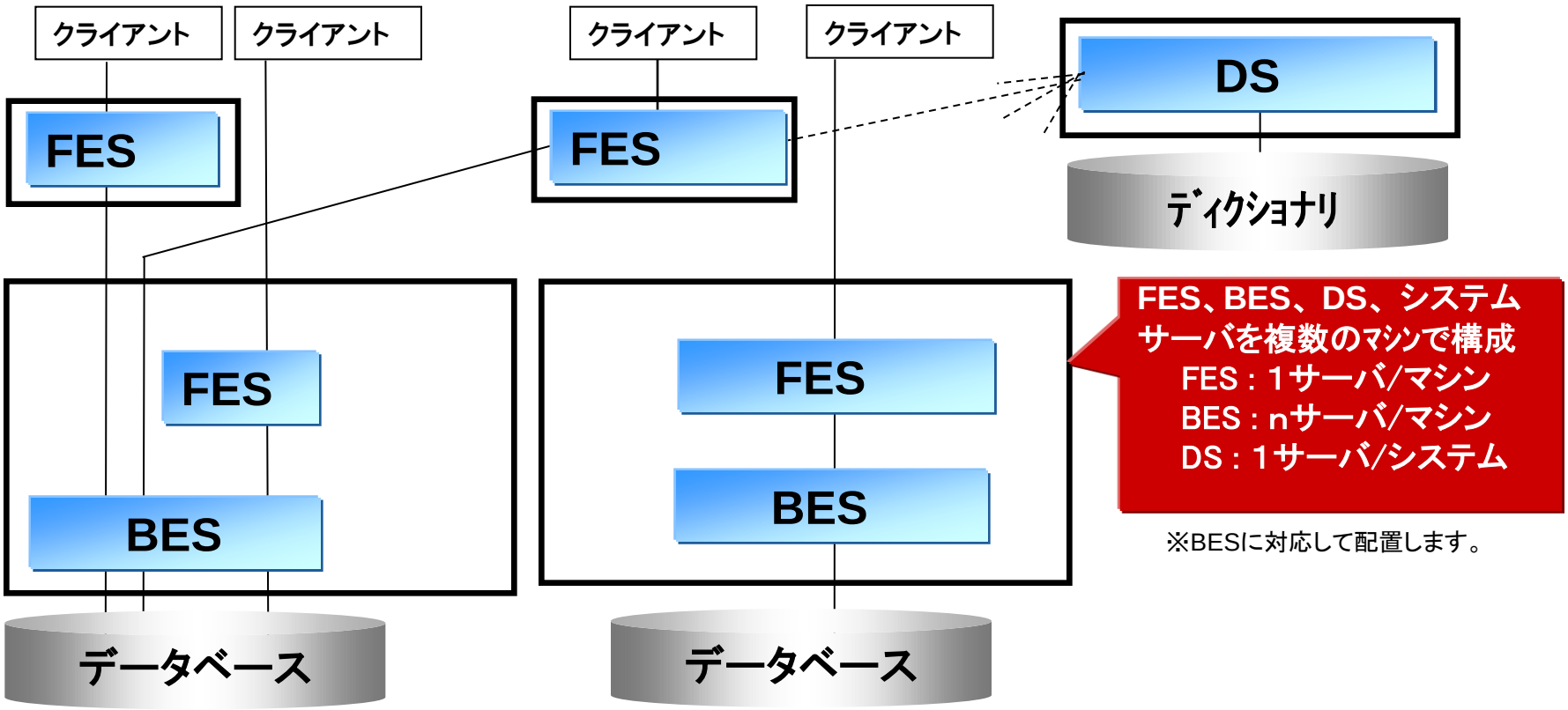
用語

SDS(Single Database Server)・・・DB処理を行う

2-1 サーバの構成の詳細

HITACHI
Inspire the Next

解説 Parallel Server 構成のサーバ構成について説明します。



※BESに対応して配置します。

用語

FES(Front End Server)･･･ SQL受付処理を行う
BES(Back End Server)･･･ DB処理を行う
DS(Dictionary Server) ･･･ディクショナリの管理を行う

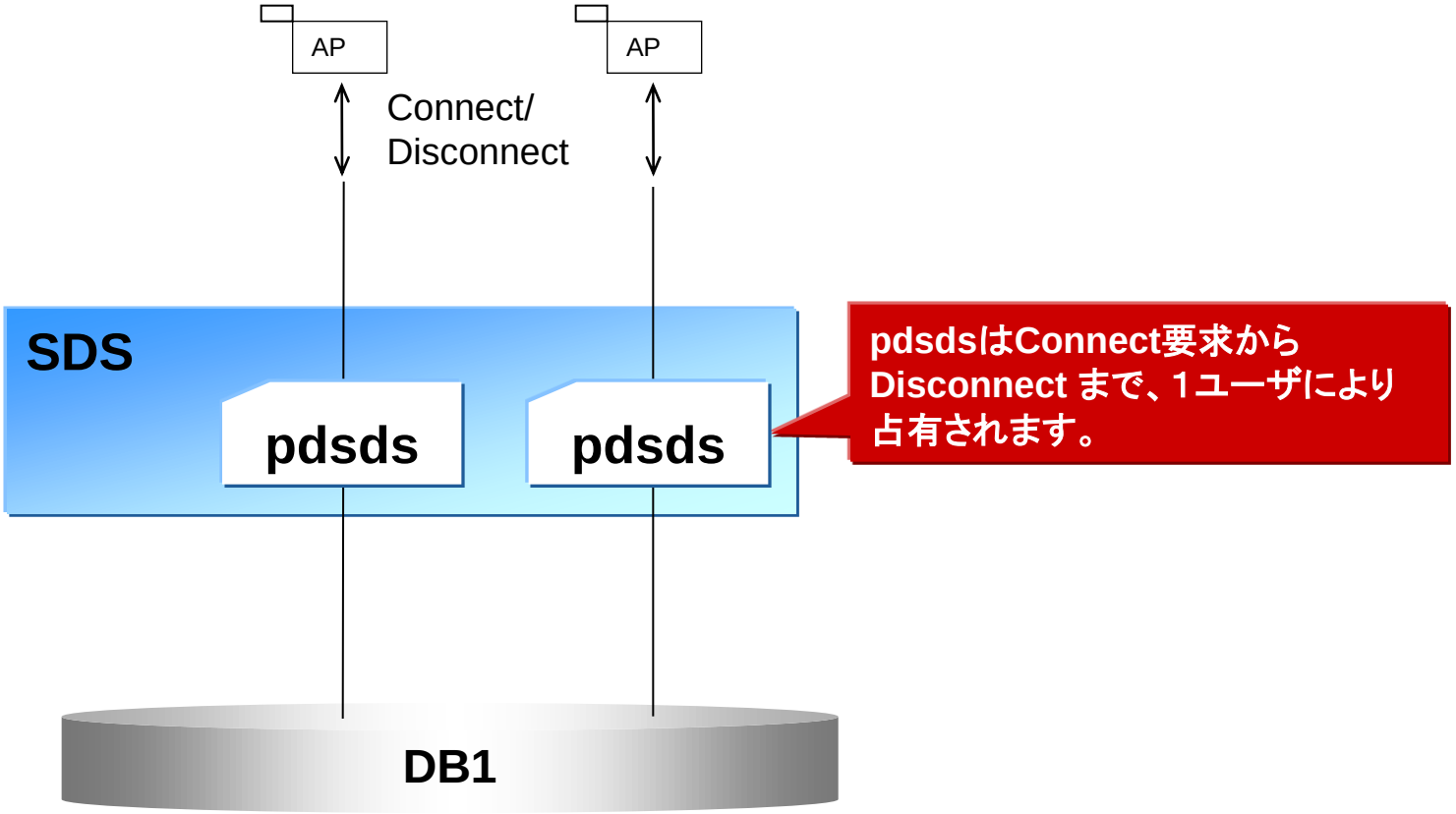
Single Server構成
ユーザサーバプロセスとアプリケーションの関係

HITACHI
Inspire the Next

解説

Single Server構成のユーザサーバプロセス(※1)とアプリケーションの関係を示します。
アプリケーションはユーザサーバプロセスのpdsdsに接続します。

※1 DB処理をするプロセス。Single Server構成のときはpdsds



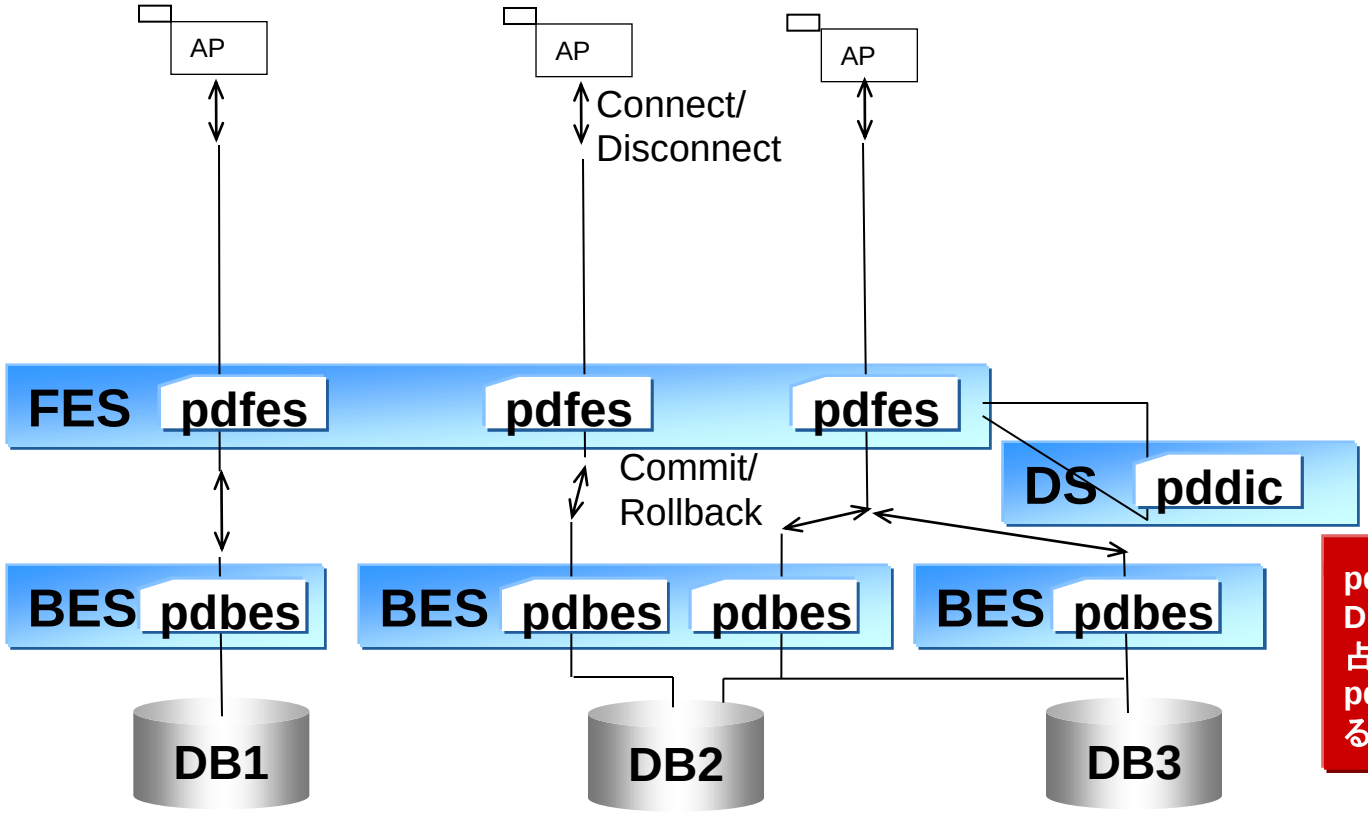
ユーザサーバプロセスとアプリケーションの関係

HITACHI
Inspire the Next

解説

Parallel Server 構成のユーザサーバプロセス(※1)とアプリケーションの関係を示します。
アプリケーションはユーザサーバプロセスのpdfesに接続します。

※1 DB処理をするプロセス。Parallel Server構成のときはpdfes、pdbes、pddic



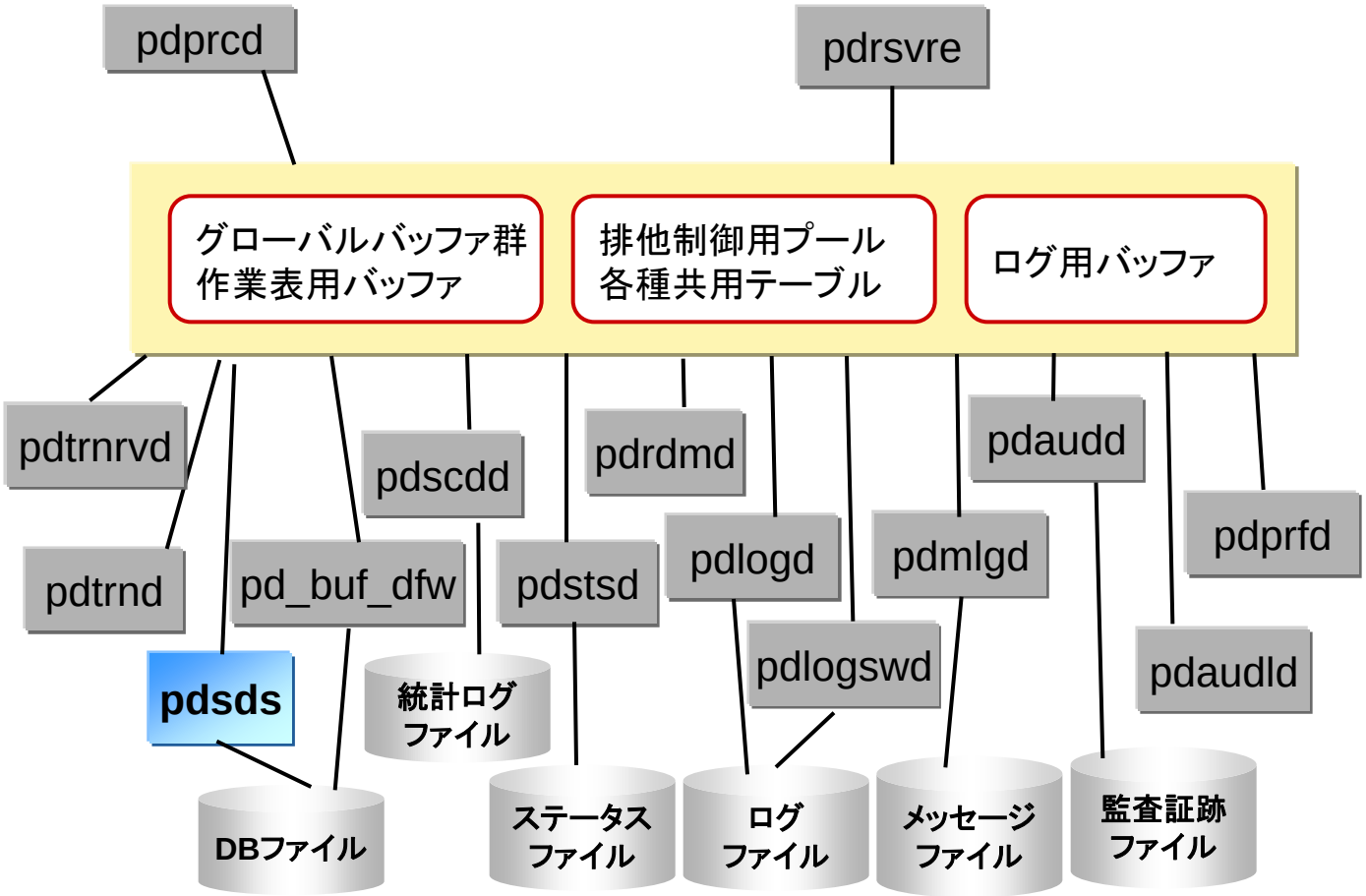
pdfesはConnect要求から
Disconnectまで、1ユーザにより
占有されます。
pdbesはトランザクションが終了す
るまで占有されます。

ユーザサーバプロセスとシステムサーバプロセスの関連

HITACHI
Inspire the Next

解説 Single Server構成のユーザサーバプロセスとシステムサーバプロセス(※)の関連を示します。

※メモリ管理やシステムファイルの制御をするプロセス

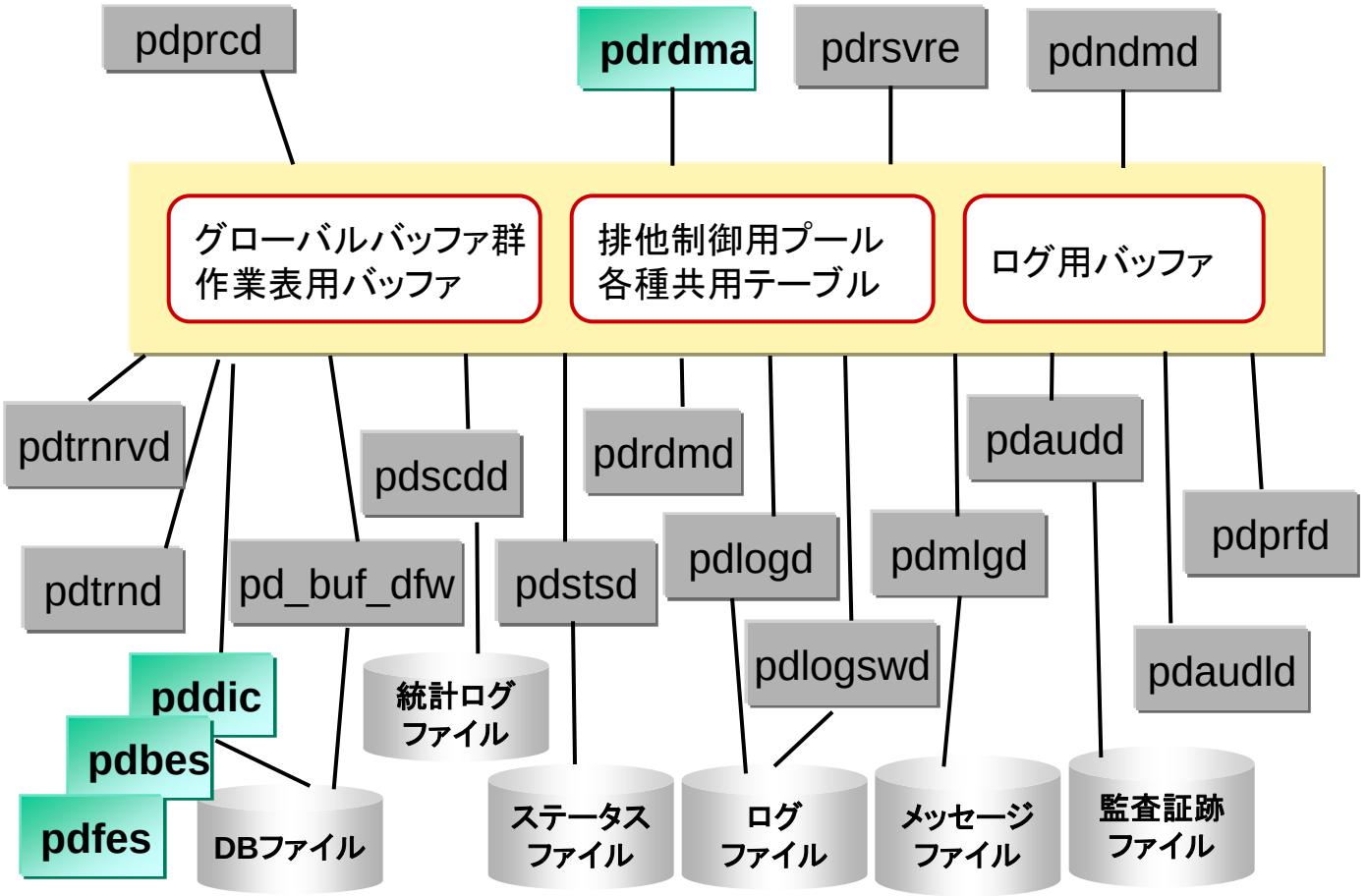


ユーザサーバプロセスとシステムサーバプロセスの関連

HITACHI
Inspire the Next

解説 Parallel Server構成のユーザサーバプロセスとシステムサーバプロセス(※)の関連を示します。

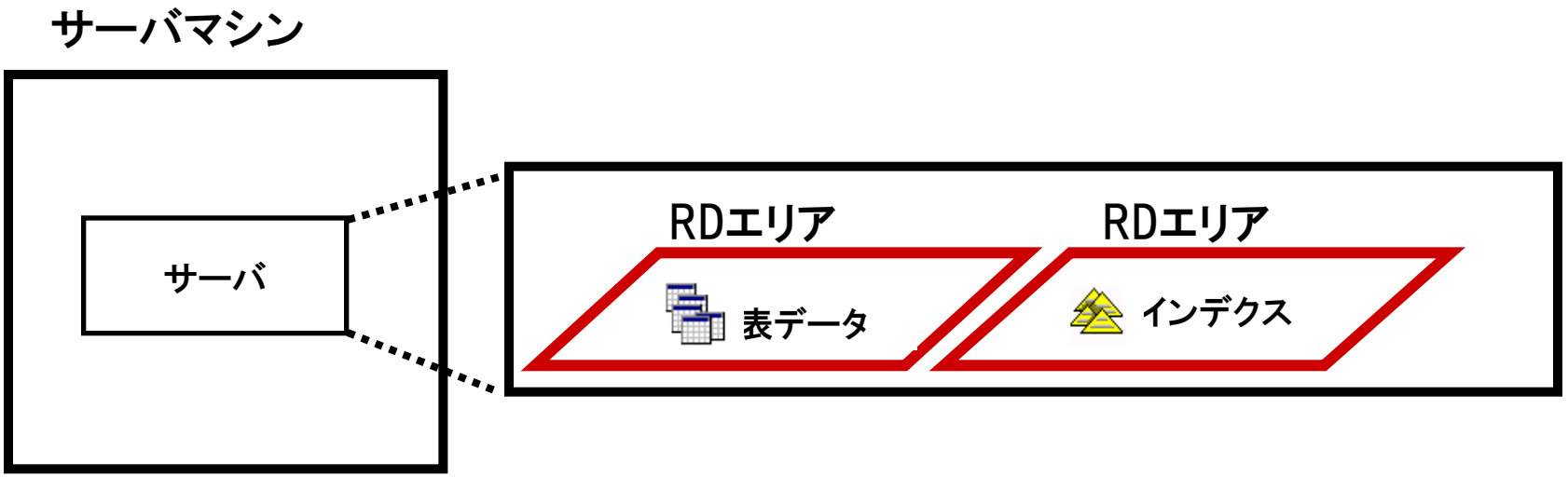
※メモリ管理やシステムファイルの制御をするプロセス



2-2 データベースの格納領域の概要

HITACHI
Inspire the Next

解説 RDエリアとは、表およびインデクスを格納するディスク上の論理的な領域です。

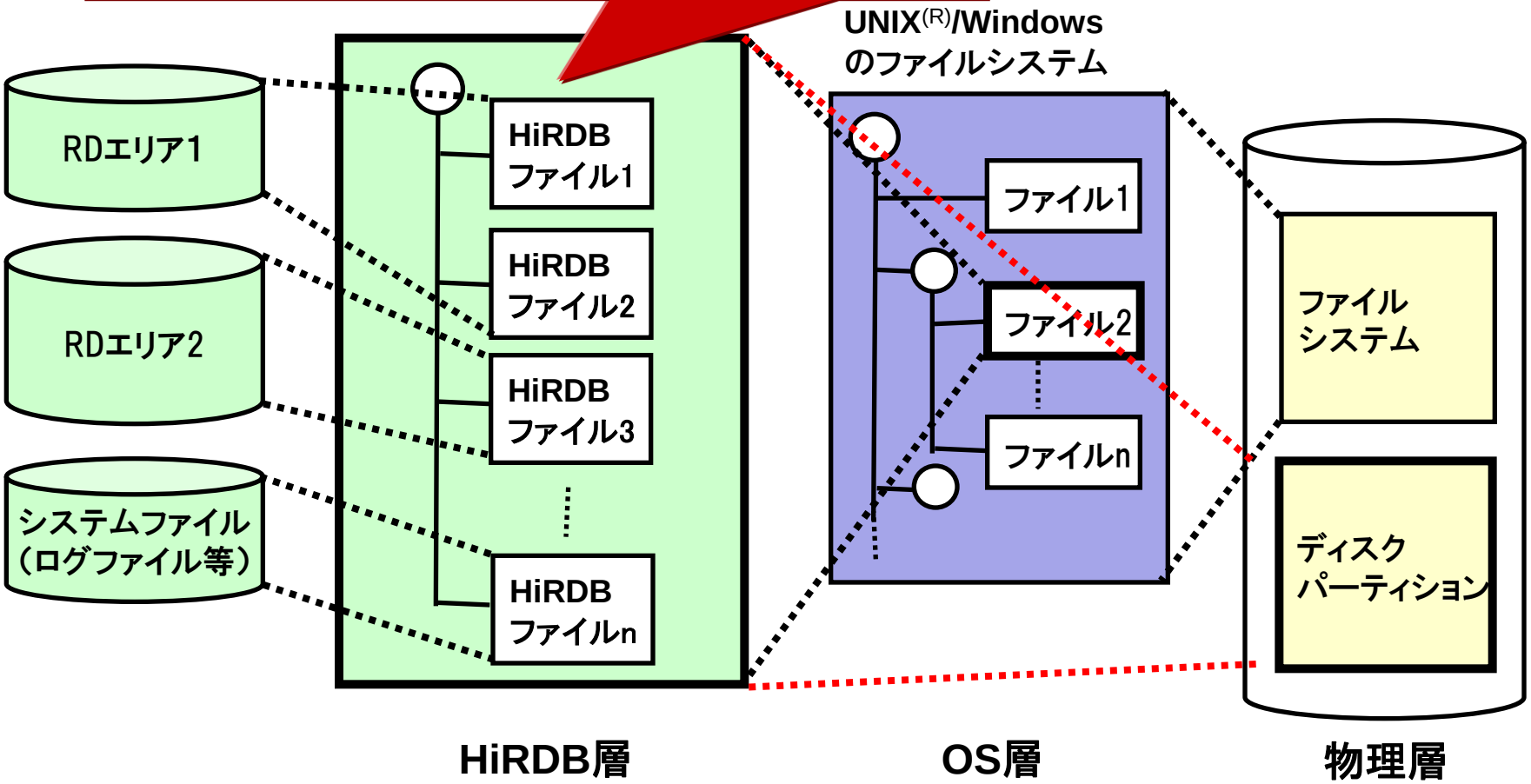


2-2 RDエリアとファイル構成

HITACHI
Inspire the Next

解説 RDエリアはHiRDBファイルから構成されます。HiRDBファイルはHiRDBファイルシステム領域に確保され、HiRDBファイルシステム領域はOSのファイルシステムのファイル上もしくはディスクパーティション上に確保します。

HiRDBファイルシステム領域：
→ HiRDB自身で独自のファイルシステムを実現しています。



2-3 RDエリア(1):種類

HITACHI
Inspire the Next

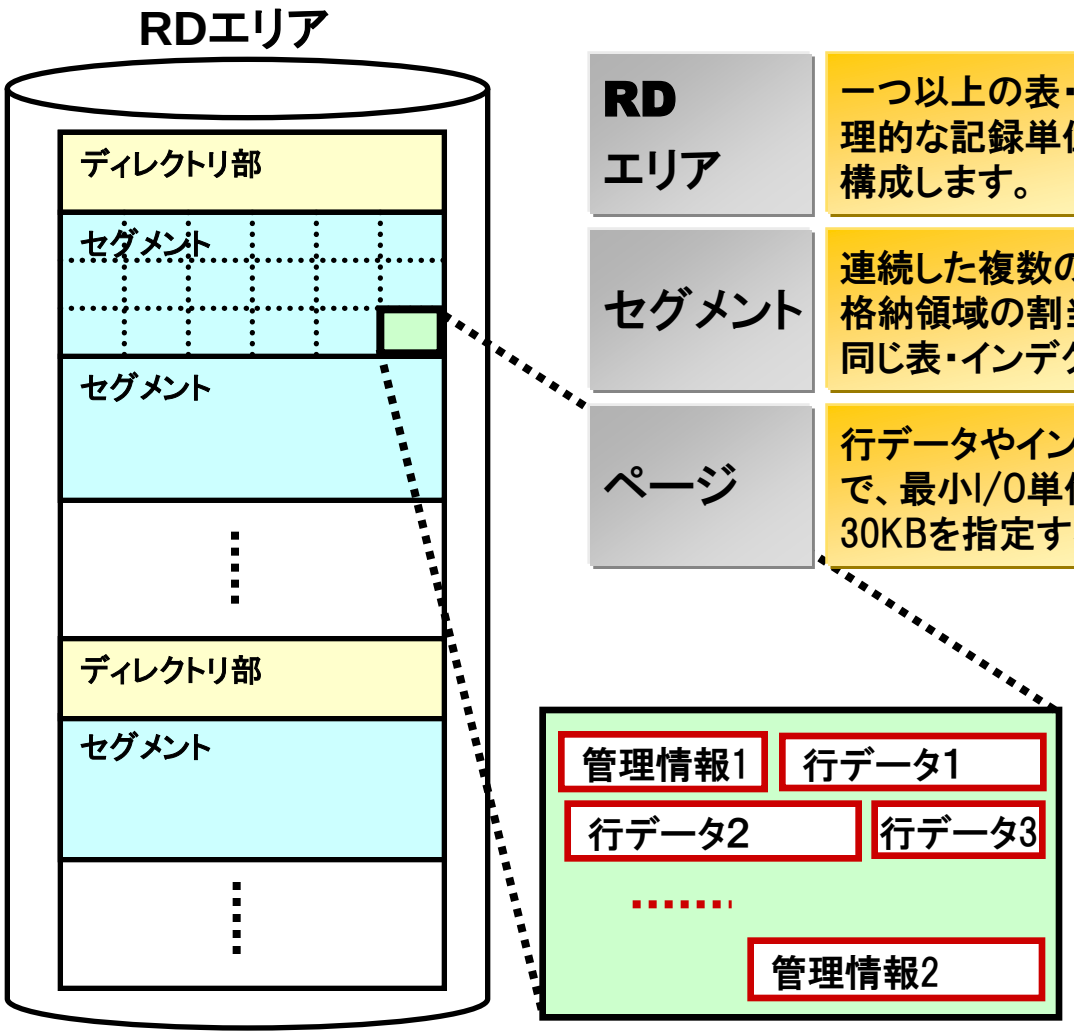
解説 RDエリアには種類があり、用途に応じて使い分けます。RDエリアの種類を示します。

#	種別	RDエリアの種類 (括弧内はパラレルサーバの場合の管理元サーバ)	説明
1	システムのデータを格納	マスタディレクトリ(DS)	システムの内部情報を格納します。
2		ディクショナリ(DS)	ディクショナリ表およびディクショナリ表のインデクスを格納します。
3		データディレクトリ(DS)	システムの内部情報を格納します。
4		ディクショナリLob用(DS)	ストアドプロシジャまたはストアドファンクションの定義ソースおよびオブジェクトを格納します
5		レジストリ(DS)	レジストリ情報を格納します。
6		レジストリ用 Lob(DS)	レジストリ情報に登録したキーのうち、キー長が32,000バイトを超えるものを格納します
7	ユーザのデータを格納	ユーザ(BES)	表およびインデクスを格納します
8		ユーザLob用(BES)	文書、画像、音声などの長大な可変長データを格納します
9		リスト(BES)	ASSIGN LIST文で作成するリストを格納します

2-3 RDエリア(2):構造

HITACHI
Inspire the Next

解説 RDエリアは、表やインデクスを格納するために必要な論理的な単位です。RDエリアはセグメントで構成され、セグメントはページから構成されます。

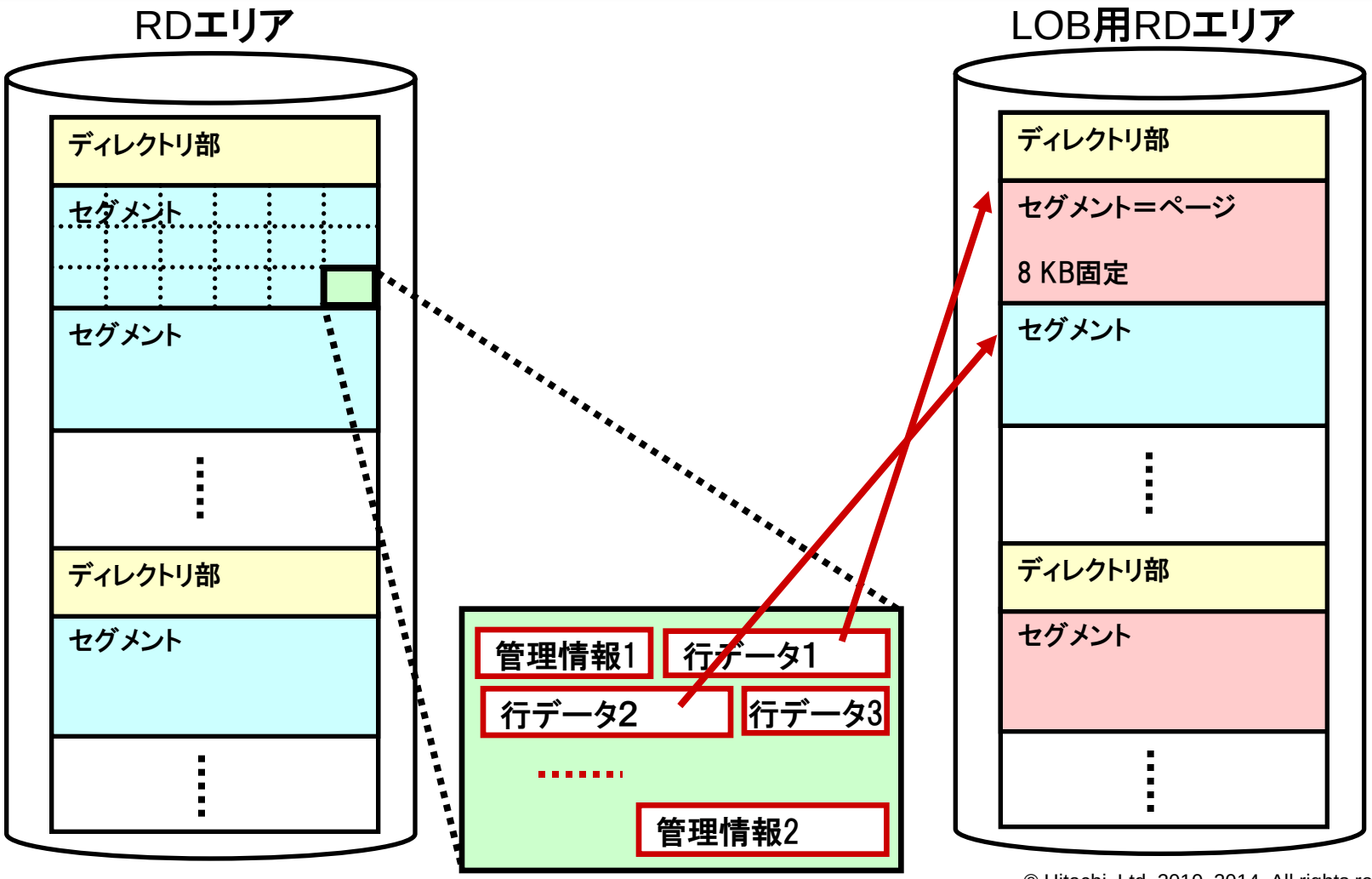


RDエリア	一つ以上の表・インデクスを記録する為の論理的な記録単位。1～16個のHiRDBファイルで構成します。
セグメント	連続した複数のページからなる表・インデクス格納領域の割当単位。一つのセグメントには同じ表・インデクスしか格納しません。
ページ	行データやインデクスキー値を格納する単位で、最小I/O単位になります。大きさは4～30KBを指定することができます。

2-3 RDエリア(3):BLOB用

HITACHI
Inspire the Next

解説 LOB用RDエリアとは、BLOBデータを格納する専用のエリアのことです。BLOBの列ひとつに対して一つのRDエリアが必要になります。「1セグメント＝1ページ」で、ページサイズは8KB固定となります。

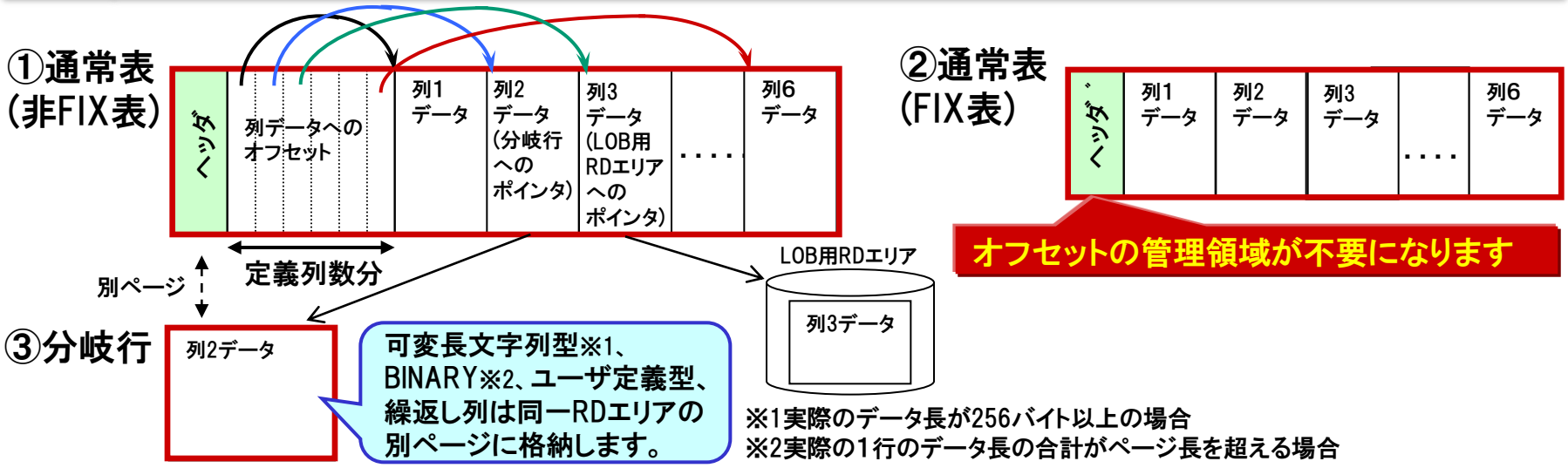


2-4 行の構造

HITACHI
Inspire the Next

解説

HiRDBの行のデータの構造には、非FIX表に対応する列データへのオフセットを持つ①の形式と、オフセットを持たない②の形式があります。①の形式において、列データが可変長文字列型のデータなどの場合、列データを別のページに格納する③の形式があります。BLOBの場合にはLOB用RDエリアに格納します。

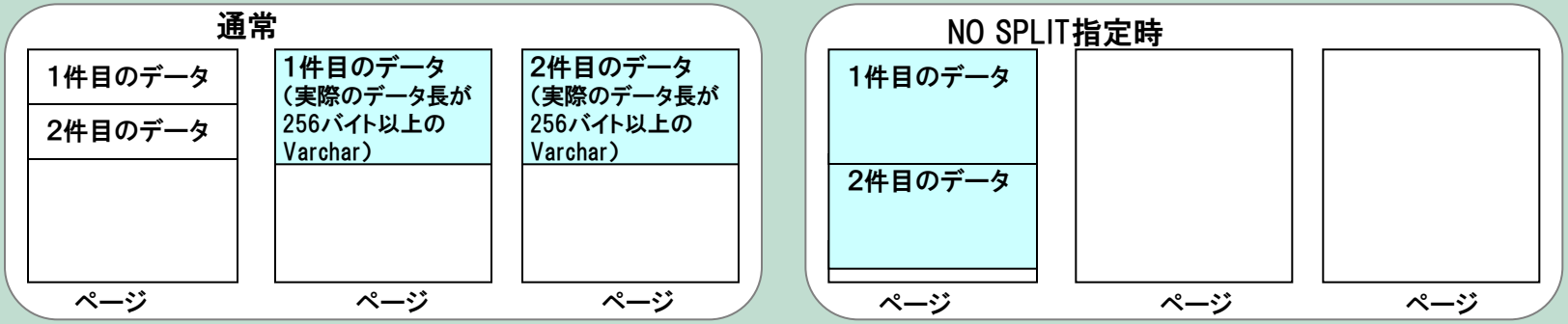


※1実際のデータ長が256バイト以上の場合
※2実際の1行のデータ長の合計がページ長を超える場合

補足事項

可変長文字列型の場合、ノースプリットオプション(CREATE TABLE時にNO SPLIT)を指定した場合は、基本行に長データも格納します。

例



2-5 更新処理方式

HITACHI
Inspire the Next

解説 HiRDBでは、どのように更新処理を行っているかの概要をUPDATEを例に説明します。

UPDATE （A→B ）

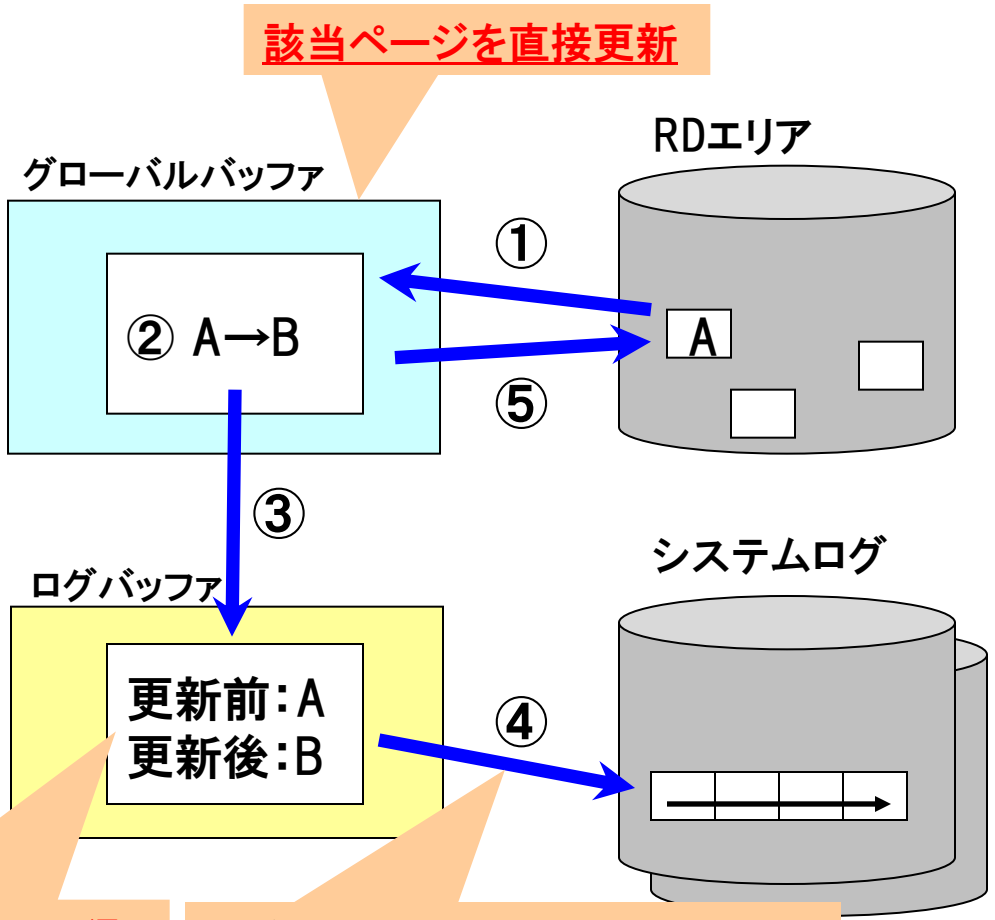
① 該当ページをグローバルバッファに読み込みます。

② グローバルバッファ上で対象行データをAからBに更新します。

③ 更新前・後ログを取得します。

④ ログバッファが溢れた時、またはコミット時にログレコードを書き込みます。

⑤ シンクポイントダンプ処理の時、または更新ページが一定の値に達した時にDBに書き込みます。



更新前後情報を1回のログで取得

ログレスモードでは、DB更新ジャーナルの取得抑止が可能

HITACHI
Inspire the Next

HiRDB

3. HiRDBの運用管理

HITACHI
Inspire the Next

HiRDB

3-1. データベース運用

3-1 データベース運用の一覧

HITACHI
Inspire the Next

解説

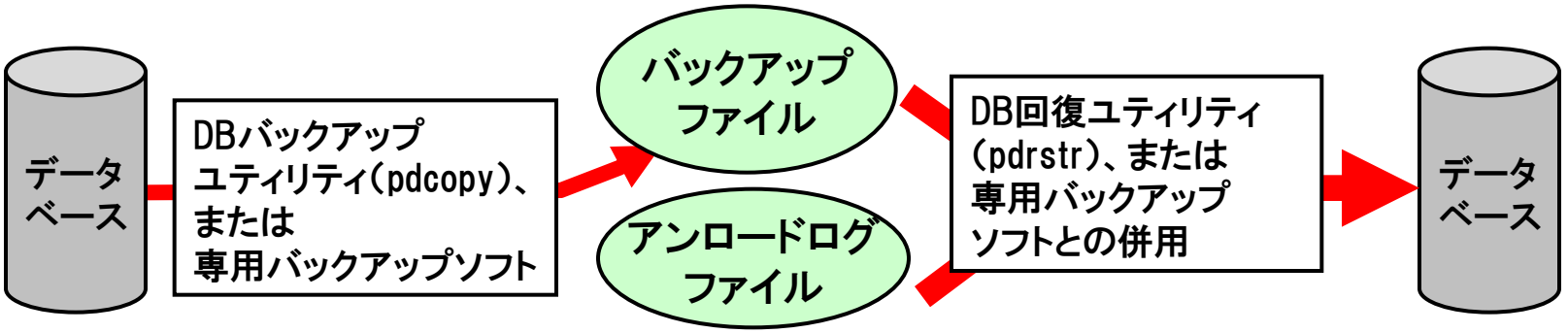
HiRDBの主要なデータベース運用には、以下のものがあります。
本節では、これらについて解説します。

#	運用の種別	解説箇所
1	バックアップ・リストアの運用	3-1-1
2	データロード・アンロードの運用	3-1-2
3	自動増分の運用	3-1-3
4	再編成の運用	3-1-4
5	空きページ解放の運用	3-1-4
6	構成変更の運用	3-1-5

3-1-1 バックアップ・リストア運用の目的

HITACHI
Inspire the Next

解説 データベースの障害に備えて、バックアップを取得しておき、バックアップからリストアしてデータベースを回復できるようにしておくことが大切です。



＜バックアップ運用の考慮事項＞
バックアップの要否、バックアップサイクル、バックアップ方法(オン中、業務停止中)、回復要件(いつの時点に戻れば良いか？)、バックアップ時間・回復時間　など

3-1-1 バックアップ方式の概要と特徴

HITACHI
Inspire the Next

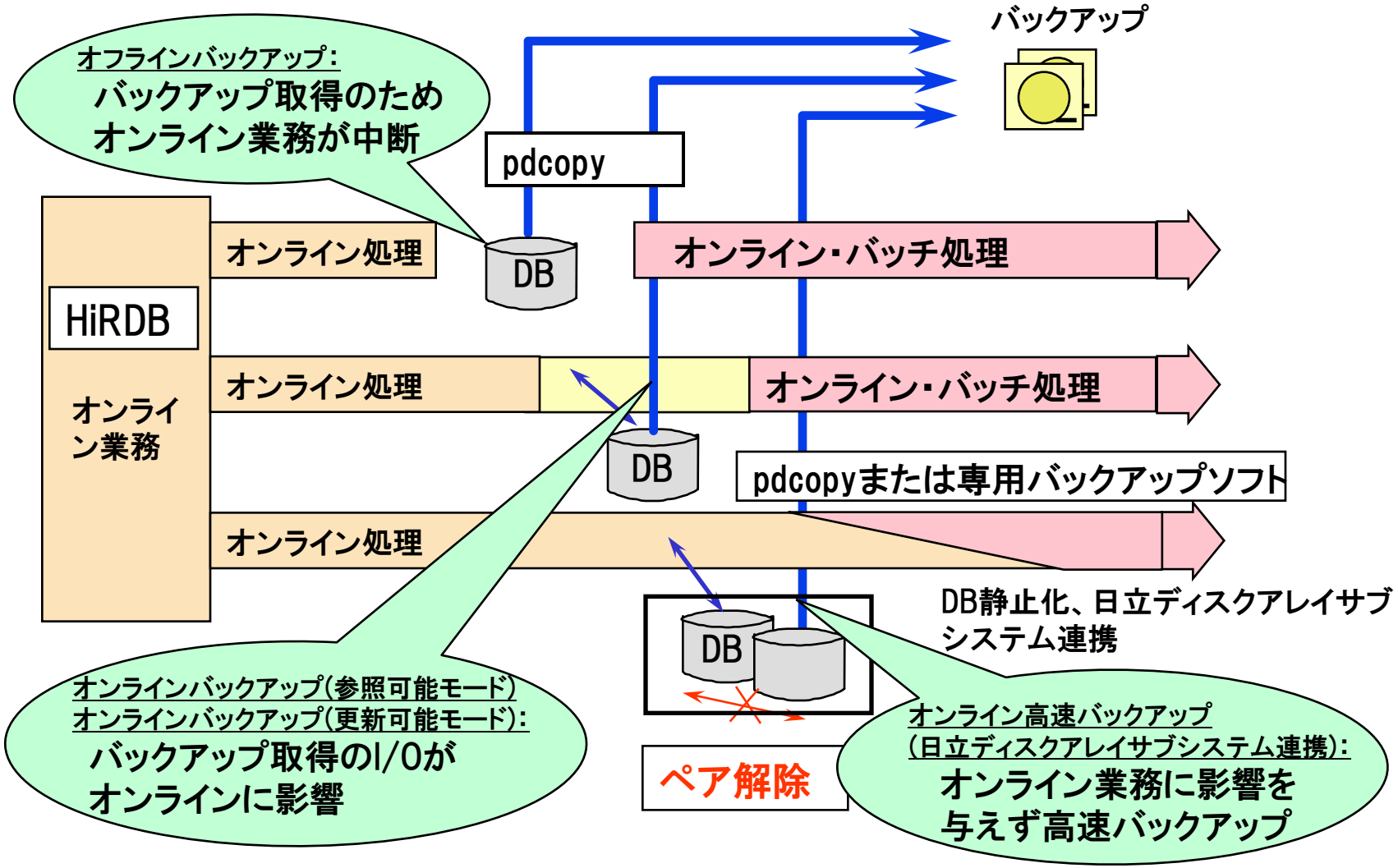
解説 バックアップ方式の概要と特徴についてまとめます。

#	方式	説明	方法	特徴
1	オフライン バックアップ	業務を停止した状態でバックアップします。	・pdcopy –M x ・pdstart –rで起動 ・JP1/VERITAS NetBackup連携	業務は停止する必要がありますが、業務とバックアップ処理とのディスクI/O競合が発生しません。
2	オンライン バックアップ(参照可能モード)	参照業務のみ継続した状態でバックアップします。	・pdcopy –M r ・JP1/VERITAS NetBackup連携	業務とバックアップ処理とのディスクI/O競合が発生しますが、参照業務は継続できます。
3	オンライン バックアップ(更新可能モード)	参照および更新を含む業務を継続した状態でバックアップします。	・pdcopy –M s ・JP1/VERITAS NetBackup連携	業務とバックアップ処理とのディスクI/O競合が発生しますが、参照および更新業務は継続できます。 ただし、この状態で取得したバックアップは、必ずシステムログと組み合わせて回復します。
4	オンライン高速 バックアップ (日立ディスクアレイサブシステム連携)	参照および更新を含む業務を継続した状態でバックアップします。	・pdcopy –M x ・pdcopy –M r ・JP1/VERITAS NetBackup連携	業務とバックアップ処理とのディスクI/O競合が発生しますが、参照および更新業務は継続できます。 レプリカDBを作成するためのディスク容量が必要です。

3-1-1 各バックアップ方式の特徴

HITACHI
Inspire the Next

解説 各バックアップ方式のオンライン業務への影響について説明します。

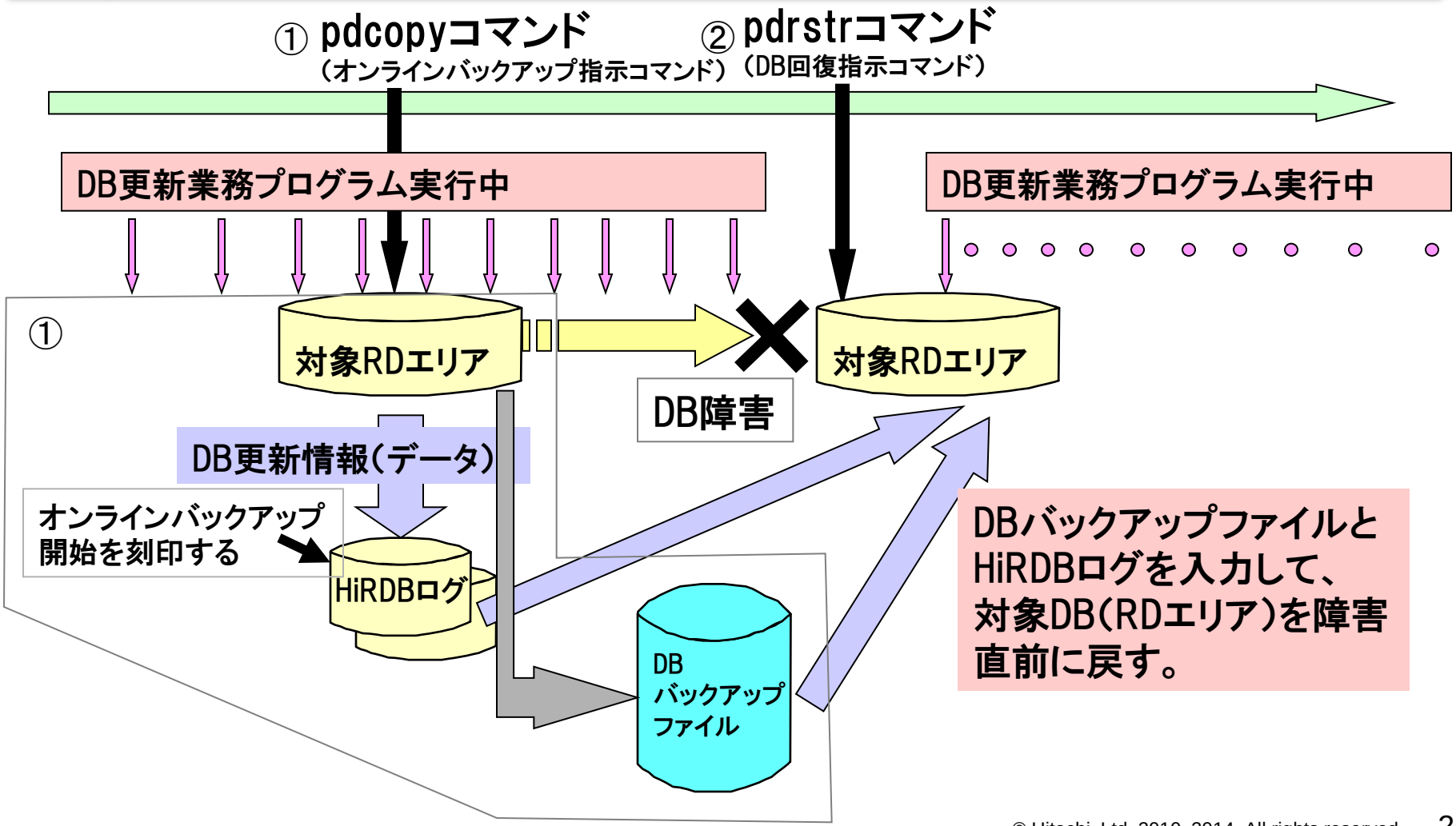


3-1-1 オンラインバックアップ＆回復

HITACHI
Inspire the Next

解説

オンラインバックアップ方式のバックアップと回復について説明します。
pdcopyコマンドの実行により、①の処理が行われます。DB障害後はpdrstrコマンドでDBバックアップファイルとHiRDBログを指定して障害直前に戻せます。

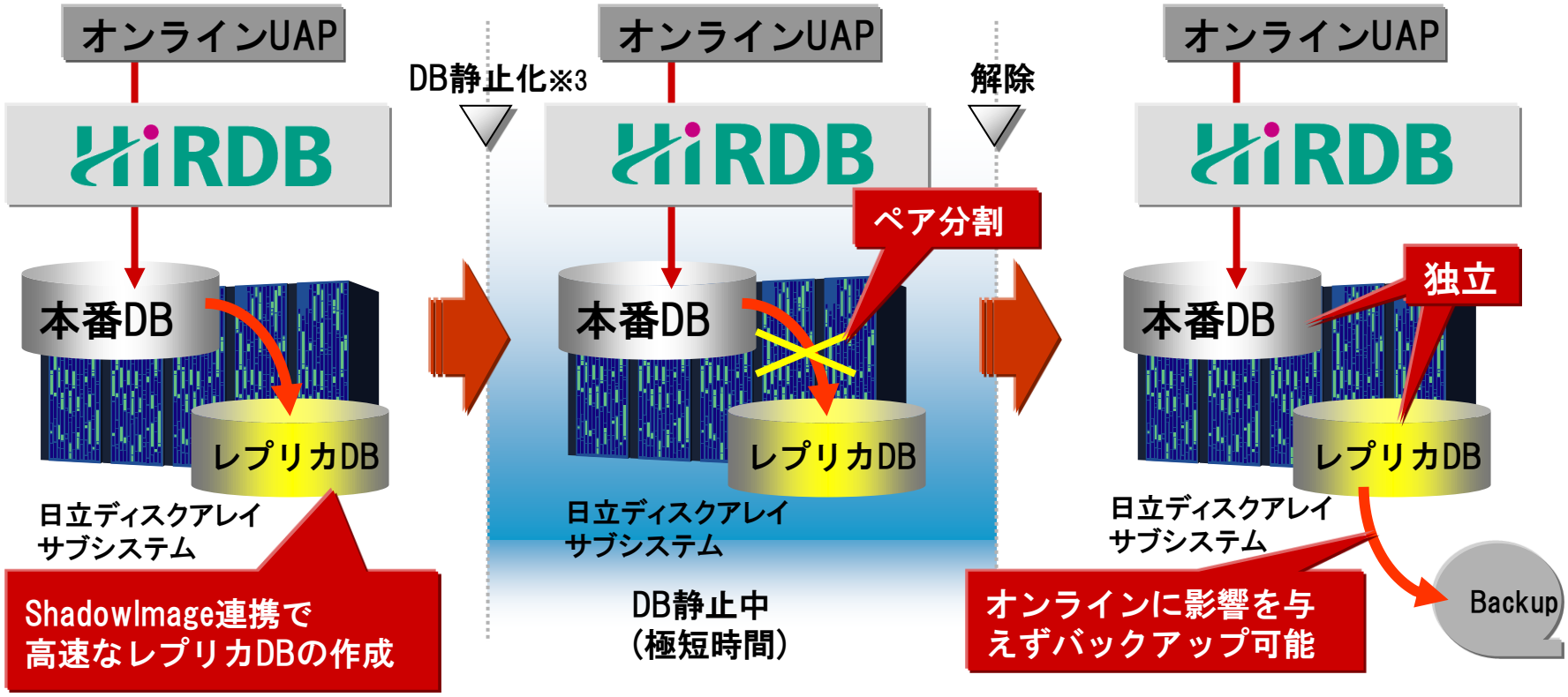


3-1-1 オンラインでの高速バックアップ

HITACHI
Inspire the Next

解説

DB静止化※1と日立ディスクアレイサブシステムのShadowImageとの連携で整合性のとれたレプリカDB※2を高速に作成できます。
整合性の取れたレプリカDBでバックアップを取得できます。
整合性の取れたレプリカDBなのでログとペアでのバックアップは不要です。



※1 バッファ内の更新情報のディスクへの書き込みとトランザクションの完結待ちを行い整合性を確保する機能

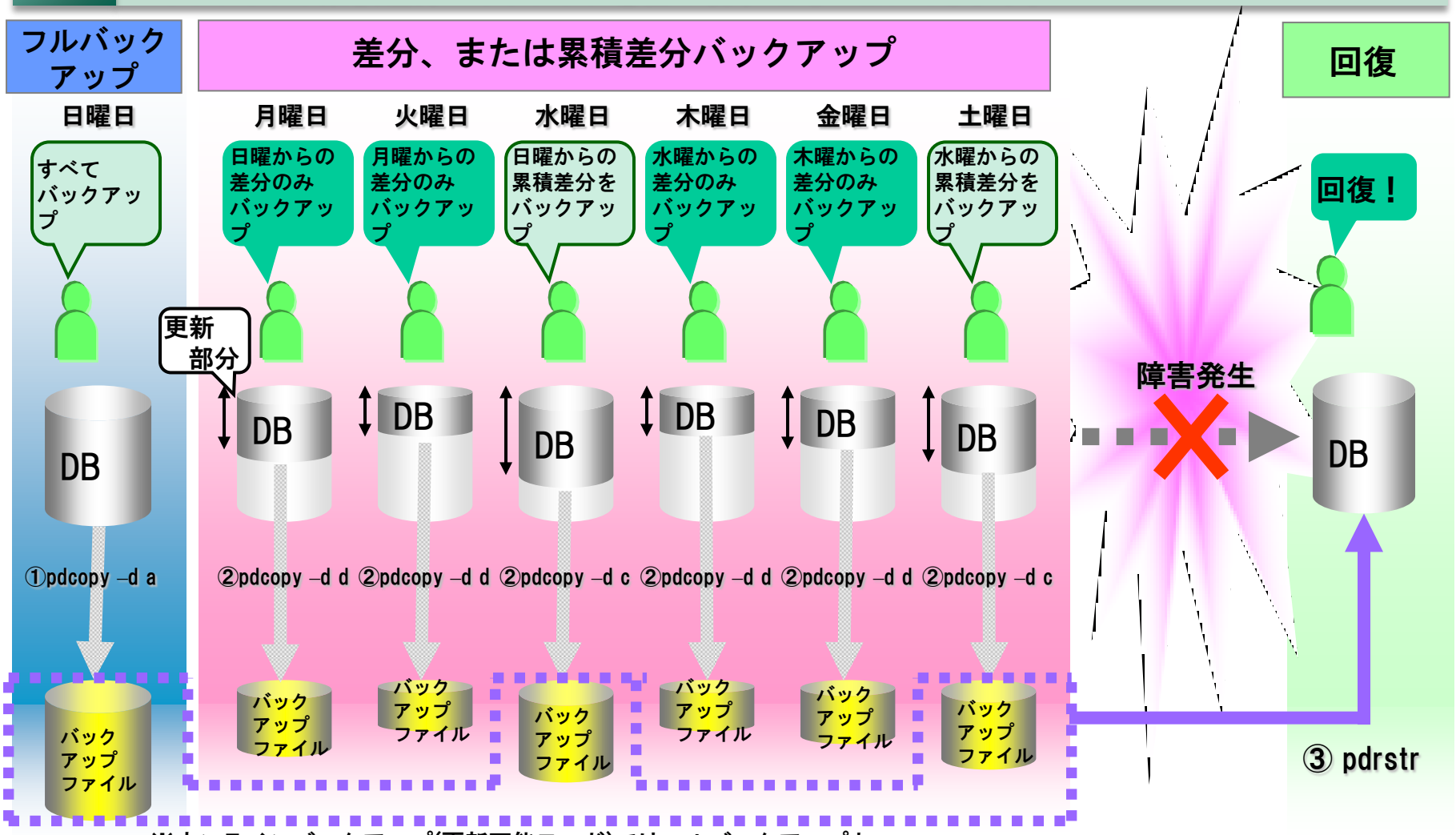
※2 レプリカDBを使う場合には、HiRDB Staticizer Option が必要です。

※3 静止化中での、更新トラン扱いを選択可能（受付可／エラー）

3-1-1 差分バックアップと回復運用

HITACHI
Inspire the Next

解説 各バックアップ方式※では、フルバックアップのほかに、更新部分のみをバックアップする差分バックアップ、累積差分バックアップを組み合わせることができます。



※オンラインバックアップ(更新可能モード)ではフルバックアップと累積差分バックアップのみ可能です。

3-1-1 バックアップとリストアの注意事項

HITACHI
Inspire the Next

解説

定期的なバックアップ以外にも、以下に示すような操作を実行した場合は、必ず実行前後でバックアップを取得してください。該当する操作を実行中あるいは実行後に障害が発生すると、システムログを使った回復ができないため、十分注意してください。

(1)ログレスモードによるDB更新

◆ログレスモード(クライアント環境変数PDDBLOG=NO)でアプリケーションを実行する場合
◆ログレスモードまたは更新前ログ取得モードで、データベース作成ユティリティ(pdload)
もしくはデータベース再編成ユティリティ(pdrorg)を実行する場合

ログレスモードとは、データ更新に関するシステムログを出力しない機能です。また、更新前ログ取得モードとは、更新後の値(ロールフォワードに必要な情報)をシステムログに記録しない機能です。大量の更新を実行するときは、システムログの出力量を削減することで、実行時間が短縮できるメリットがありますが、システムログを使った回復ができません。

(2)データベース構成変更ユティリティ(pdmod)の実行

RDエリアに対してデータベース構成変更ユティリティを実行すると、ユティリティ実行前に取得したバックアップとシステムログ(ユティリティ実行前に出力されたシステムログ＋ユティリティ実行後に出力されたシステムログ)を使用した回復ができません。

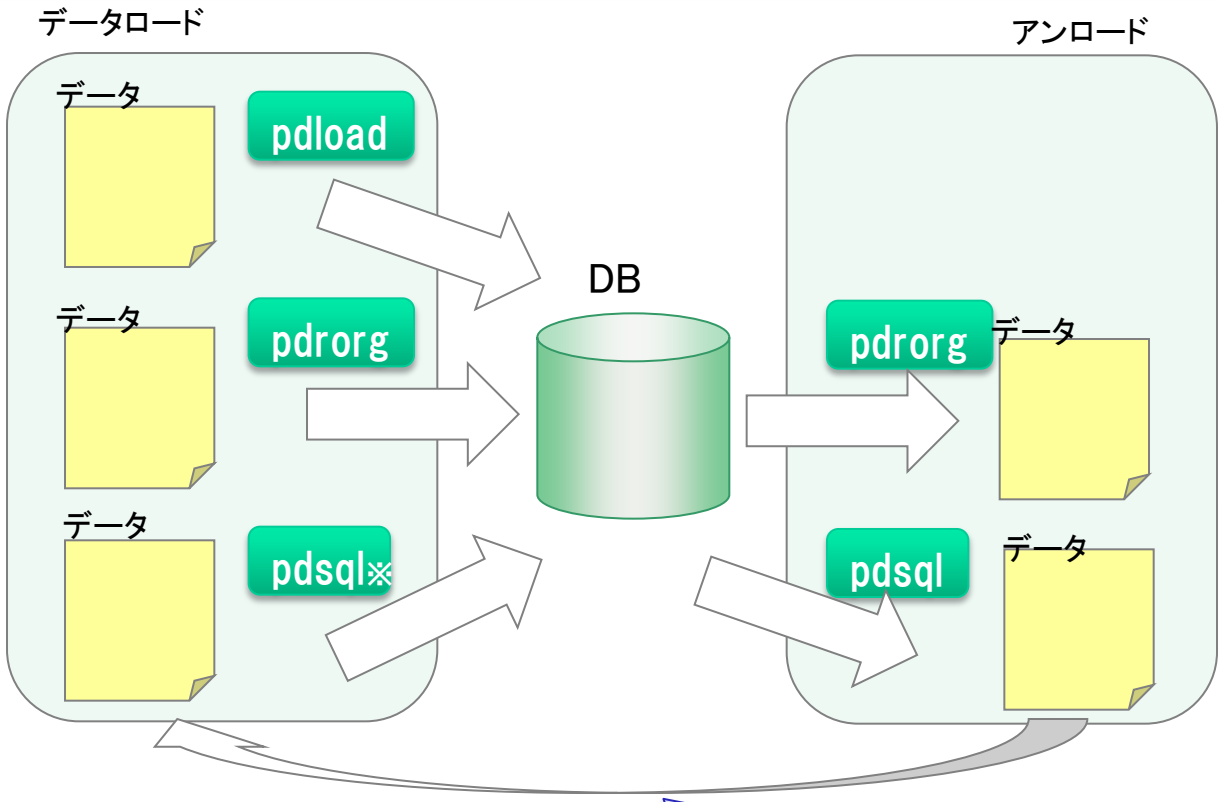
※上記以外のバックアップが必要な操作については、HiRDBマニュアル「システム運用ガイド」ー「バックアップの取得時期」を参照してください。

3-1-2 データロードとアンロード

HITACHI
Inspire the Next

解説

一括してデータを登録するときや他の表にデータを移行する場合にデータロードとアンロードを行います。pdload、pdrorg、pdsqlを使います。



データ形式
・DAT形式
CSV形式と呼ばれている方式で一般に広く使われている。
・バイナリ形式
性能面で優れている。

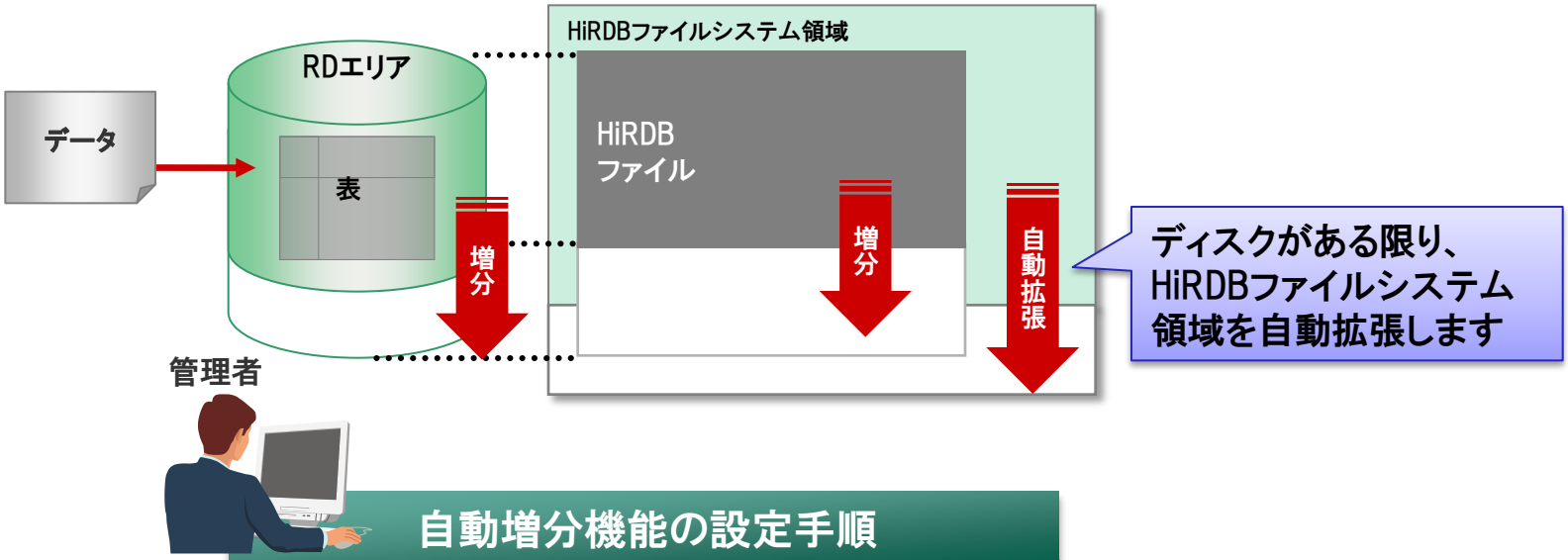
pdrorg、pdsqlでアンロードしたデータをpdloadでデータロードすることも可能

※バイナリ形式の入力をサポートしていません。

3-1-3 RDエリアの自動増分運用

解説

データが増えて、RDエリアの容量不足が発生した場合、当初確保した領域を自動的に拡張することができます。予期せぬデータの増加が想定される場合に設定してください。



自動増分機能の設定手順

①自動増分機能の使用を指定

pdfmkfsコマンドの-aオプションの指定例

```
> pdfmkfs ～ -a
```

②増分セグメント数を指定

pdinitコマンドの制御ファイル

```
create rdarea RDエリア名
  extension use 増分セグメント数 segments
  ・・・
```

3-1-4 RDエリアの再編成

HITACHI
Inspire the Next

解説

データの追加、更新、および削除を繰り返すことによって、データベース中のデータ、インデクスの配置に乱れが生じ、格納効率の低下や性能劣化を引き起こします。これを防ぐために、定期的にpdrorgコマンド(データベース再編成ユティリティ)で表の再編成を実施してください。

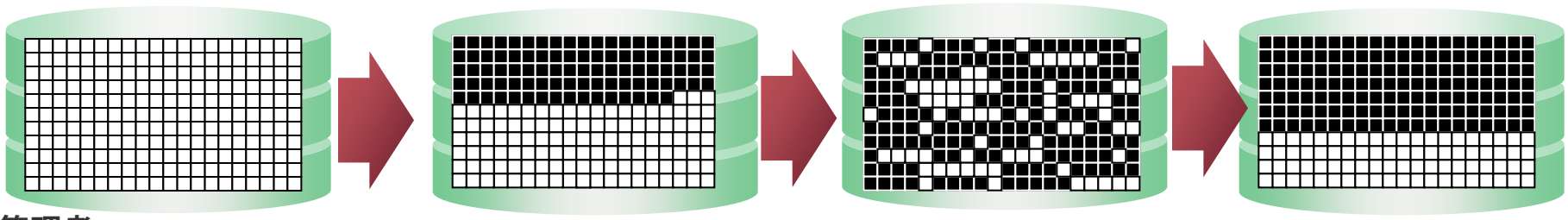
■データベースの格納状態

①初期状態
データが格納されていない状態です。

②初期データの登録
初期データの登録直後は、データが連続した状態で格納されています。

③断片化の発生
データの追加・削除を実行する業務を継続していると、データベースが断片化されます。

④データベースの再編成
データベースの再編成を実行することで、データベースの断片化を解消します。



管理者

pdrorgの使い方

```
pdrorg -k rorg -t テーブル名 ・・・
```

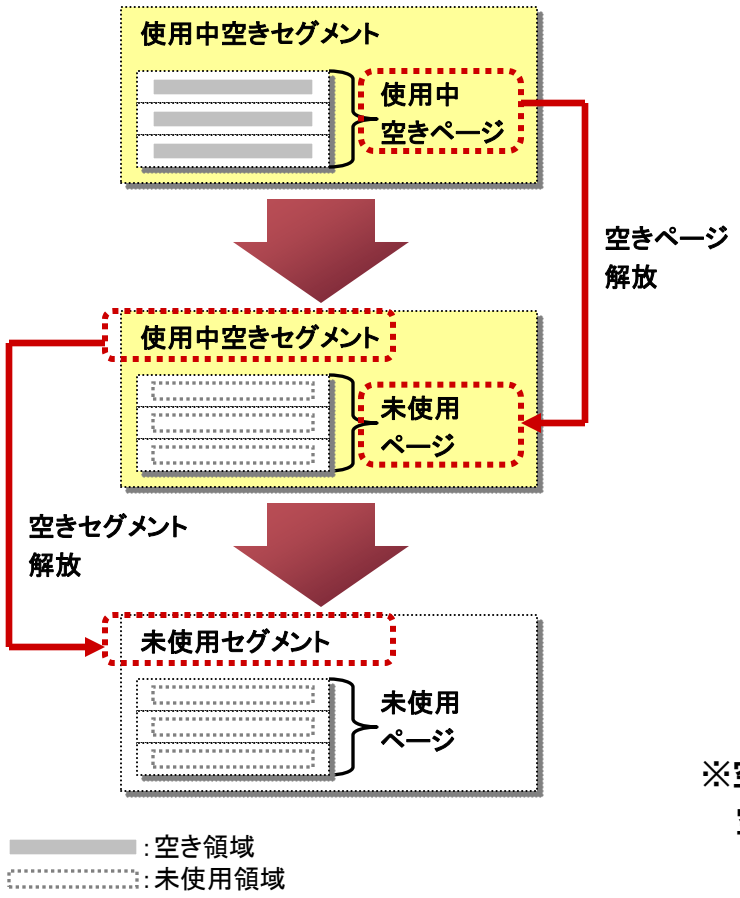
再編成サイクルを長くするために、以下を考慮してください。
・REUSE表を利用してください。
・インデクスキー値更新はしないでください。

3-1-4 空きページ解放

HITACHI
Inspire the Next

解説

空きページ解放ユティリティ(pdreclaim)を使うと、オンライン処理中に、空きページ解放と空きセグメント解放を行うことができます。



■空きセグメント解放の実行例■

(1)空きページ解放

pdreclaim
-k table　　:対象資源の種別。表の場合はtableを指定。
-t TABLE1　　:表識別子。

(2)空きセグメント解放

pdreclaim
-k table　　:対象資源の種別。表の場合はtableを指定。
-t TABLE1　　:表識別子。
-j　　:空きセグメント解放を行う場合に指定。

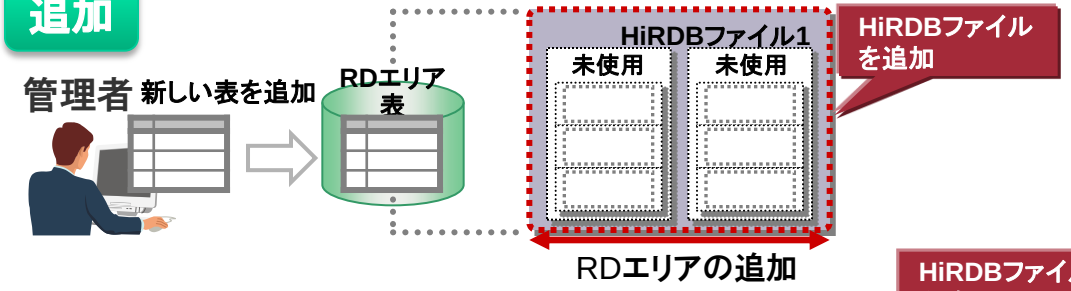
※空きページ解放処理に続けて、空きセグメント解放処理行う場合は、
空きページ解放ユティリティ(pdreclaim)に -a オプションを指定してください。

3-1-5 RDエリアの構成変更

HITACHI
Inspire the Next

解説 表の追加、データの追加、表の削除などに応じてRDエリアを追加、拡張、削除などを行えます。
RDエリアの構成変更はpdmodで行い、構成変更の詳細は制御ファイルに記載します。

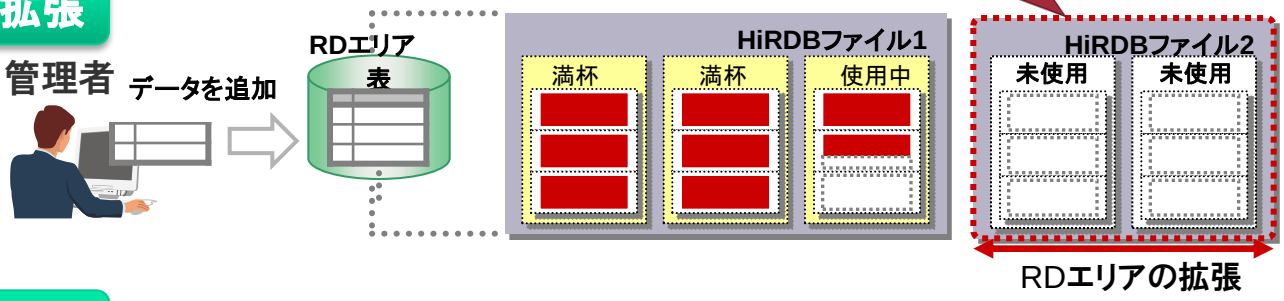
追加



追加方法
pdmod -a 制御ファイル
制御ファイル

```
create rdarea RDエリア名
  ・・・
```

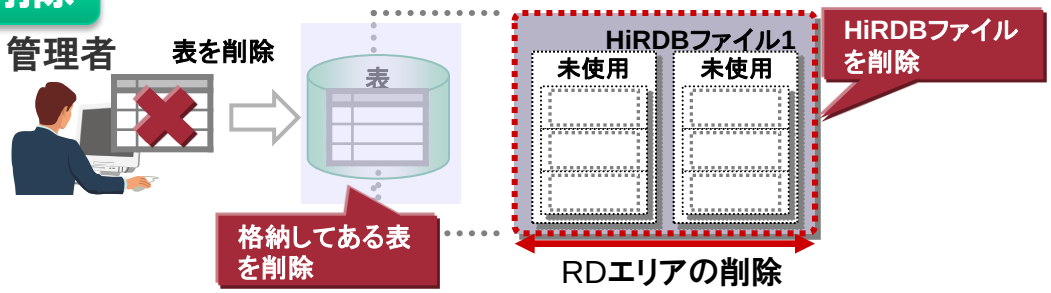
拡張



拡張方法
pdmod -a 制御ファイル
制御ファイル

```
expand rdarea  RDエリア名
  ・・・
```

削除



削除方法
pdmod -a 制御ファイル
制御ファイル

```
remove rdarea  RDエリア名
```

pdmodコマンドは稼働中に実行できます

HITACHI
Inspire the Next

HiRDB

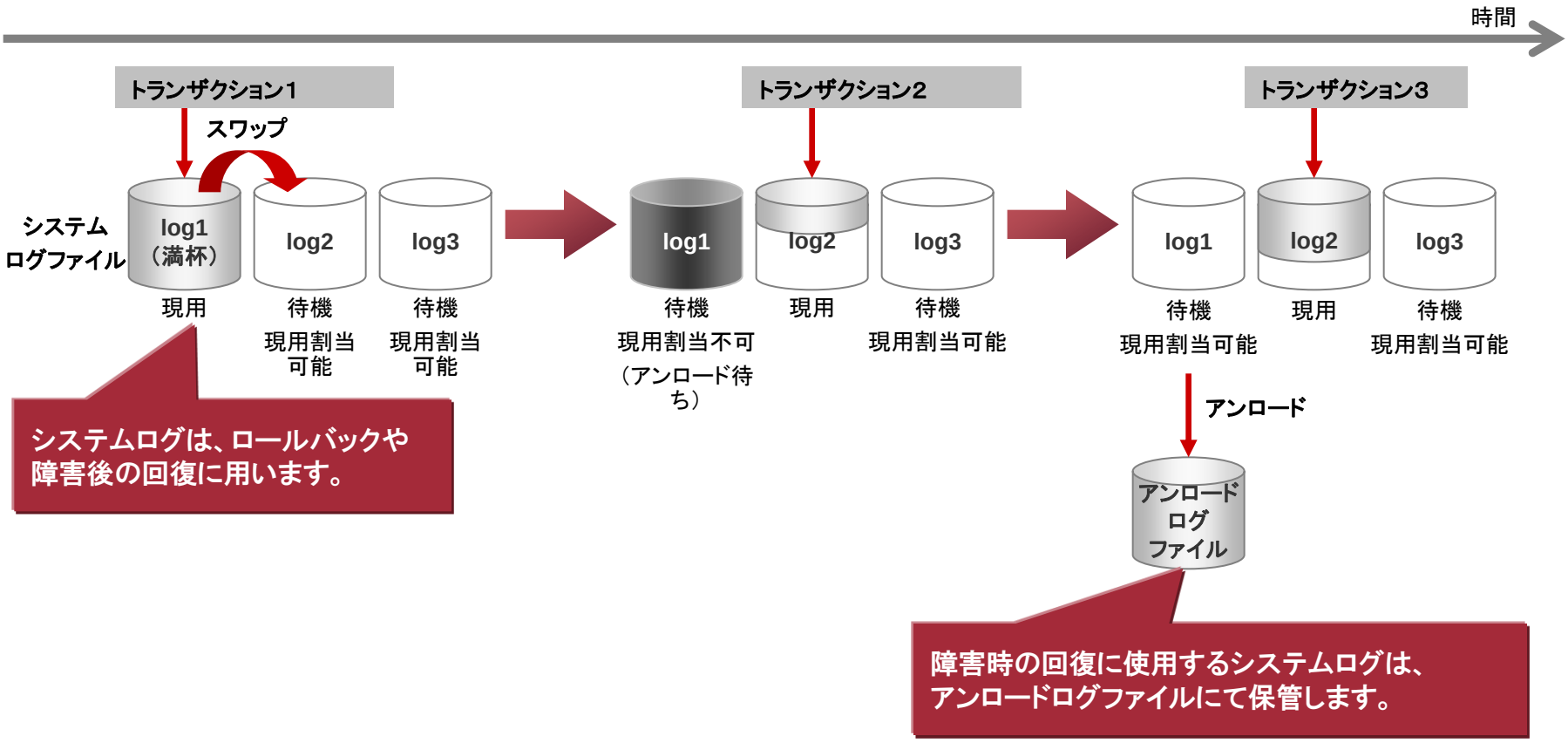
3-2. システムログ運用

3-2-1 HiRDBのシステムログの目的

HITACHI
Inspire the Next

解説

システムログはデータベースの更新履歴情報で、ロールバックや障害後の回復に用います。システムログを格納するファイルはN世代分あり、有限のファイルを使いまわしているので、古いシステムログはアンロードが必要です。（運用方法によります）

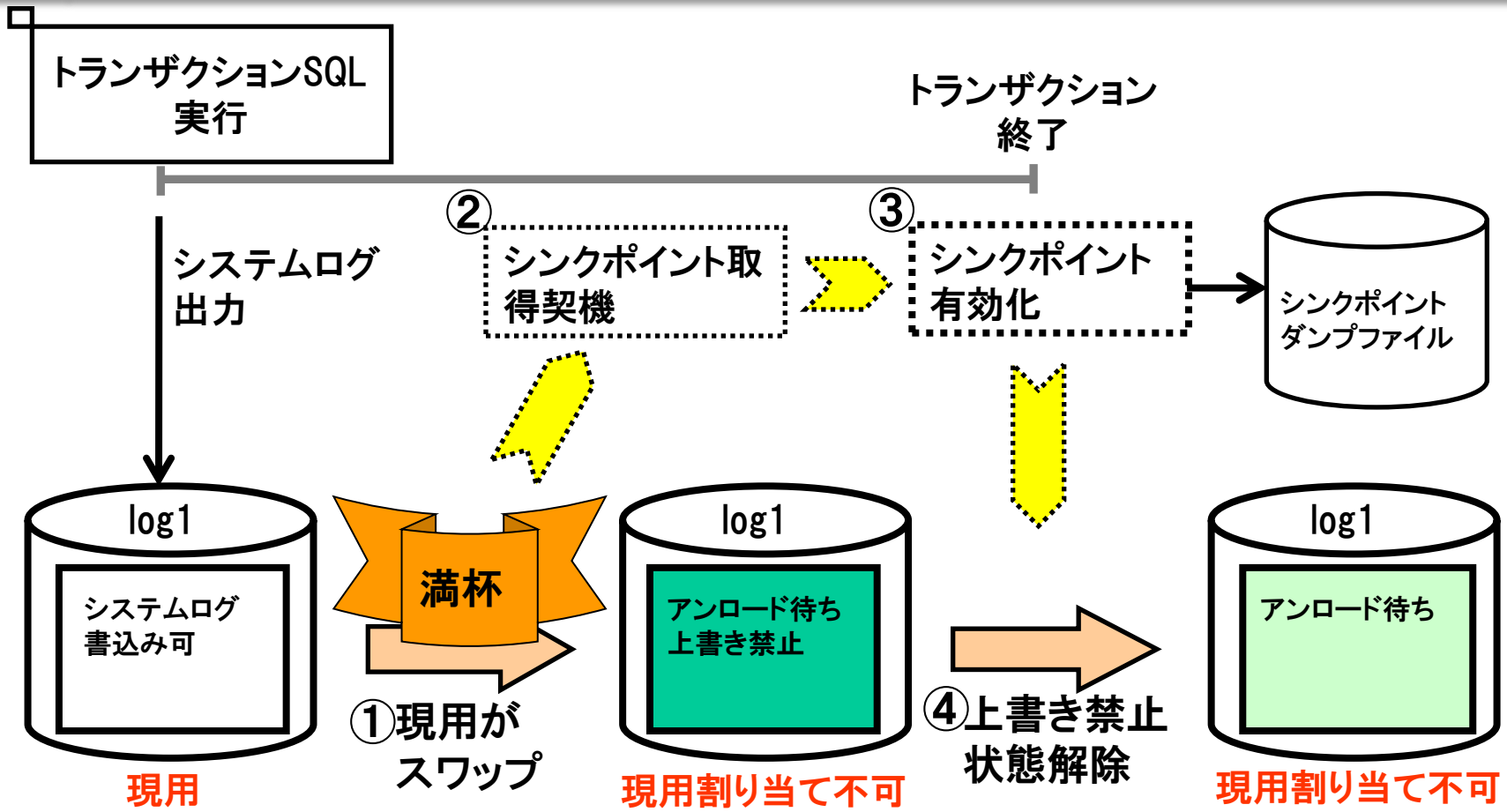


3-2-1 HiRDBのシステムログ制御　上書き禁止状態の遷移

HITACHI
Inspire the Next

解説

システムログファイルはデータベースの回復に重要なファイルですので、その状態を管理しています。
システムログlog1が満杯になるとスワップして①、アンロード待ち上書き禁止状態になります。この時点でシンクポイント取得契機が発生し②、仕掛かり中のトランザクションの待ち合わせを行います。トランザクションが終了し、シンクポイントが有効化され③、上書き禁止状態が解除されます④。

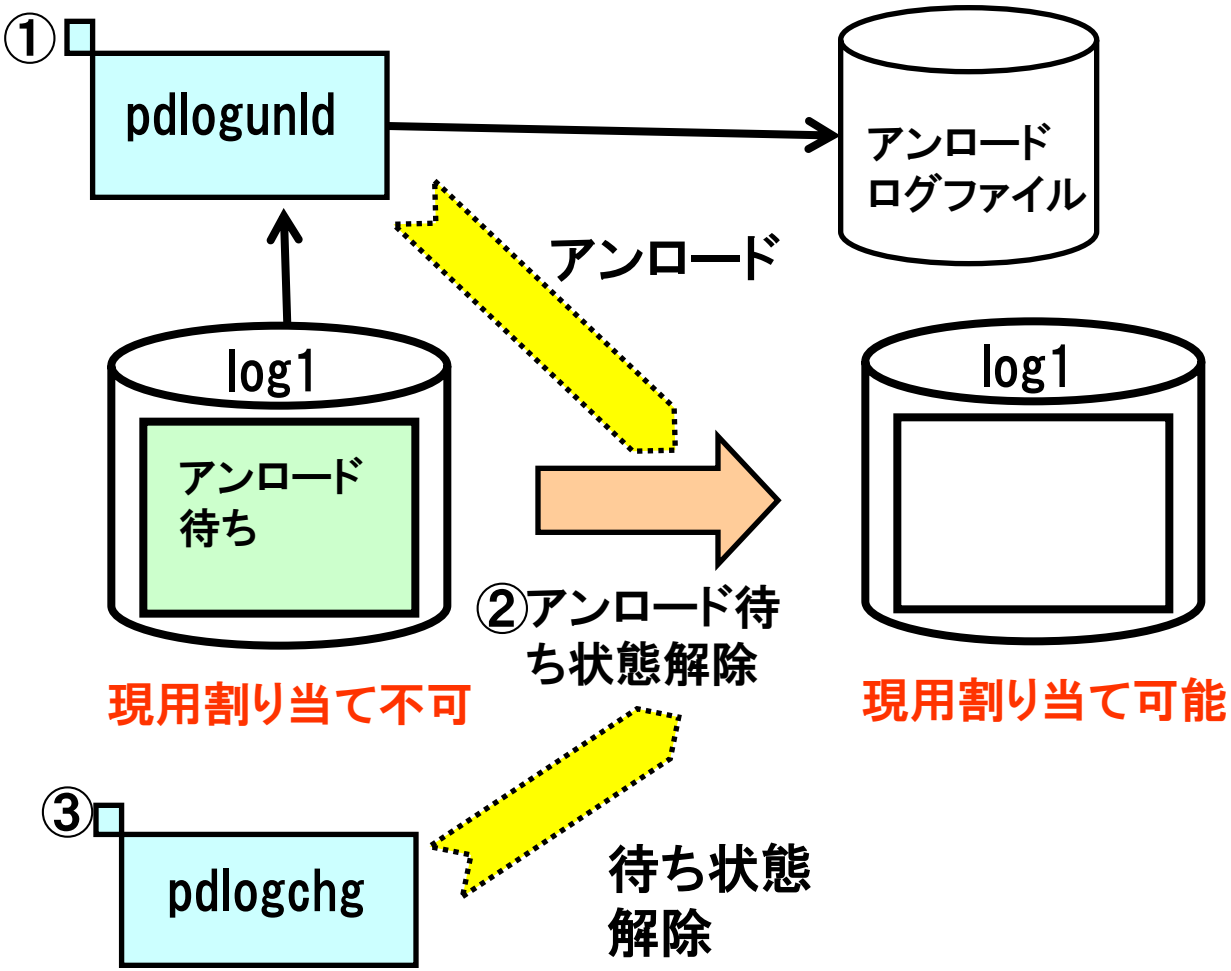


3-2-1 HiRDBのシステムログ制御　アンロード待ち状態の遷移

HITACHI
Inspire the Next

解説

アンロード待ち状態は、取得したシステムログをアンロードログファイルにアンロードしていない状態です。pdlogunldを使ってシステムログをアンロードし①、アンロード待ち状態を解除します。アンロードが不要な場合はpdlogchgを使って③、待ち状態を解除することもできます。



3-2-1 HiRDBのシステムログ運用の一覧

HITACHI
Inspire the Next

解説 システムログの運用の一覧を示します。

#	運用の種類		運用方法	データベースの回復方法
1	システムログをアンロードする運用	手動ログアンロード運用 自動ログアンロード運用	•アンロード待ち状態のファイルをアンロードします。	•バックアップおよびアンロードログファイルを入力情報にしてデータベースを回復します。 •データベースは、バックアップ取得時点、およびバックアップ取得時点以降の任意の同期点に回復できます。
2	アンロードレスシステムログ運用		•アンロード待ち状態のファイルを解放します(アンロードする必要はありません)。 •サーバ単位にバックアップを取得する必要があります。	•バックアップおよびバックアップ取得以降のシステムログを入力情報にしてデータベースを回復します。 •データベースは、バックアップ取得時点、およびバックアップ取得時点以降の任意の同期点に回復できます。
3	アンロード状態のチェックを解除する運用		•アンロード待ち状態がなくなります。したがって、システムログをアンロードする必要はありません。	•バックアップを入力情報にしてデータベースを回復します。 •データベースは、バックアップ取得時点にしか回復できません。

3-2-1 自動ログアンロード運用　システムログのアンロード

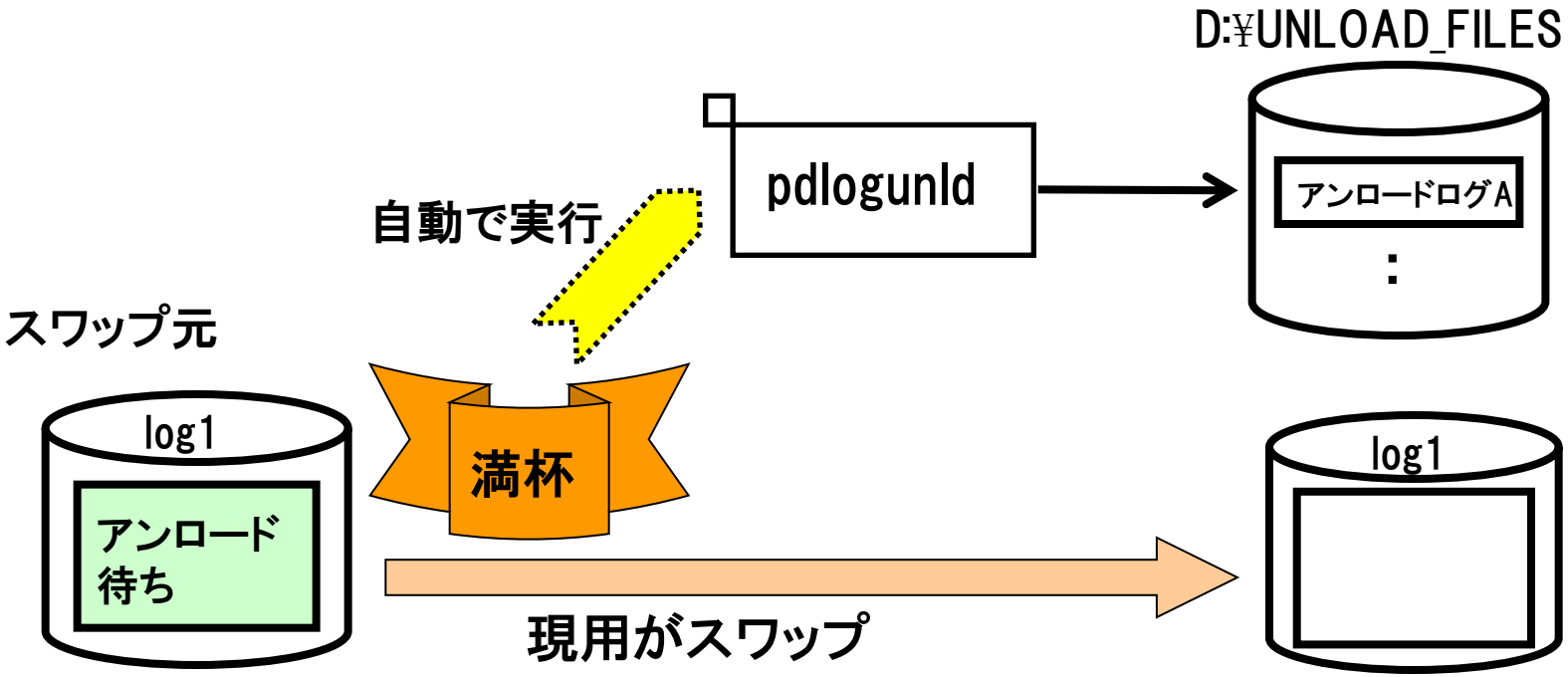
HITACHI
Inspire the Next

解説

アンロード待ちのシステムログファイルを自動的にアンロードする運用について説明します。自動ログアンロード機能を利用することで、システムログファイルが満杯になった際に自動的にシステムログをアンロードすることができます。

HiRDBサーバ定義

```
set pd_log_auto_unload_path = D:¥UNLOAD_FILES
```

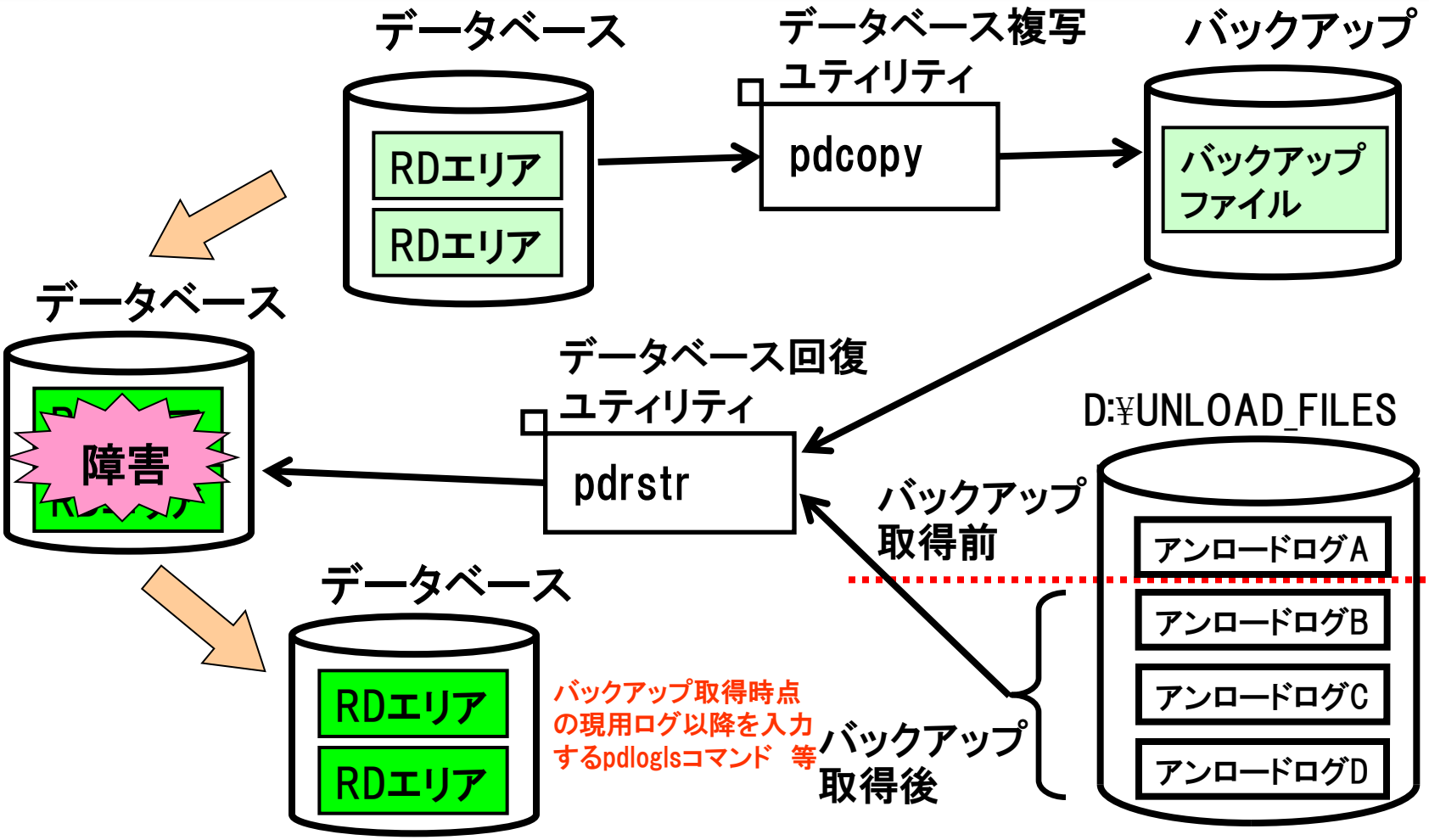


3-2-1 自動ログアンロード運用時のデータベースの回復

HITACHI
Inspire the Next

解説

自動ログアンロード運用時のデータベースの回復では、障害が起きた場合、取得済みのバックアップとアンロードログをpdrstrの入力にしてデータベースを回復します。

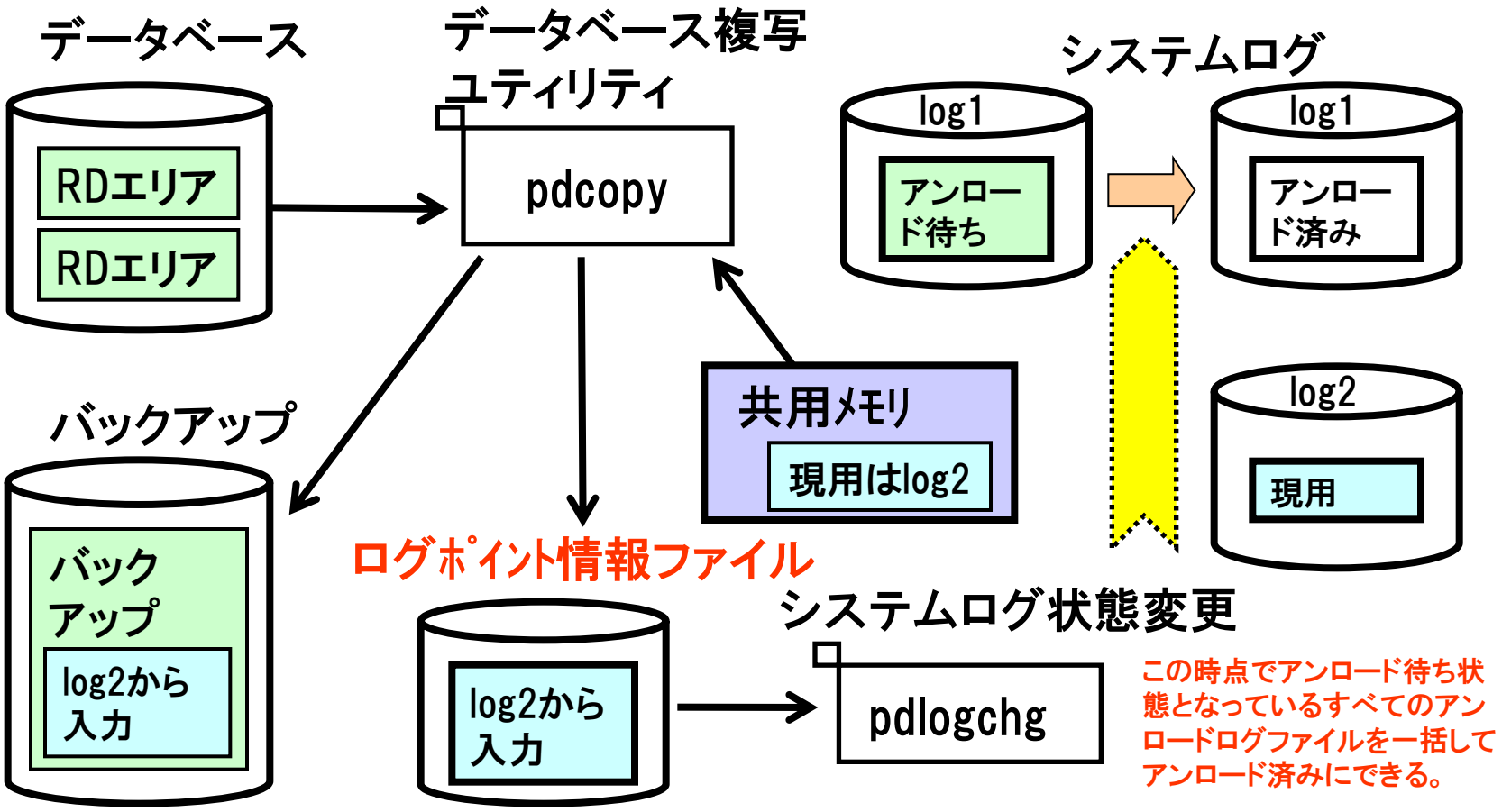


3-2-1 アンロードレスシステムログ運用

HITACHI
Inspire the Next

解説 アンロードレスシステムログ運用では、システムログをアンロードしません。pdcopyでバックアップを取得し、データベースの回復に必要なシステムログの位置を記録します。回復に不要なシステムログはpdlogchgでアンロード済み状態にします。

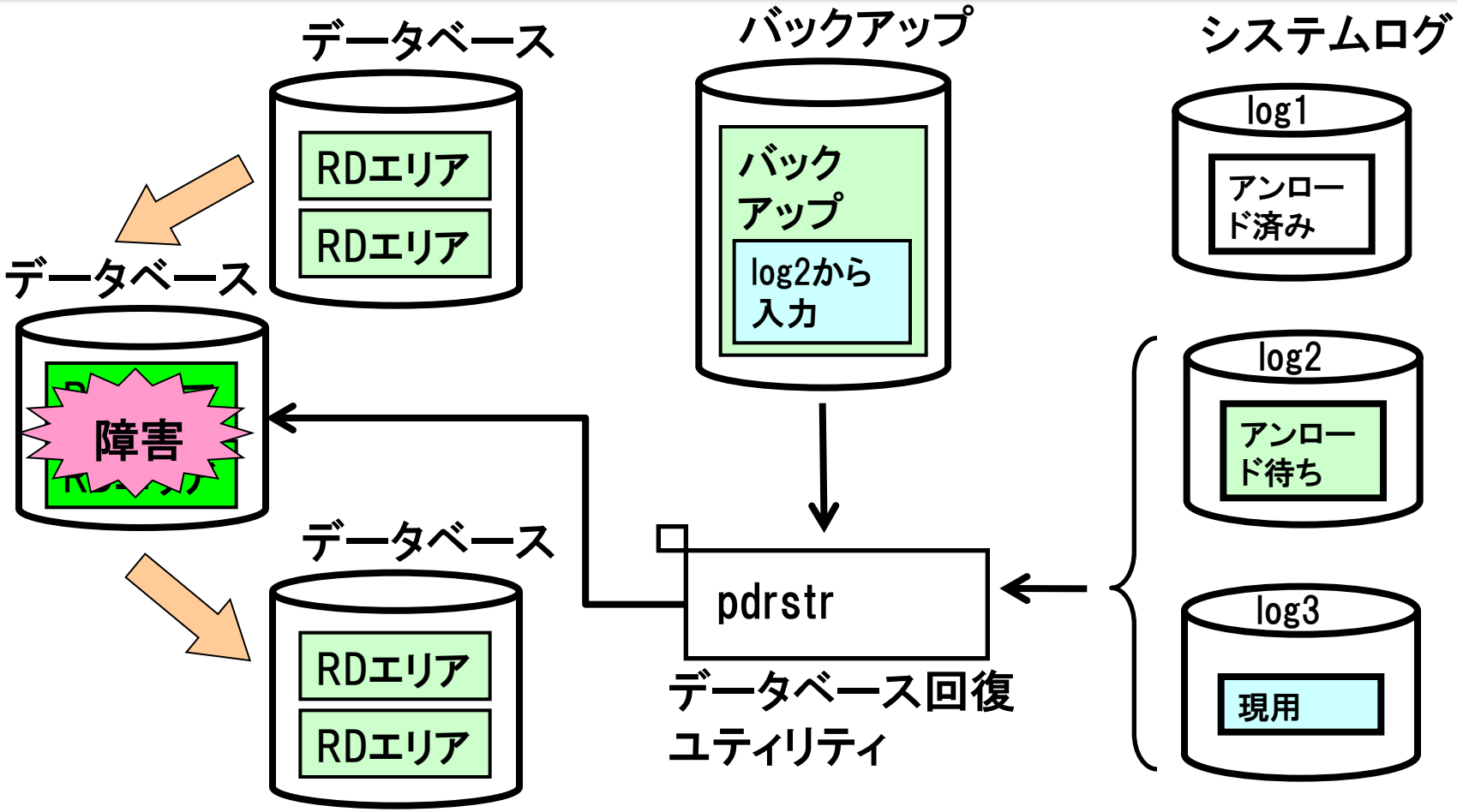
アンロード待ち状態の解除



3-2-1 アンロードレスシステムログ運用時のデータベースの回復

HITACHI
Inspire the Next

解説 アンロードレスシステムログ運用時に、障害が起きた場合は、バックアップとシステムログファイルをpdrstrの入力としてデータベースを回復します。



HITACHI
Inspire the Next

HiRDB

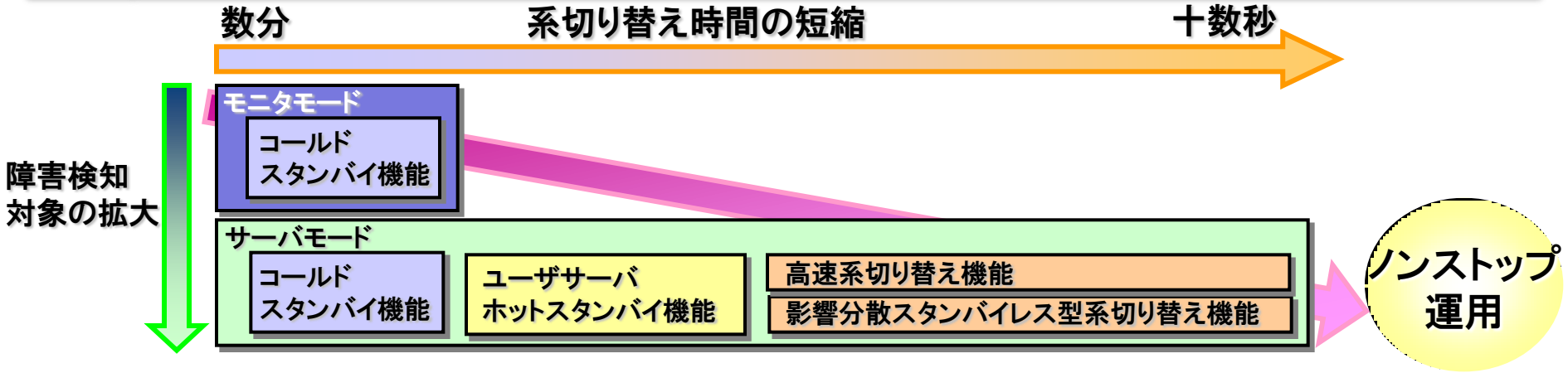
3-3. 系切り替え

3-3-1 系切り替え機能ごとの障害検知と系切り替え時間

HITACHI
Inspire the Next

解説

系切り替え機能を適用することで、万一の時でも自動かつ短時間で業務が再開でき、安定したオンラインサービス性能を維持できます。
系切り替えに要する時間は障害検知対象と系切り替え機能の組合せによって異なっており、それらの関係を以下にまとめます。



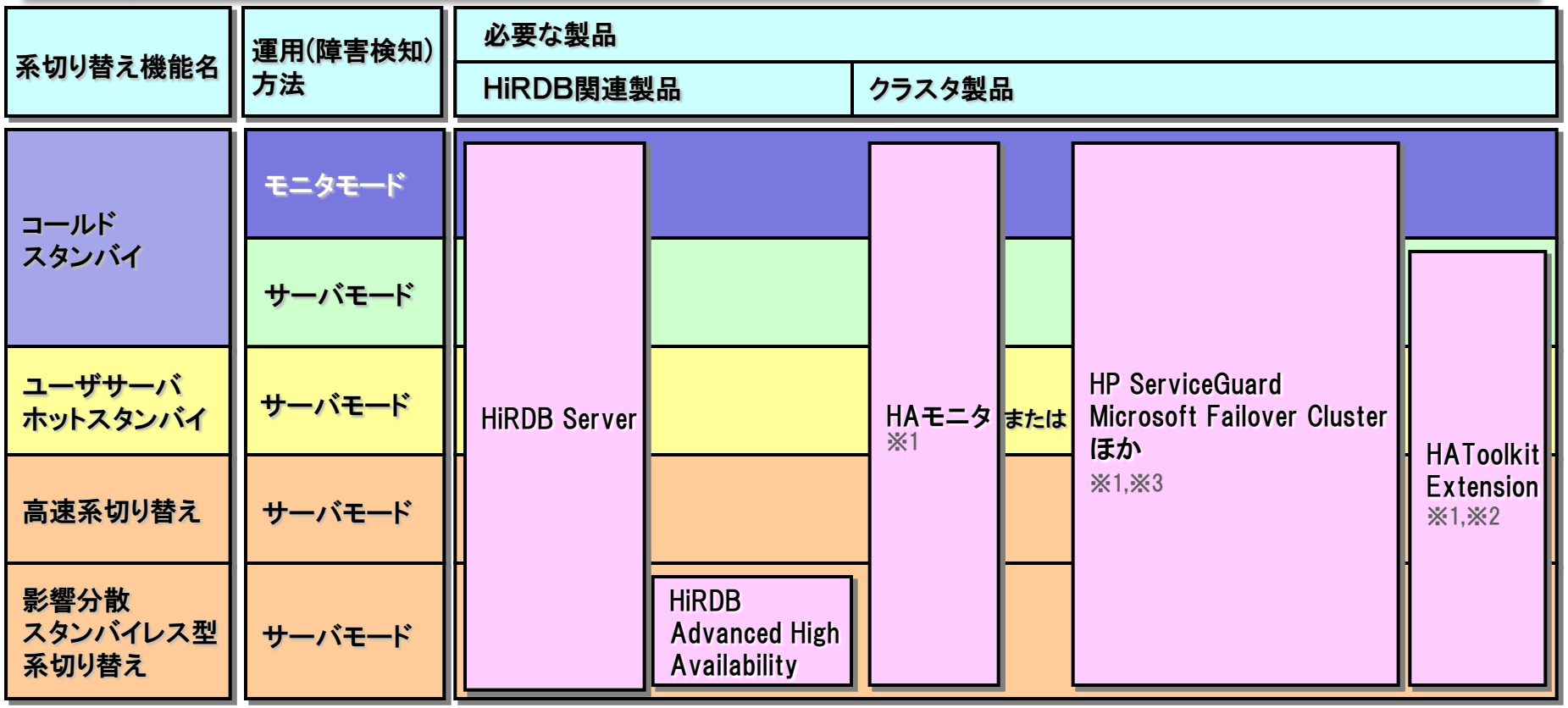
系切り替え機能名		運用(障害検知)方法			系切り替えに要する時間
			系障害	サーバ障害	
スタンバイ型系切り替え	コールドスタンバイ	モニターモード	検知できる	検知できない	数分～
		サーバモード	検知できる	検知できる	数分～
	ユーザサーバホットスタンバイ	サーバモード	検知できる	検知できる	数十秒 ～ 数分
	高速系切り替え	サーバモード	検知できる	検知できる	数秒 ～ 十数秒
スタンバイレス型系切り替え	影響分散スタンバイレス型系切り替え	サーバモード	検知できる	検知できる	数秒 ～ 十数秒

3-3-1 系切り替え機能の種類と特徴

HITACHI
Inspire the Next

解説

系切り替え機能を適用するためには、クラスタ製品が必要です。さらにスタンバイレス型系切り替えの場合はHiRDB Advanced High Availabilityが必要です。HAモニタ以外のクラスタ製品でサーバモードを適用するためには、HAToolkit Extensionが必要です。



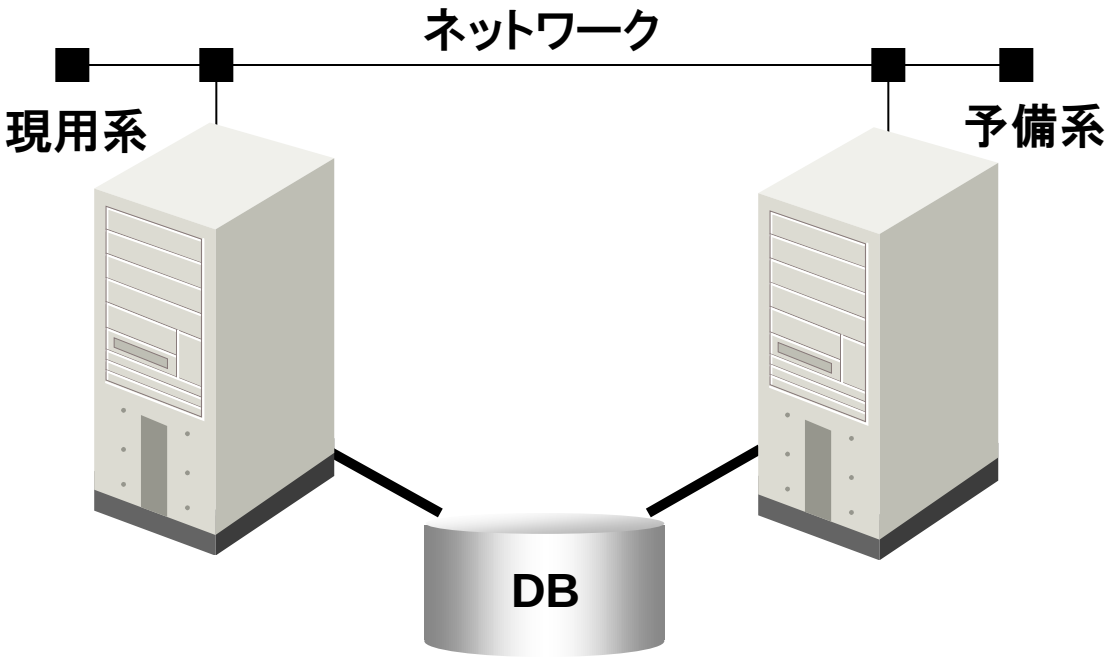
※1 対応するプラットホーム、バージョンについては、マニュアル等で確認してください。
※2 HAモニタ以外のクラスタ製品を用いてサーバモードを適用する場合に必要になります。
※3 HAToolkit Extensionが対応していないクラスタ製品では、サーバモードを適用できません。

3-3-2 HAクラスタ構成

HITACHI
Inspire the Next

解説 HiRDBの系切り替えは、障害が発生すると自動的に系を切り替えるフェイルオーバー型です。

HA(High Availability)クラスタ
フェイルオーバー型(共有ディスク)

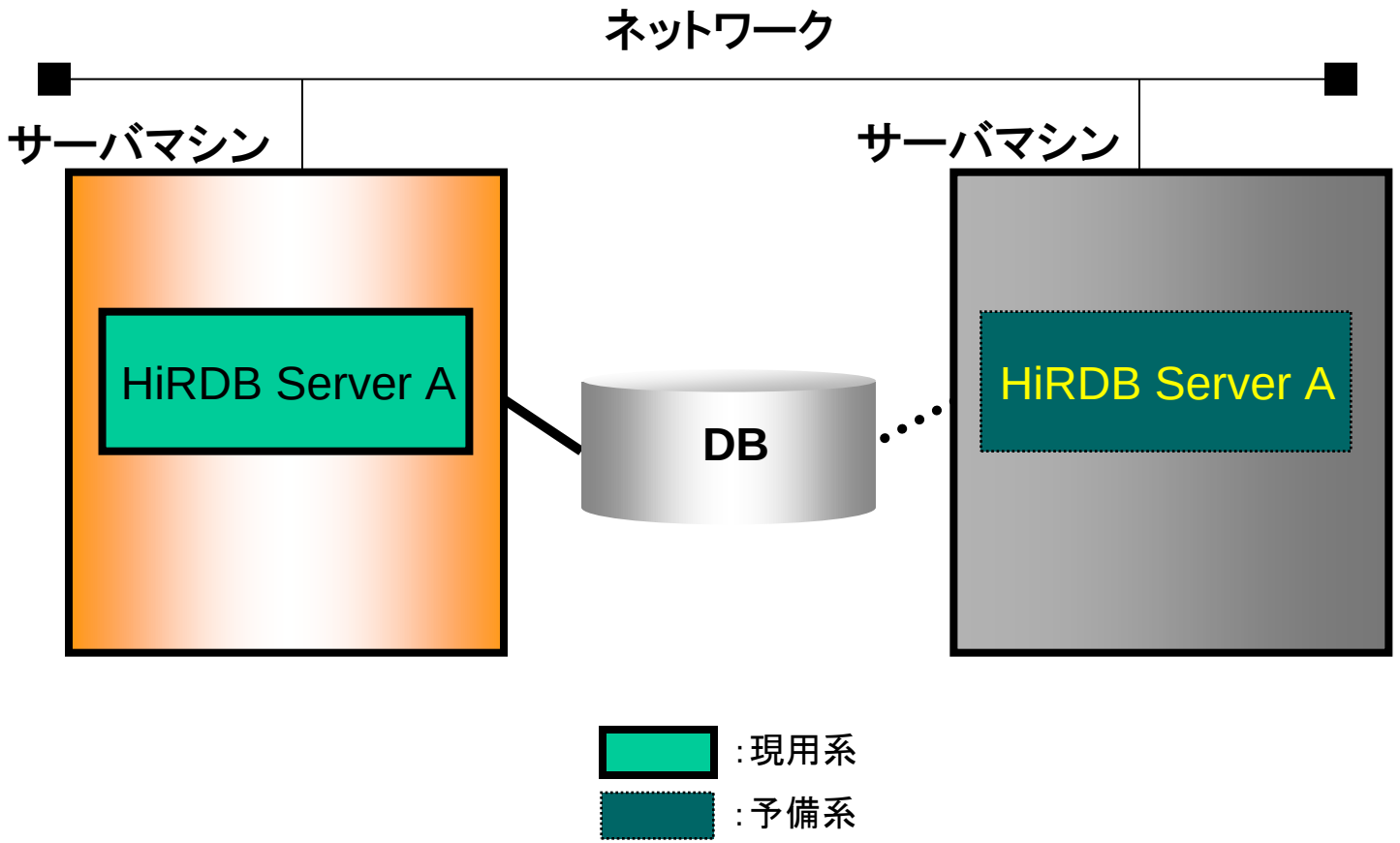


3-3-2 HAクラスタ構成：完全待機

HITACHI
Inspire the Next

解説

HAクラスタ構成の一つとして、2台のサーバを用意し、予備系の1台は稼働させずに完全に待機させておく構成があります。

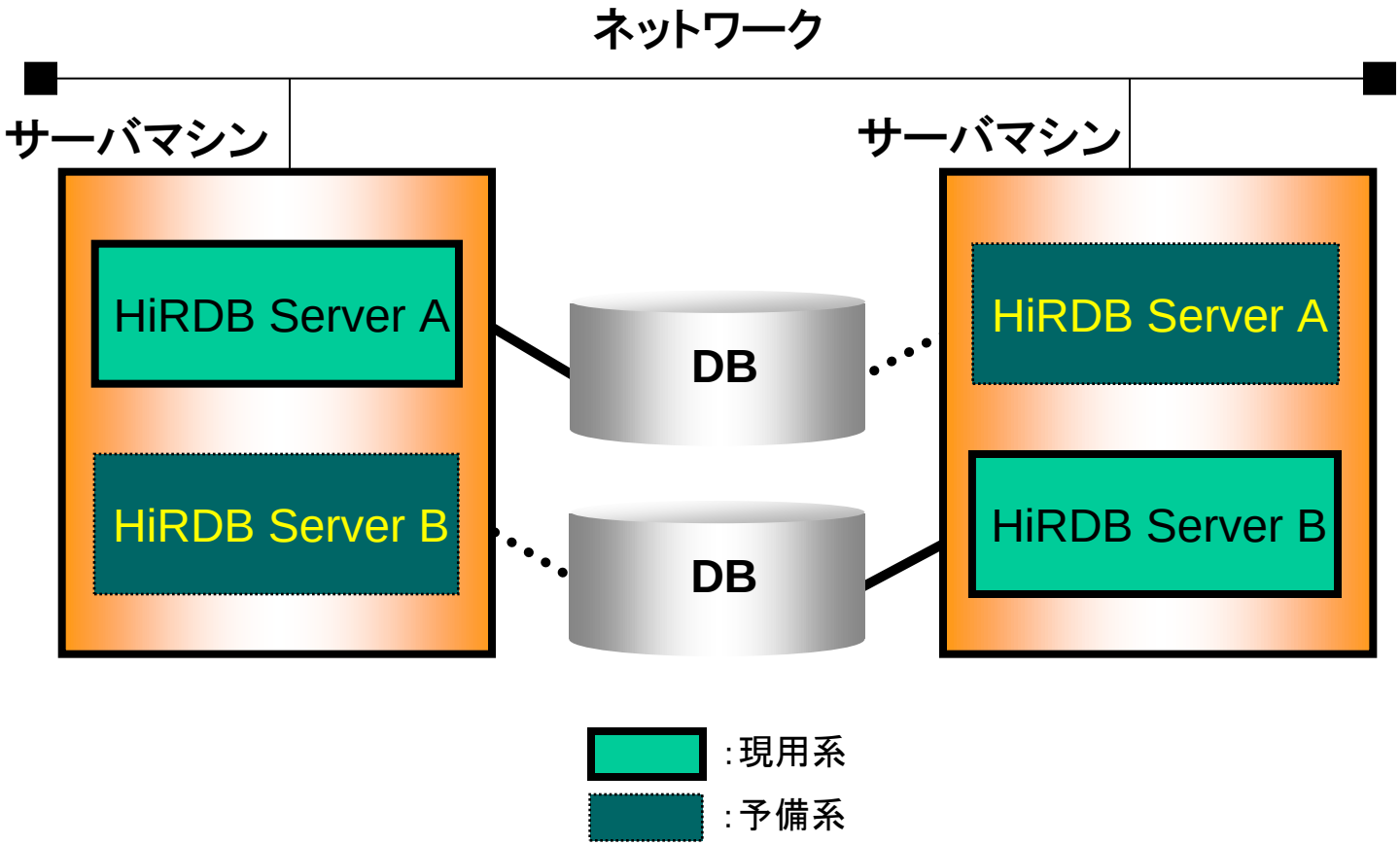


3-3-2 HAクラスタ構成：相互待機

HITACHI
Inspire the Next

解説

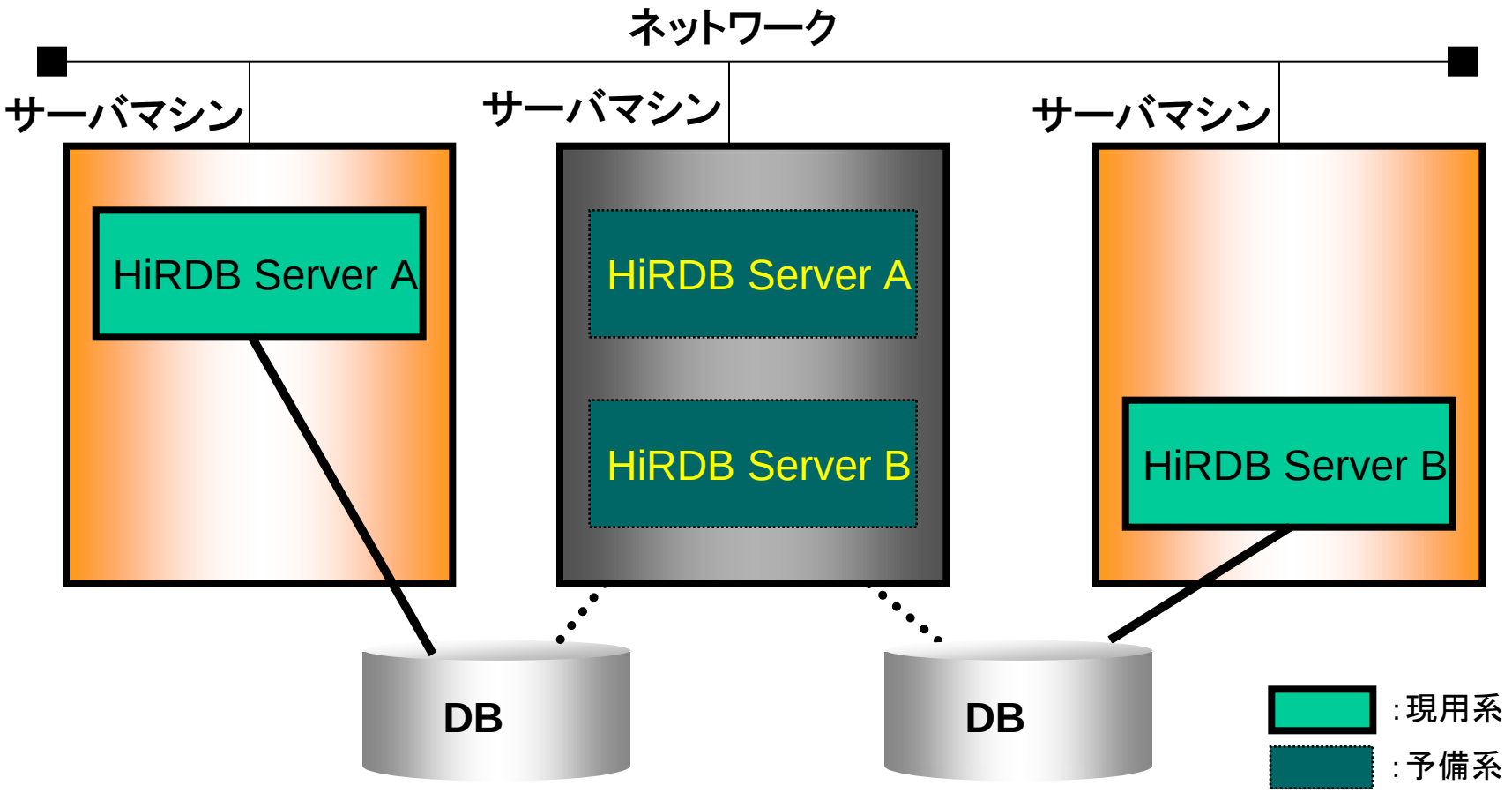
HAクラスタ構成の一つとして、2台のサーバを用意し、2台とも別のデータベースを稼働させ、相互に現用系と予備系にする構成があります。



3-3-2 HAクラスタ構成：n対1待機

HITACHI
Inspire the Next

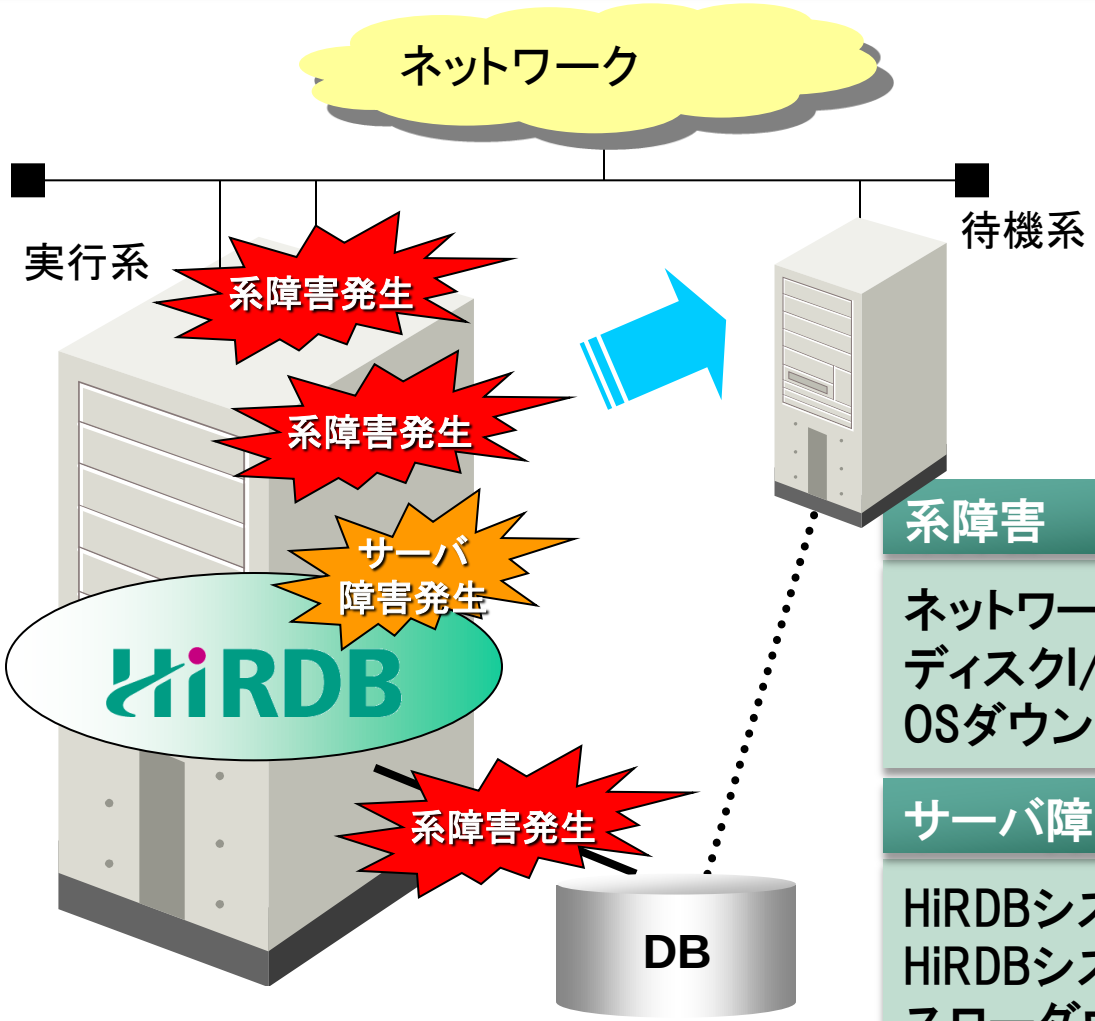
解説 HAクラスタ構成の一つとして、3台のサーバを用意し、2台を現用系として、1台を共通の予備系にする構成があります。



3-3-3 切り替えの運用（障害検知）方法

HITACHI
Inspire the Next

解説 切り替え対象となる障害には、系障害とサーバ障害があります。



系障害

ネットワーク、CPU、メモリ、
ディスクI/O障害等による
OSダウン・ハングアップ

サーバ障害

HiRDBシステム内部障害による
HiRDBシステム停止
スローダウン

3-3-3 系切り替えの運用（障害検知）方法：モードの種類

HITACHI Inspire the Next

解説 系切り替えの運用(障害検知)方法には、モニターモードとサーバモードの2つがあります。

●モニターモード

サーバマシンの障害のみによる系切り替え

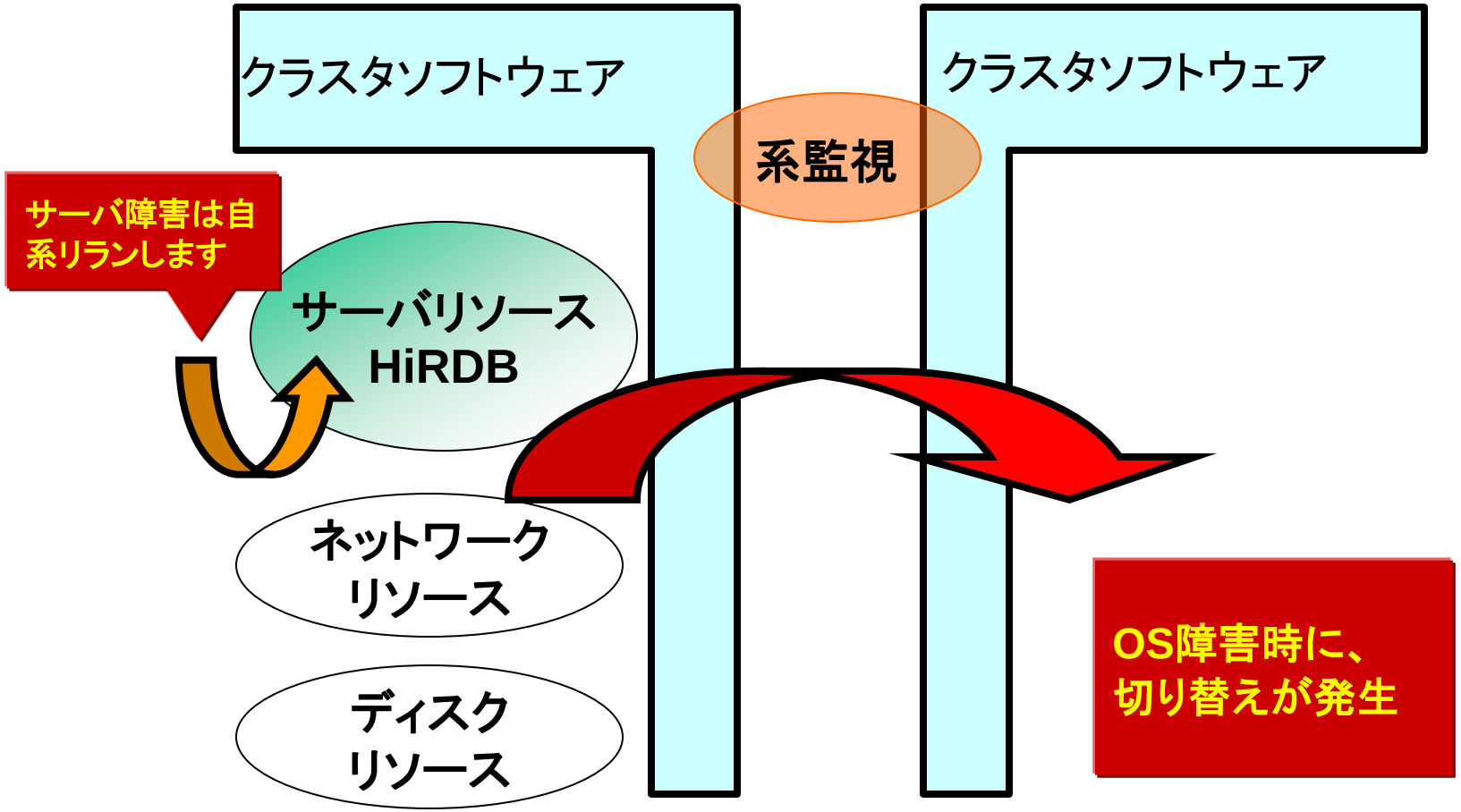
●サーバモード

サーバマシンの障害および、HiRDBプロセスの内部障害による
HiRDBスローダウン、HiRDBユニット停止時の系切り替え
（HAモニタ、HAToolkit使用時のみ）

3-3-3 系切り替えの運用（障害検知）方法：モニターモード

HITACHI
Inspire the Next

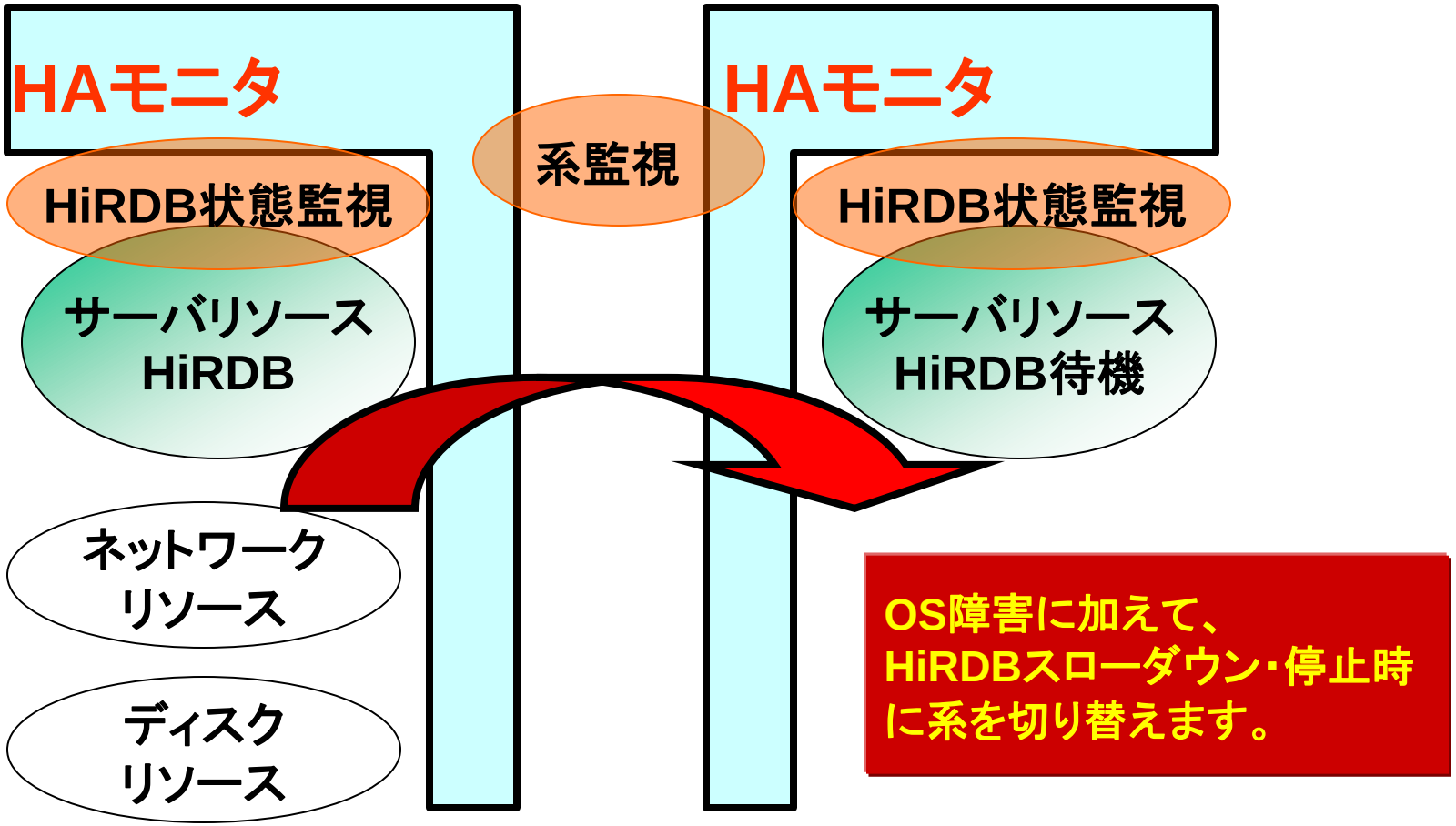
解説 モニターモードでは、OS障害など系障害を検知すると系を切り替えます。
サーバ障害は系切り替えを行わず、自系でリランします。



3-3-3 系切り替えの運用（障害検知）方法：サーバモード

HITACHI
Inspire the Next

解説 サーバモードでは、系障害に加えて、サーバ障害を検知すると系を切り替えます。



3-3-4 HiRDB系切り替え機能

HITACHI
Inspire the Next

解説 系切り替え機能は大きく分けて、スタンバイ型とスタンバイレス型の2つに分類できます。

●スタンバイ型系切り替え機能

業務処理中のHiRDBのほかに待機用のHiRDBを準備して、業務処理中のサーバマシンまたはHiRDBに障害が発生した場合、待機用のHiRDBに業務処理を自動的に切り替えます。この分類ではさらに以下の3つの機能に分けられます。

- ✓コールドスタンバイ
- ✓ユーザサーバホットスタンバイ
- ✓高速系切り替え

●スタンバイレス型系切り替え機能※

障害が発生した場合は待機系ユニットに系を切り替えるのではなく、ほかのユニットに系を切り替えて稼働中のバックエンドサーバに処理を代行させます。この分類ではさらに以下の機能に分けられます。

- ✓影響分散スタンバイレス型系切り替え

※ HiRDB Advanced High Availabilityが必要

3-3-5 スタンバイ型系切り替え

HITACHI
Inspire the Next

解説 スタンバイ型の系切り替えとして3つあります。

●コールドスタンバイ

切り替え発生後にHiRDBを待機側で起動します。

●ユーザサーバホットスタンバイ

HiRDBのユーザサーバプロセスを事前に起動しておくことで、コールドスタンバイに比べ、切り替え時間を短縮します。(HAモニタ使用時のみ)

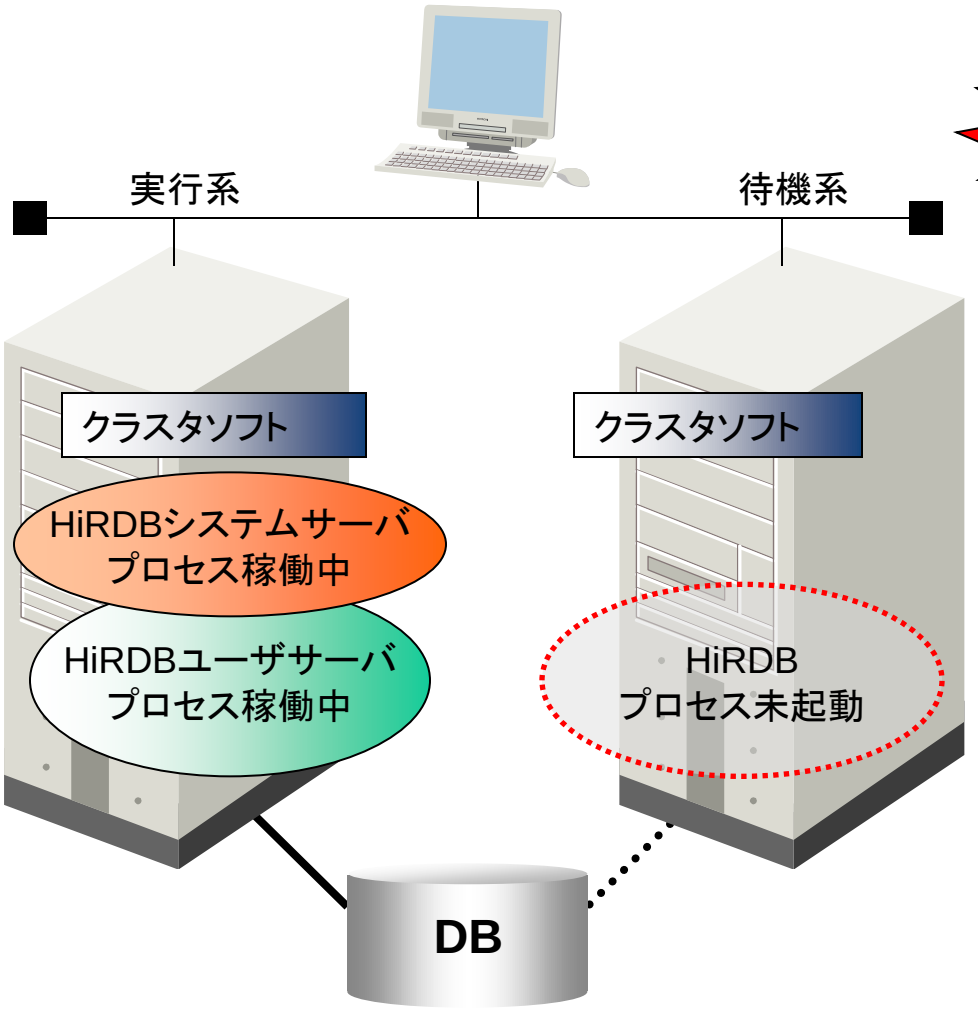
●高速系切り替え

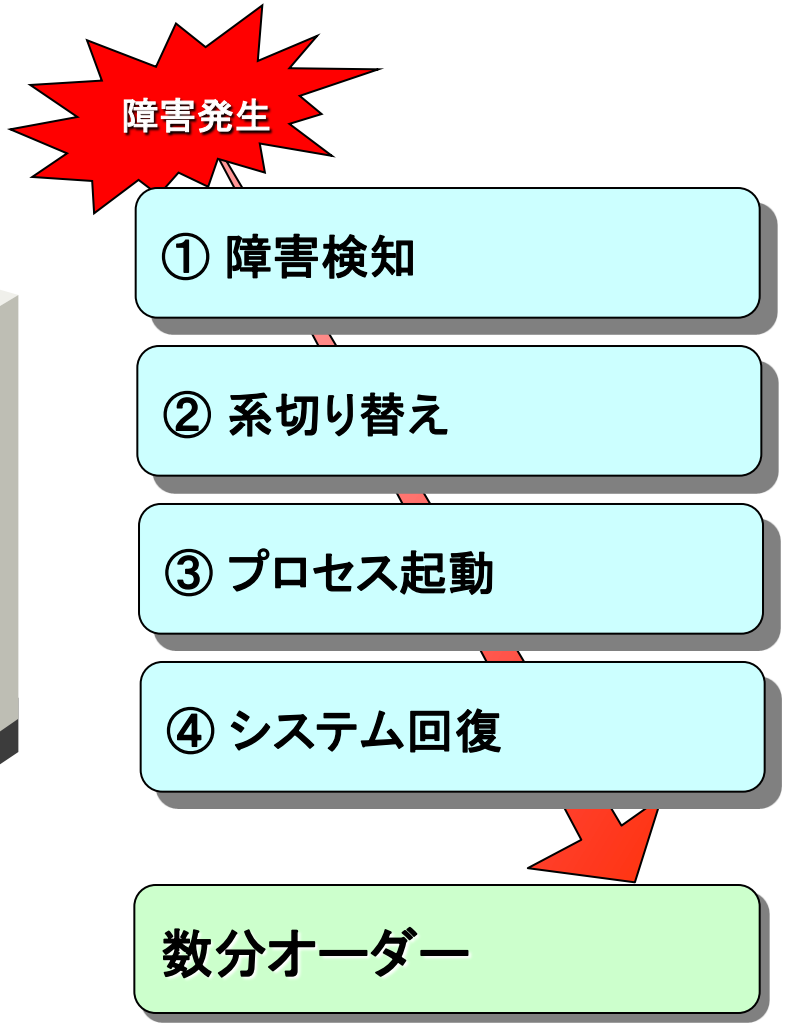
ユーザサーバプロセスに加え、HiRDBのシステムプロセスも含めた全プロセスを事前に起動し、さらに、HiRDB開始処理の一部を事前に準備しておくことで、切り替え時間を十数秒オーダーにします。(HAモニタ使用時のみ)

3-3-5 スタンバイ型系切り替え：コールドスタンバイ

HITACHI
Inspire the Next

解説 コールドスタンバイでは、HiRDBプロセスは未起動で、系切り替えに数分オーダー掛かります。

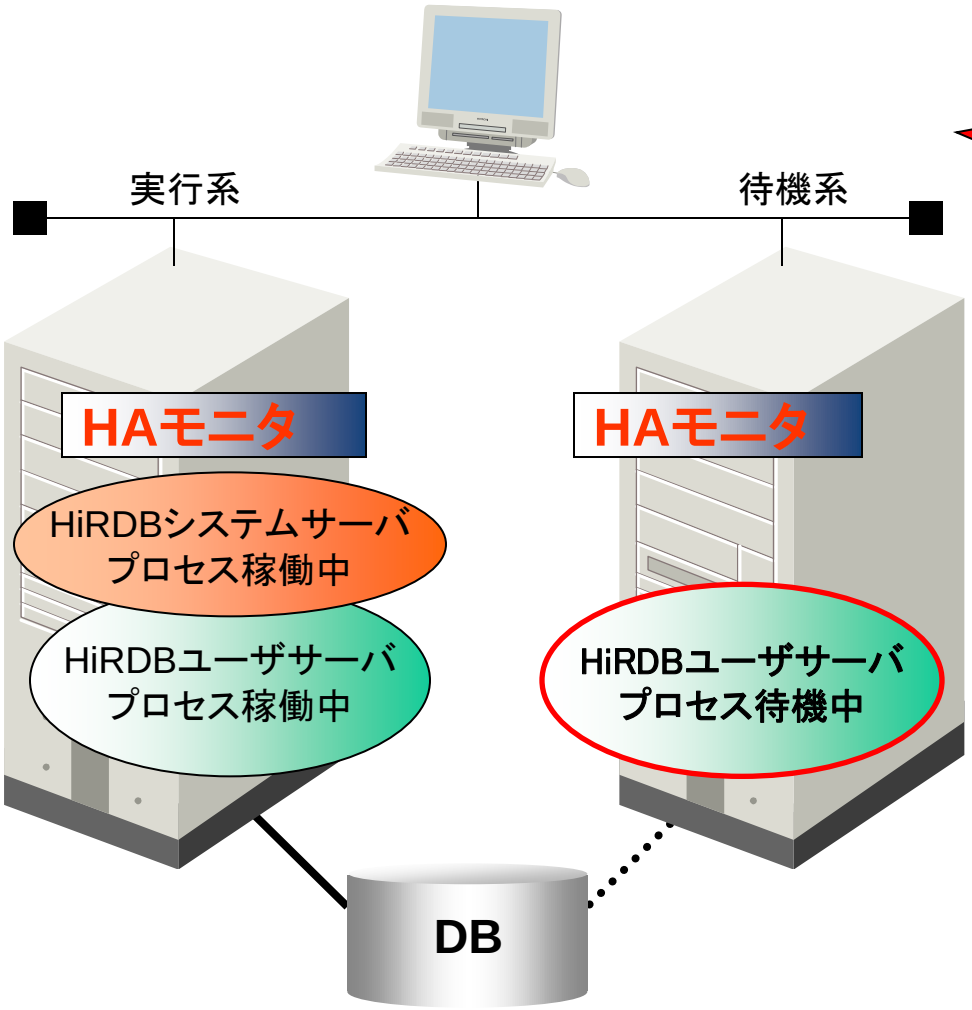


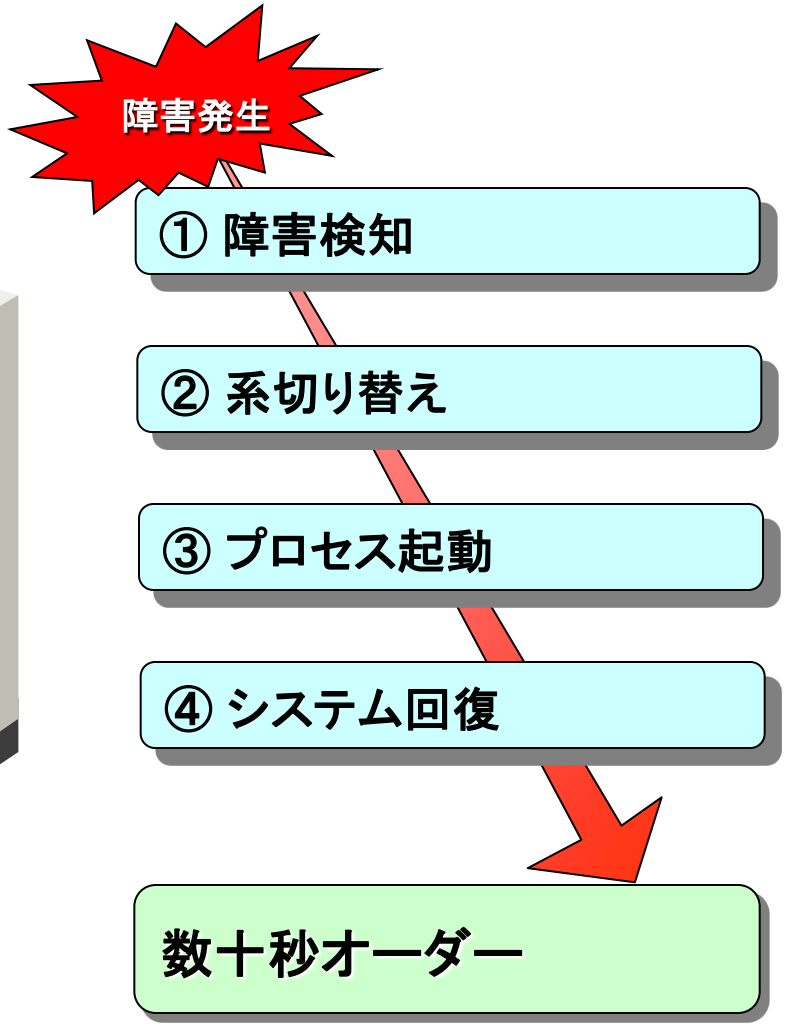


3-3-5 スタンバイ型系切り替え：ユーザサーバホットスタンバイ

HITACHI
Inspire the Next

解説 ユーザサーバホットスタンバイでは、待機系のHiRDBユーザサーバプロセスを待機させておき、系切り替えの時間を数十秒オーダーに短縮できます。

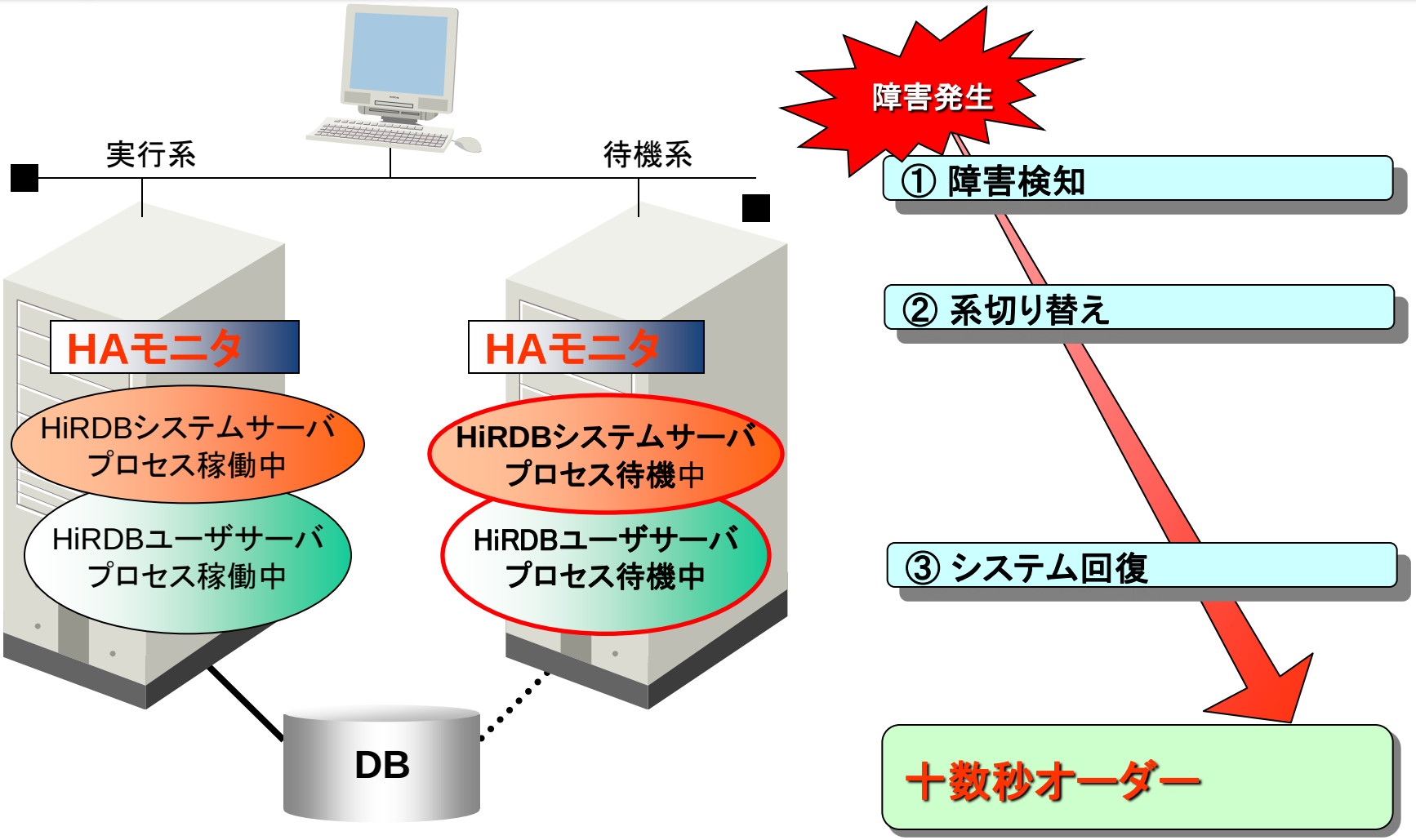




3-3-5 スタンバイ型系切り替え：高速系切り替え

HITACHI
Inspire the Next

解説 高速系切り替えでは、待機系のHiRDBユーザサーバプロセスに加えて、システムサーバプロセスを待機させておき、系切り替えの時間を十数秒オーダーに短縮できます。



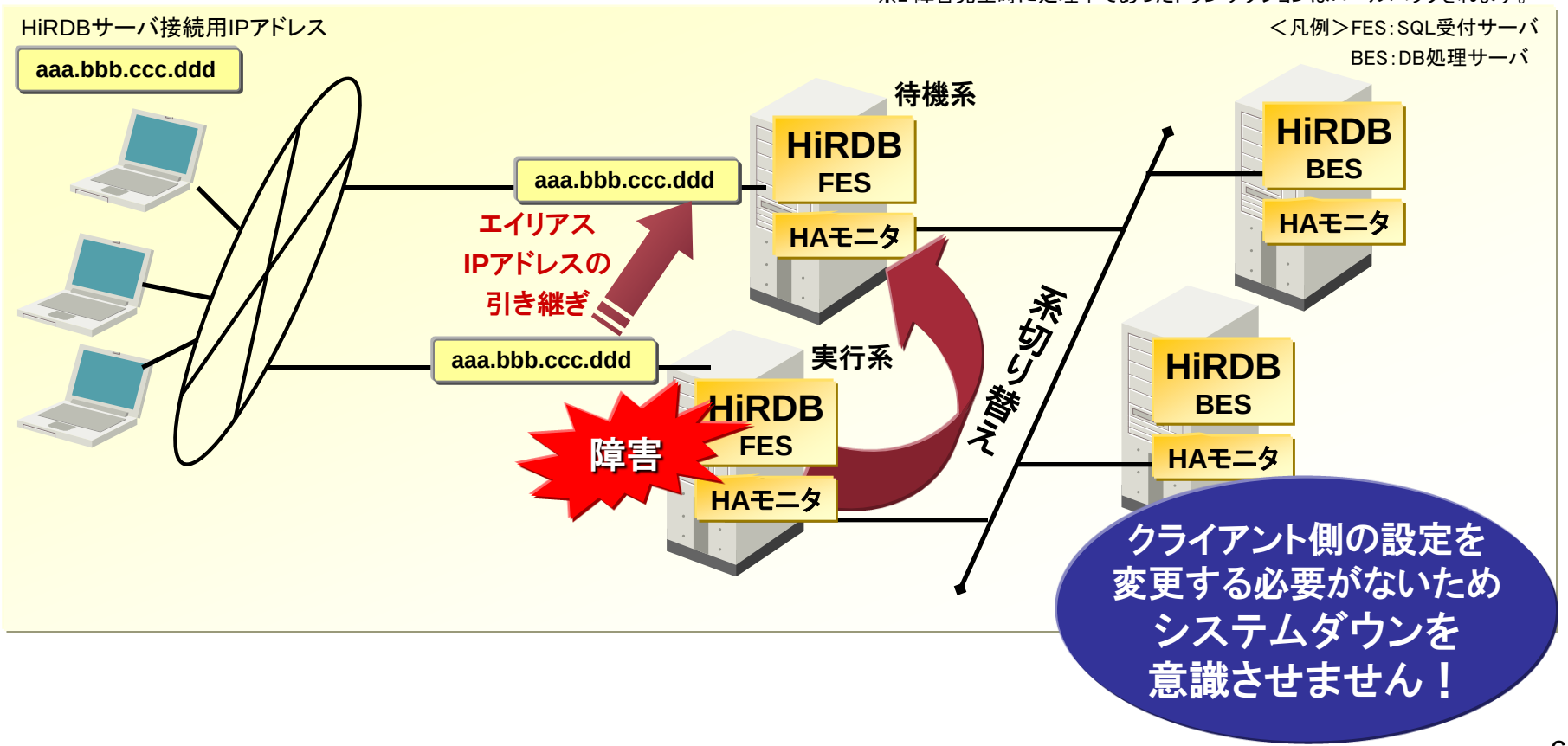
3-3-5 スタンバイ型系切り替え：高速系切り替えのIPアドレス引き継ぎ

HITACHI
Inspire the Next

解説

高速系切り替えの場合、系を切り替える前から待機系のHiRDBが起動処理を実行しているため、系切り替え時に、HiRDBサーバのIPアドレスを引き継ぐことはできません。
そこで、HiRDBクライアントからHiRDBサーバへの接続用IPアドレスを設定しておき、系切り替え時にそのIPアドレスを引き継ぐことで、クライアント側の設定を変更することなく※1、待機系に接続できます。※2

※1 予め予備系のIPアドレスを指定可能ですが、現用系に接続を試みる時間を要するので推奨しません
※2 障害発生時に処理中であったトランザクションはロールバックされます。

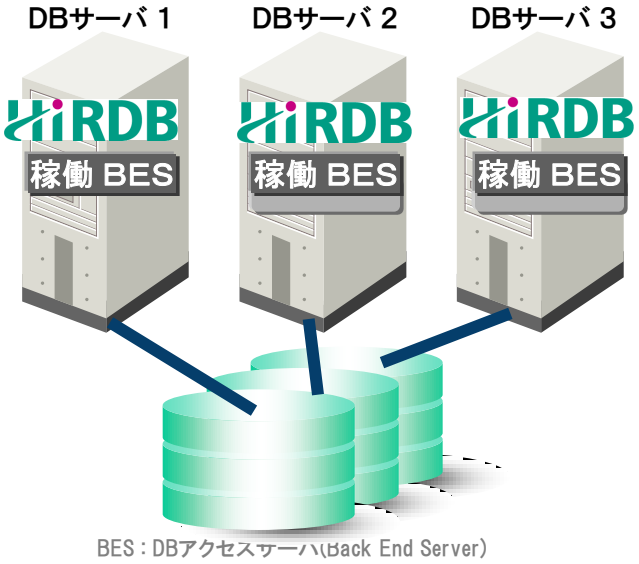


HITACHI
Inspire the Next

ノード障害の影響を分散

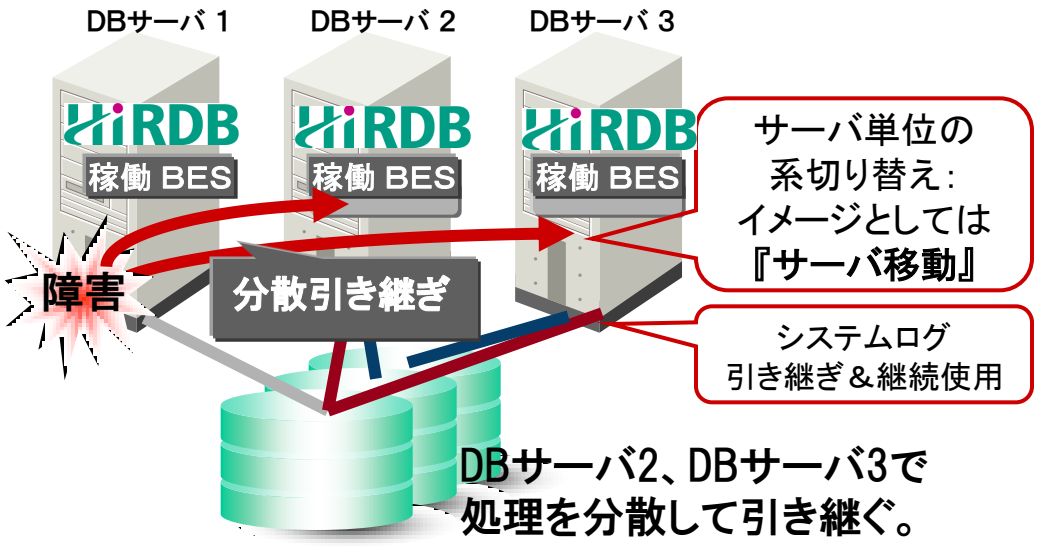
特長	切り替え先をBESごとに複数の稼働中ノードに分散させることにより、障害時における切り替え先のユニットの処理性能劣化を最小限に抑え、安定性能を維持できます。
利点	ノード障害時でも、HiRDBシステム全体での安定稼働を実現できます。待機専用サーバマシンは不要であり、費用効率の高い(経済的な)ソリューションを提供できます。

通常時



BES：DBアクセスサーバ(Back End Server)

ノード障害発生



DBサーバ2、DBサーバ3で
処理を分散して引き継ぐ。
それぞれの負荷はプラス50%。

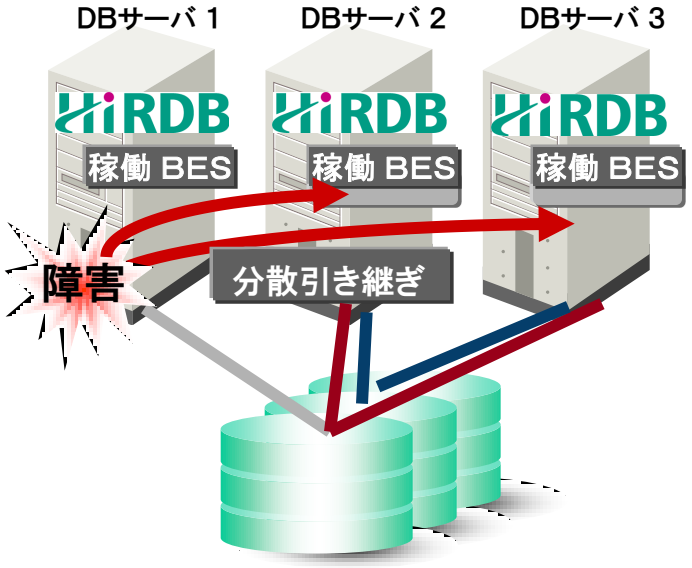
影響分散スタンバイレス型系切り替え（2）

HITACHI
Inspire the Next

多段系切り替え対応

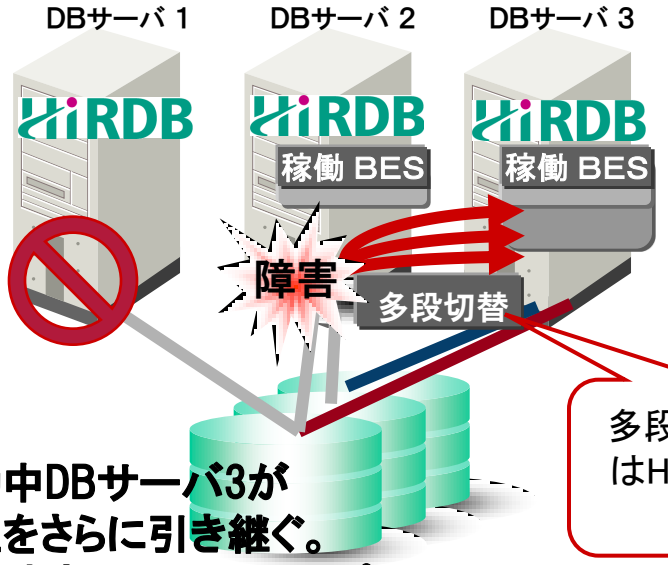
特長	HAモニタ(マルチスタンバイ機能)との連携により、複数ノードに障害が発生した場合にも、第2・第3の稼働中ユニットで障害発生対象BESの処理要求を引き継ぐことができます。
利点	多点障害でも、HiRDBシステムは停まることなく100%のサービスを提供し続けます。

1ノード障害発生



BES：DBアクセスサーバ(Back End Server)

2ノード障害発生



稼働中DBサーバ3が
処理をさらに引き継ぐ。
多点障害でもノンストップ。

多段切替の指定はHAモニタ定義で指定

3-3-6 スタンバイレス型系切り替え：所要リソースを有効活用

HITACHI
Inspire the Next

解説

下の表は待機系ユニットの待機時の所要リソースと系切り替え後の所要リソースを比較したものです。スタンバイレス型系切り替え機能では、HiRDBサーバプロセスとグローバルバッファプール用共用メモリを共用するので、待機中の所要リソースを有効活用できます。

項　目		HiRDBシステム サーバプロセス	HiRDB サーバプロセス	ユニット コントローラ用 共用メモリ	ロックプール用 共用メモリ	グローバル バッファプール用 共用メモリ
スタンバイ型 系切り替え 機能	コールド スタンバイ	○	○	○	○	○
	ユーザ サーバホット スタンバイ	○	●	○	○	○
	高速系 切り替え 機能	●	●	●	●	●
スタンバイ レス型 系切り替え 機能	影響分散 スタンバイレス型 系切り替え機能	△	ー（共用）	●	●	ー（共用）

（凡例）
●：待機完了までに確保して、系の切り替え後も使用します。
○：系の切り替え後、実行系になった時点で確保して使用します。
△：一部のリソースを系切り替え後に実行系になった時点で確保し、使用します。
ー：確保しません。

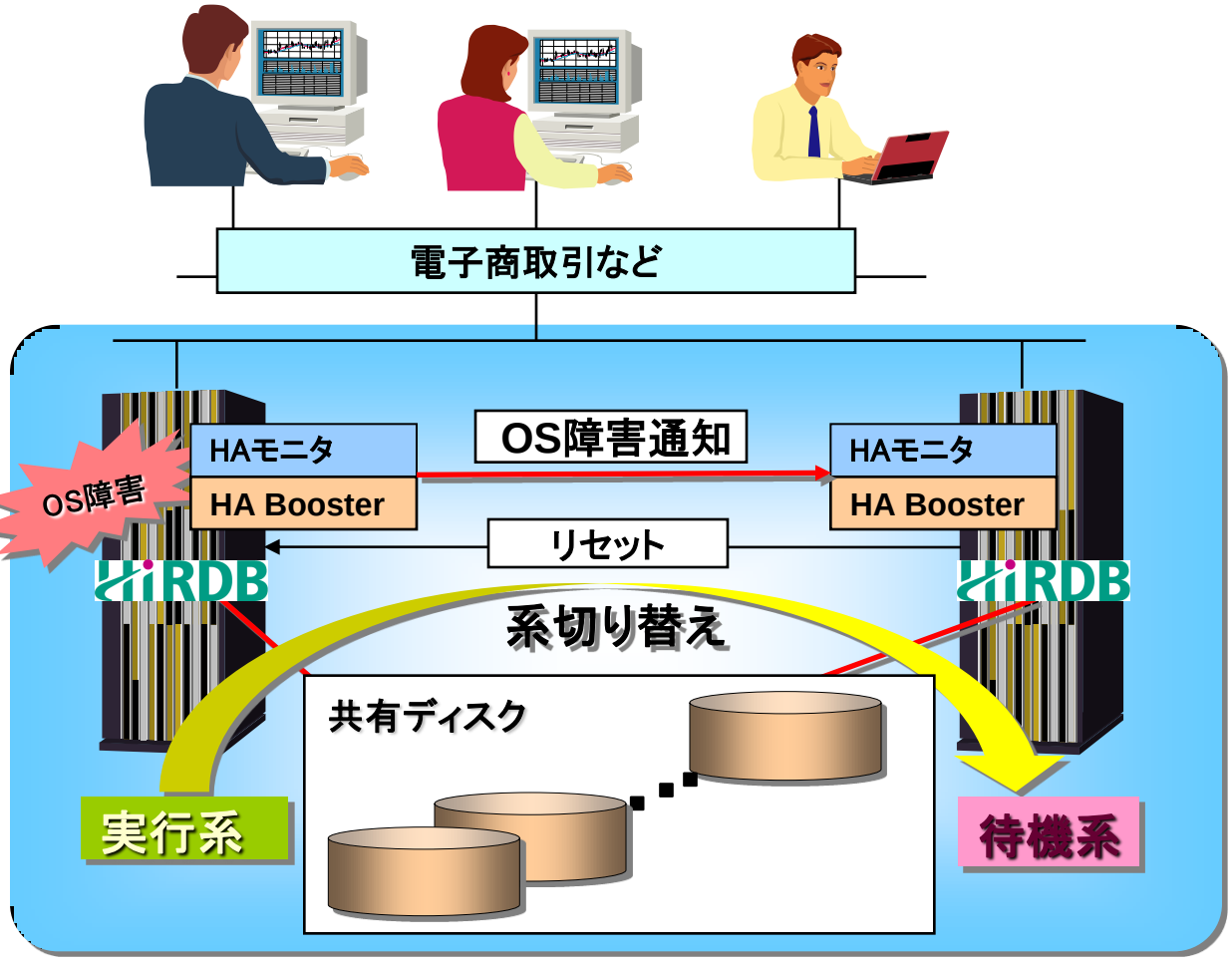
3-3-7 HA Booster Packの概要

HITACHI
Inspire the Next

EP8000(AIX)・BladeSymphony(Linux)
付加機能

解説

さらにHA Boosterにより、十数秒の系切り替え時間を実現できます。
～業務システムは継続してサービスを提供し続けます



HA Booster Packの提供機能

OS障害即時通知機能

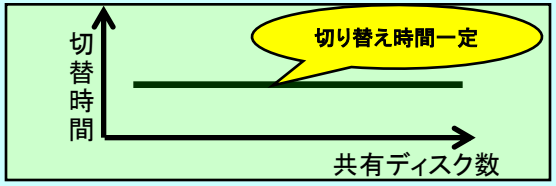
実行系でOS障害を検知し、待機系に通知可能

実行系のOS障害発生と同時に待機系の切り替え処理を開始
⇒切り替え時間短縮

高速ディスク切り替え機能

共有ディスク数に依存しない
安定したディスク切り替え時間を実現

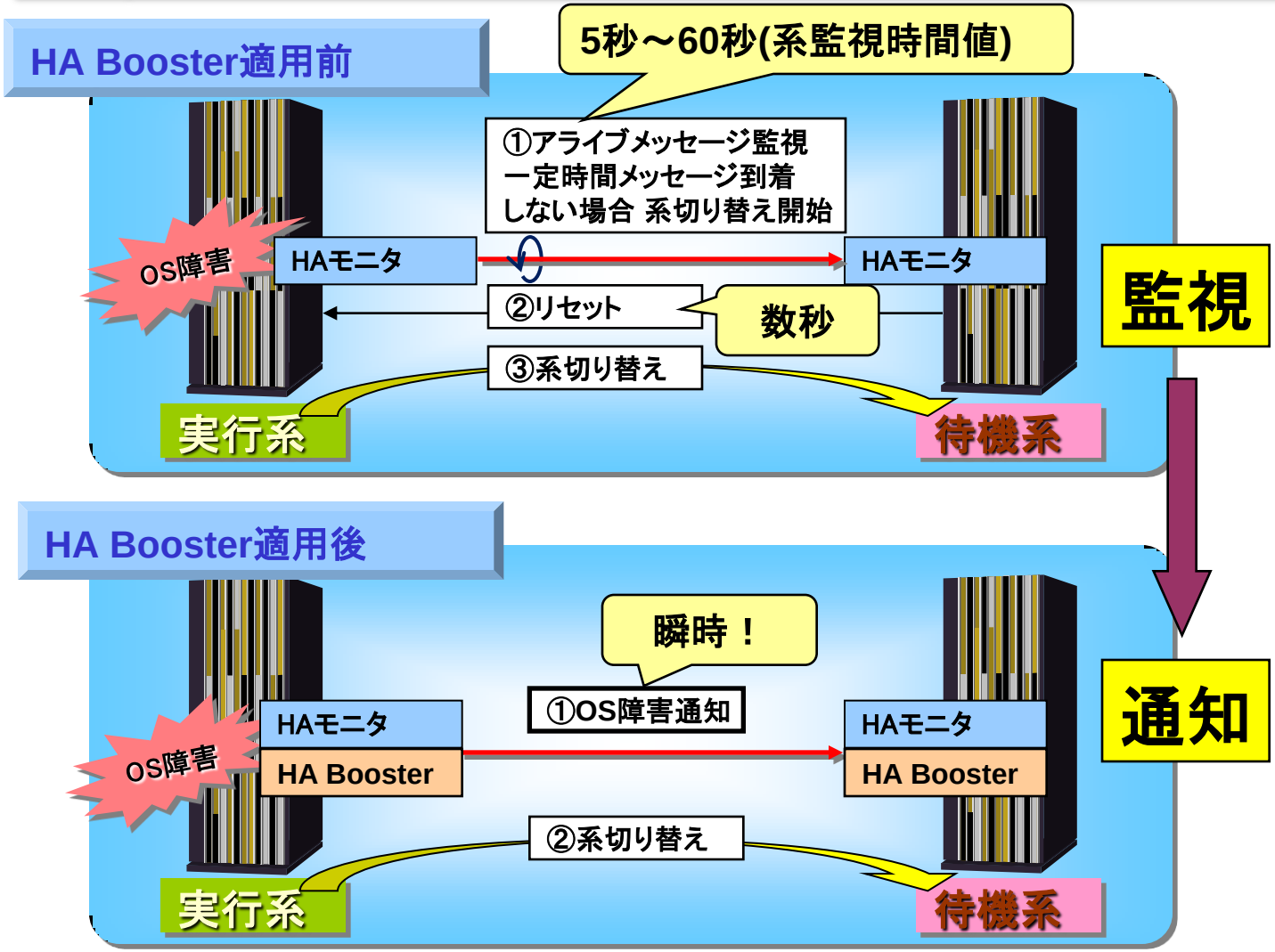
大規模システムにおいても
「数秒～十数秒」
の高速系切り替えを可能とします



3-3-7 HA Booster Packの概要(OS障害即時通知機能)

HITACHI
Inspire the Next

解説 OS障害の即時通知によりヘルスチェック&リセット時間を短縮できます。



適用効果

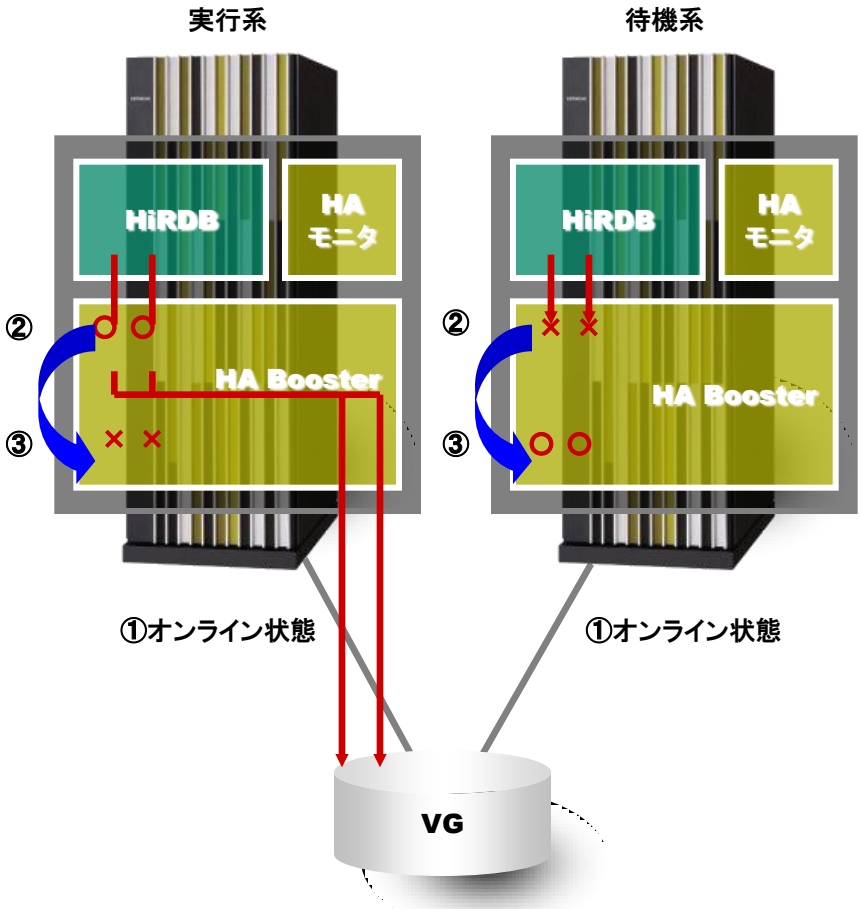
- 系切り替え時間の確実な短縮(系監視時間とリセット時間を短縮)
- OS障害に系監視で対応する必要がなくなり、 処理の沈み込みを考慮 した安全な系監視時間値の設定が可能

3-3-7 HA Booster Packの概要（高速ディスク切り替え機能）

HITACHI Inspire the Next

解説

系切り替え時、実行系から共用ディスクへのアクセスを高速に遮断します。共有ディスク数に依存しない高速なディスク切り替えを実現できます。

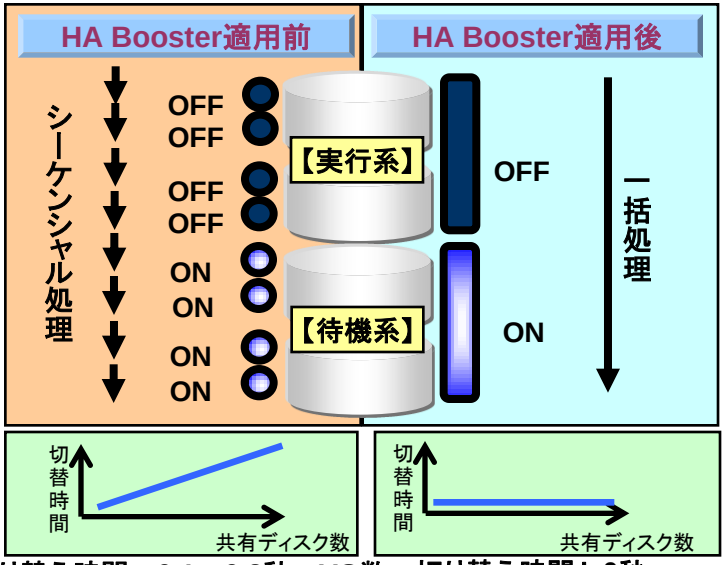


①システム起動時に実行系、待機系ともに共用ディスクをオンライン状態にする。

②HA Boosterがアプリケーションの I/Oに介在し、実行系からのアクセスのみを許可する。

③系切り替え時には、実行系から共用ディスクへのアクセスを高速に遮断する。待機系から共用ディスクへのアクセス許可を一括処理する。これにより、共用ディスク数に依存しない切り替えを実現可能できる。

■EP8000の場合の例



切り替え時間＝0.1～0.2秒×VG数　切り替え時間≒0秒

HITACHI
Inspire the Next

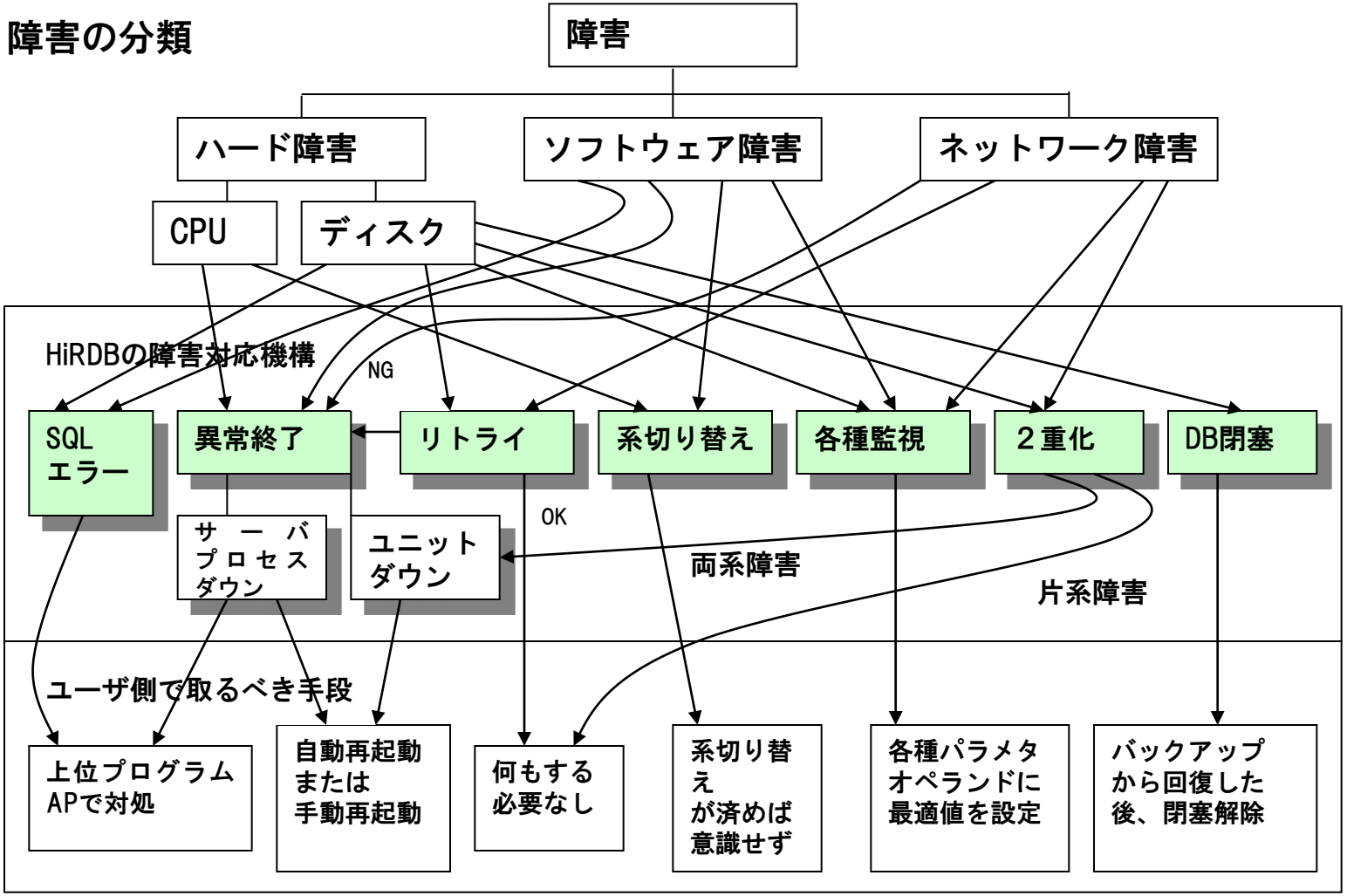
HiRDB

3-4. 障害時の運用

3-4-1 障害対策の基本的な考え方

HITACHI
Inspire the Next

解説 HiRDBの障害の分類とその対応機構について説明します。



3-4-1 障害対策:SQLエラー

HITACHI
Inspire the Next

SQL エラー | 異常終了 | リトライ | 系切り替え | 各種監視 | ２重化 | DB閉塞

解説 SQL実行中にエラーが発生すると、クライアントにエラー要因を返却し、トランザクションを無効にします。

① 外部要因エラー

DBを格納するディスク障害などの媒体障害時はＤＢ閉塞エラーとなります。
通信エラーやカーネルのサービスエラーなど。

②各種資源オーバ、タイマ監視値オーバ

各種資源（排他制御用資源、ＤＢ格納領域など）オーバ。
タイマ値（排他待ち時間、サーバ内実行時間など）オーバ。

③ 内部矛盾検知、その他要因による異常終了

運用操作手順ミス等で、データ破壊に繋がる可能性がある場合、HiRDB自身が異常終了します。

3-4-1 障害対策:異常終了

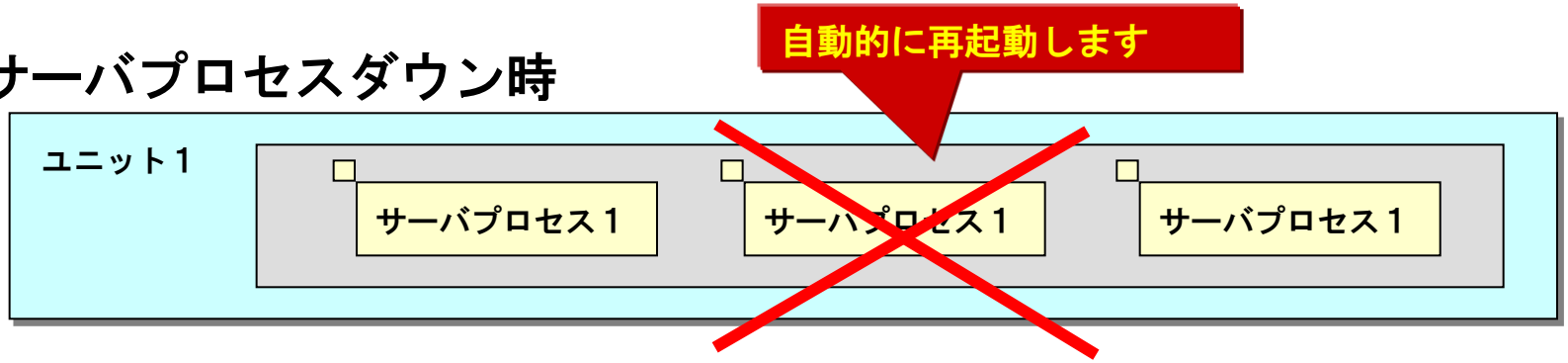
HITACHI
Inspire the Next

SQL エラー | 異常終了 | リトライ | 系切り替え | 各種監視 | ２重化 | DB閉塞

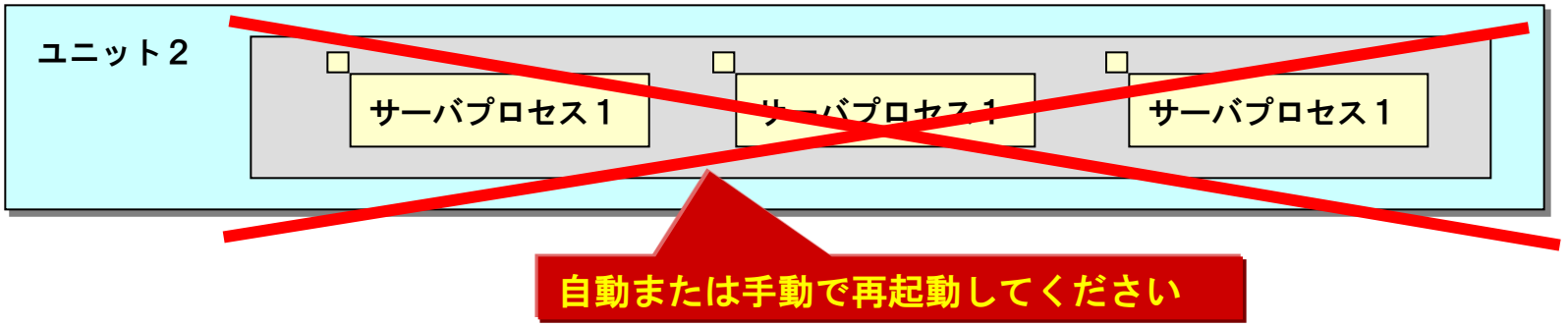
解説

障害の波及範囲に応じて、サーバプロセスを異常終了させたり、HiRDBユニットを異常終了させることができます。
再起動に関しては「自動」と「手動」の2種類があります。

① サーバプロセスダウン時



② ユニットダウン時



3-4-1 障害対策:リトライ(再実行)

HITACHI
Inspire the Next

SQL エラー | 異常終了 | リトライ | 系切り替え | 各種監視 | ２重化 | DB閉塞

解説 再実行することにより、処理続行が見込まれる処理ではシステム内部でリトライします。

リトライする処理の例

◎通信エラー

◎内部ロックエラー(ラッチ等)

◎I/Oエラー

3-4-1 障害対策：系切り替え

HITACHI
Inspire the Next

SQL エラー | 異常終了 | リトライ | 系切り替え | 各種監視 | ２重化 | DB閉塞

解説 障害が発生すると自動的に待機系に系を切り替えて、待機系で業務を続行できます。

詳しくは3-3をご覧ください。

3-4-1 障害対策:各種監視機能

HITACHI
Inspire the Next

SQL エラー | 異常終了 | リトライ | 系切り替え | 各種監視 | ２重化 | DB閉塞

解説 HiRDBの処理時間や消費資源に対して、次のような監視機能を備えています。

① タイマ監視(時間監視)

クライアントからの要求間隔やサーバの応答間隔などを、あらかじめパラメタで指定し、その時間を超えた場合にキャンセルしたりエラーリターンします。

【例】 ●SQL処理時間監視（サーバの応答間隔） ●許容最大排他待ち時間監視
●SQL発行間隔時間監視（クライアントからの要求間隔）

② 資源監視

各種管理領域などを容量を監視し、あらかじめ定義時に指定した値を超えた場合に、警告メッセージなどを出力します。

【例】 ●メモリ領域 ●DB使用領域 ●作業用ファイル使用率

③ 各種上限値監視

同時実行接続ユーザ数、作成可能ファイル数などをあらかじめ定義時に指定した上限値を超えた場合には、メッセージを出力しエラーリターンする。

【例】 ●排他資源数の上限値 ● ファイル作成数の上限値
●シンクポイントダンプの有効化処理の連続スキップの回数の上限値

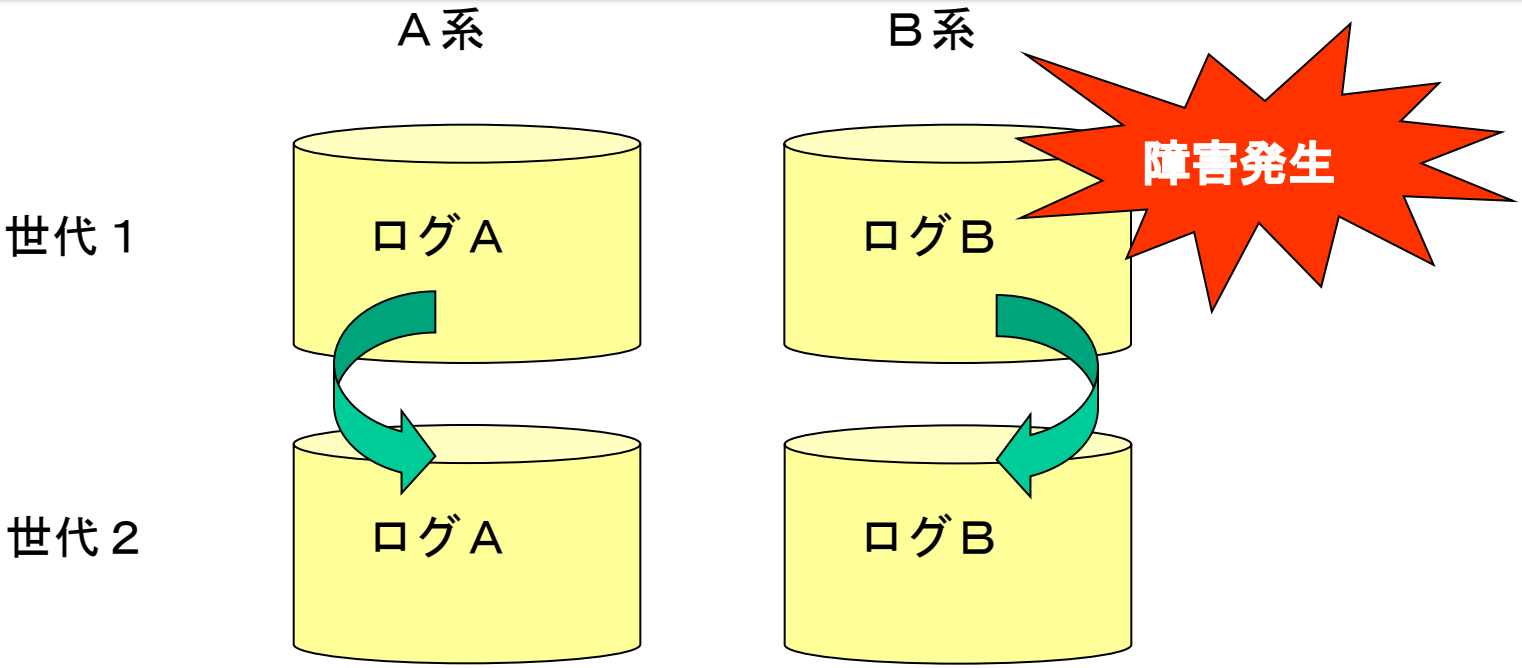
3-4-1 障害対策:2重化

HITACHI
Inspire the Next

SQL エラー | 異常終了 | リトライ | 系切り替え | 各種監視 | ２重化 | DB閉塞

解説

システムファイル(システムログ、ステータスファイル等)を2重化し、片側のディスクに障害が発生しても、もう片方で処理を継続的に運用することができるようになります。

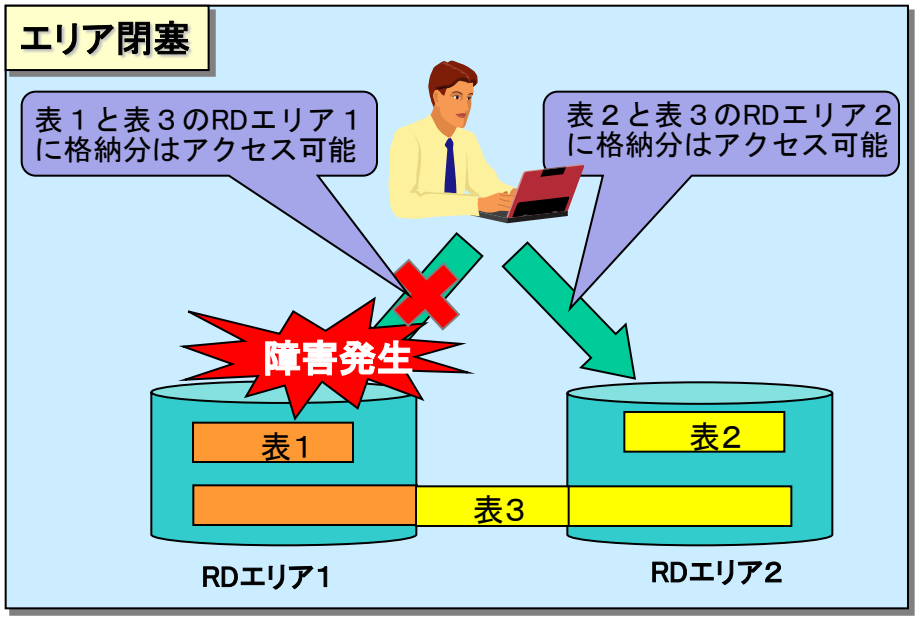


3-4-1 障害対策：DB閉塞

HITACHI
Inspire the Next

SQL エラー | 異常終了 | リトライ | 系切り替え | 各種監視 | ２重化 | DB閉塞

解説 データベース(RDエリア)に障害が発生した場合、障害によって使用できなくなる範囲をできるだけ縮小し、他の正常な範囲での業務を続けます。



●ディスクアクセスのエラーやページ内の矛盾検知した場合、対象となるRDエリアを閉塞し、当該領域へのアクセスを禁止します。

●閉塞種別の細分化を行っています。
　メモリ上の矛盾検知などは、閉塞要因を取り除き再度RDエリアをオープンすれば良いものに対しては「コマンド閉塞」にします。
　一方、障害発生によりアクセスを禁止するものに関しては「障害閉塞」にします。

3-4-2 トラブルシュート機能

HITACHI
Inspire the Next

解説 HiRDBでは、以下のトラブルシュート機能を実装しています。

① 自己診断機能

プログラム走行中に発生するシステム内の矛盾を検知し、矛盾の発生時点にできるだけ近い時点で矛盾の拡大を防止する機能です。制御ブロックの妥当性チェックやインタフェースの妥当性チェックなどを行っています。

② トレース機能

事象の遷移(タスク、プロセススイッチ、モジュール間リンクなど)が頻繁に行われる状況で矛盾が発生した場合に、どの時点での遷移が直接的な原因なのかを検出するために、遷移の軌跡を残す機能です。

③ 情報保存機能

自己診断機能、トレース機能を用いて収拾した情報を参照可能な形式で保存できる機能です。情報保存としてメモリ、ディスク、システムログファイルなど保存方法を使い分けてお使い頂けます。

3-4-2 障害調査に必要なトラブルシュート情報

HITACHI
Inspire the Next

解説 障害調査のためには、以下のトラブルシュート情報が必要となります。

トラブルシュート情報		SQLエラー	ユティリティ、コマンドエラー	無応答	レスポンス遅延
HiRDB クライアント	エラーログファイル（クライアント環境変数PDCLTPATHの下のpderr*.trc）	○			
	SQLトレース情報、UAP統計レポート情報（クライアント環境変数PDCLTPATHの下のpdsql*.trc）				△
HiRDB サーバ	ユティリティ、コマンドのメッセージ（標準出力、標準エラー出力）		○		
	HiRDB運用ディレクトリ/SPOOLの下	◎	◎	◎	◎
	HiRDBの情報（起動状態など）	◎	◎	◎	◎
OS	syslog（Windowsの場合はイベントログ）	◎	◎	◎	◎
	OSの情報（バージョンやプロセス情報など）	◎	◎	◎	◎

凡例 ○：取得が必要　◎：取得が必要（pdinfogetコマンドで取得可能）　△：場合により、取得が必要

3-4-2 pdinfogetコマンドの使用方法

HITACHI
Inspire the Next

解説 トラブルシュート情報は、pdinfogetコマンドで一括して取得してください。

1. 容量の調査※1

【コマンド実行例】
pdinfoget -m

2. 情報の取得

手順(1)で表示された空き容量を確保し、トラブルシュート情報の取得を実行します。

【コマンド実行例】
pdinfoget -e /tmp1 ：初期情報ファイル出力先ディレクトリ名
-d /tmp2 ：詳細情報ファイル出力先ディレクトリ名

3. 初期情報の送付

まず、トラブルの切り分けに必要な初期情報ファイルを送付します。次のメッセージが出力された時点で、初期情報の取得は完了していますので、コマンドの終了を待たずに、送付することができます。
KFPN10403-I init information file output, file=初期情報ファイルパス名

【初期情報ファイル名】
初期情報ファイル：pdinfoget_ホスト名称_init_日時.tar ※2
実行ログ：pdinfoget_comlog_日時 ※2

4. 詳細情報の送付

コマンド終了後、詳細情報ファイルを送付します。

【詳細情報ファイル名】
詳細情報ファイル名：pdinfoget_ホスト名称_detail_日時.tar ※2

※2：日時は「MMDDhhmmss」(MM：月、MM：日、hh：時、mm：分、ss：秒)です。

他社所有名称に対する表示

HITACHI
Inspire the Next

- Adobe，およびFlashは，Adobe Systems Incorporated(アドビシステムズ社)の米国ならびに他の国における商標または登録商標です。
- AIXは，米国およびその他の国におけるInternational Business Machines Corporationの商標です。
- BSAFEは，EMC Corporation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。
- SAP BusinessObjectsは、ドイツのSAP AGまたはSAP関連会社の商標または登録商標です。
- CORBAは、Object Management Groupが提唱する分散処理環境アーキテクチャの名称です。
- DNCWARE，ClusterPerfectは，東芝ソリューション株式会社の商標です。
- ER/Studioは，米国Embarcadero Technologies, Inc.の登録商標です。
- HACMPは，米国およびその他の国におけるInternational Business Machines Corporationの商標です。
- HP-UXは，Hewlett-Packard Development Company, L.P.のオペレーティングシステムの名称です。
- IBM,および Cognosは，米国およびその他の国におけるInternational Business Machines Corporationの商標です。
- Internet Explorerは，米国Microsoft Corporation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。
- Javaは，Oracle Corporation 及びその子会社，関連会社の米国 及びその他の国における登録商標または商標です。
- JDBCは，Oracle Corporation 及びその子会社，関連会社の米国 及びその他の国における登録商標または商標です。
- Linuxは，Linus Torvalds氏の日本およびその他の国における登録商標または商標です。
- HP Serviceguardは，Hewlett-Packard Development Company, L.P.の商品名称です。
- Microsoft .NET は，お客様，情報，システムおよびデバイスを繋ぐソフトウェアです。
- Microsoft Accessは，米国Microsoft Corporationの米国及びその他の国における登録商標または商標です。
- Microsoft Office Excelは，米国 Microsoft Corporation の商品名称です。
- Microsoftは，米国Microsoft Corporationの米国及びその他の国における登録商標または商標です。
- ODBCは、米国Microsoft Corporationが提唱するデータベースアクセス機構です。
- OLEは、Object Linking and Embeddingの略です。
- OLEは、米国Microsoft Corporationが開発したソフトウェア名称です。
- Oracleは，Oracle Corporation 及びその子会社，関連会社の米国 及びその他の国における登録商標または商標です。
- PA-RISCは，Hewlett-Packard Development Company, L.P.の商標です。
- QlikView は QlikTech International AB. の登録商標です。
- Red Hatは，米国およびその他の国でRed Hat, Inc. の登録商標若しくは商標です。
- RSAは，EMC Corporation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。
- SAPは，SAP AGのドイツ及びその他の国における登録商標または商標です。
- Solarisは，Oracle Corporation 及びその子会社，関連会社の米国 及びその他の国における登録商標または商標です。
- UNIXは，The Open Groupの米国ならびに他の国における登録商標です。
- VERITASおよびVERITASロゴは，米国Symantec Corporationの登録商標です。
- Windows Serverは，米国Microsoft Corporationの米国及びその他の国における登録商標または商標です。
- Windows Vistaは，米国Microsoft Corporationの米国及びその他の国における登録商標または商標です。
- Windowsは，米国Microsoft Corporationの米国及びその他の国における登録商標または商標です。
- X/Openは，The Open Groupの英国ならびに他の国における登録商標です。
- その他記載の会社名，製品名は，それぞれの会社の商標もしくは登録商標です。

HITACHI
Inspire the Next

END

HiRDB

HiRDBアーキテクチャと運用管理解説

2011/01

株式会社 日立製作所 情報・通信システム社
ITプラットフォーム事業本部 開発統括本部 DB設計部

Human Dreams.
Make IT Real.

私たちは、ITと制御技術、そして社会インフラシステムで
人々の夢をかなえるイノベーションを起こしていきます。


HITACHI
Inspire the Next