JAVCOM技術研究委員会7月勉強会　資料

Japan Data Storage Forum

～今さら聞けないストレージデバイスの基礎知識　第2回～

平成21年7月28日

JDSF技術交流ワーキンググループ

部会長　岡本　隆行（Atix LLC）

JDSF
Japan Data Storage Forum

Agenda

Ⅰ．今さら聞けないストレージデバイスの基礎

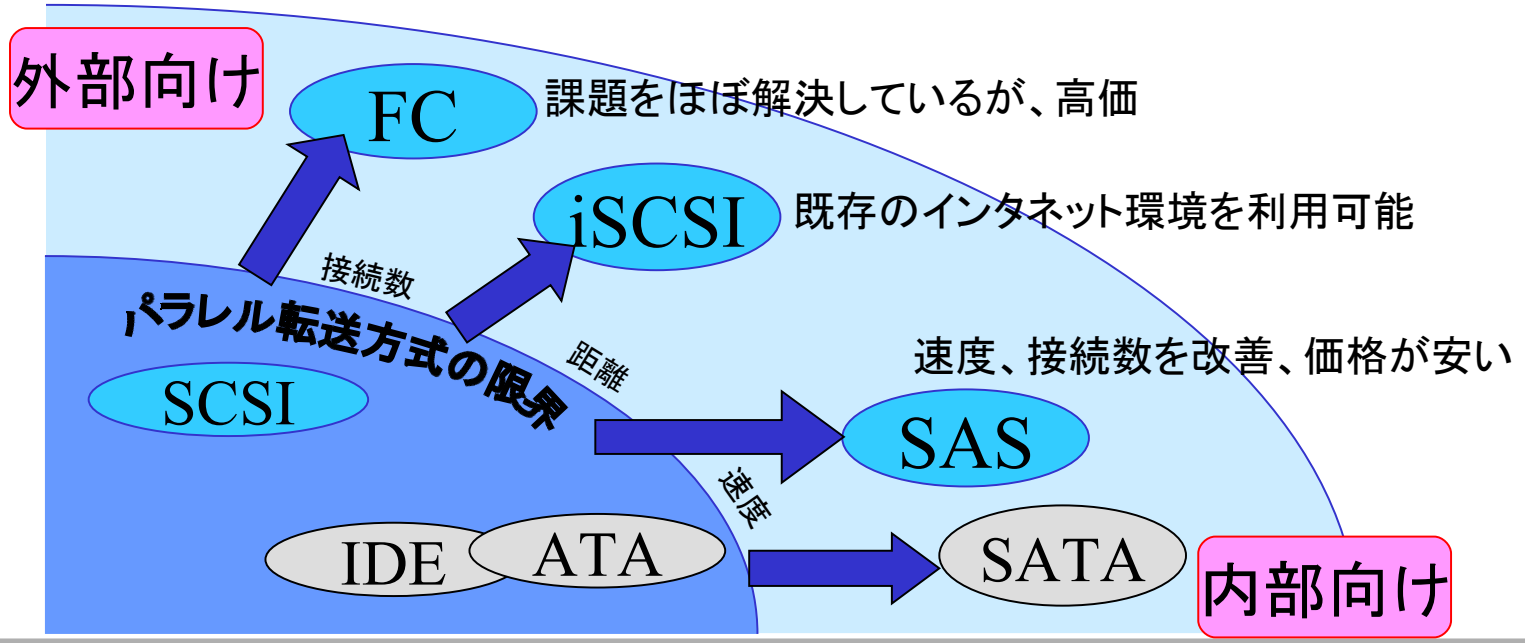
1．インターフェース

2．RAID

3．SSD

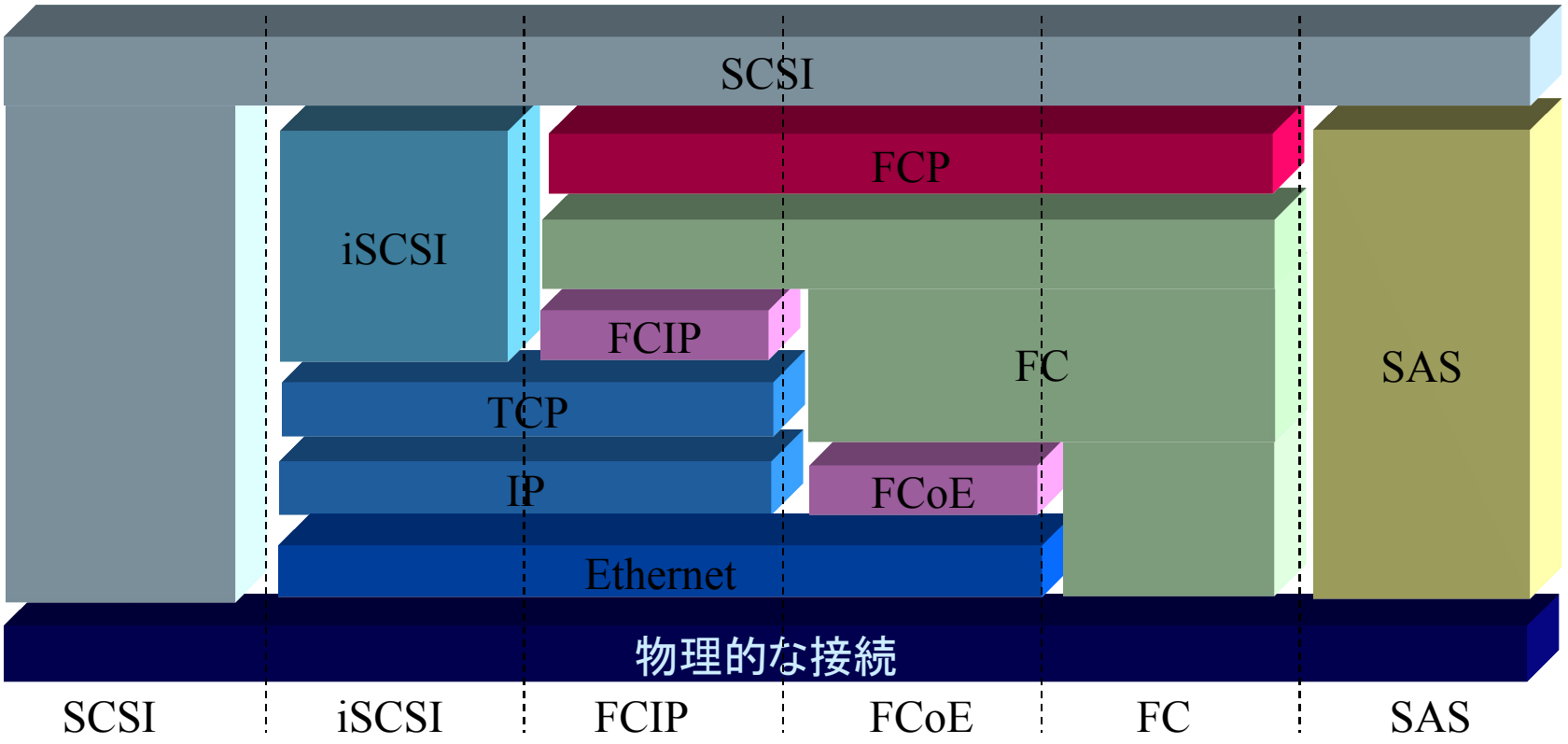
1. インターフェース

- SATA、SAS、FC、iSCSIの違い・特長
  - SAS, FC, iSCSIは、SCSIから継承、派生したインタフェースで、機器間のコマンドインタフェースは、SCSIコマンドをベースとしている
  - 物理的な接続(ケーブル、コネクタなど)、電気的仕様、データ転送方法は異なるが、論理的にSCSIコマンドを使用しているところは共通



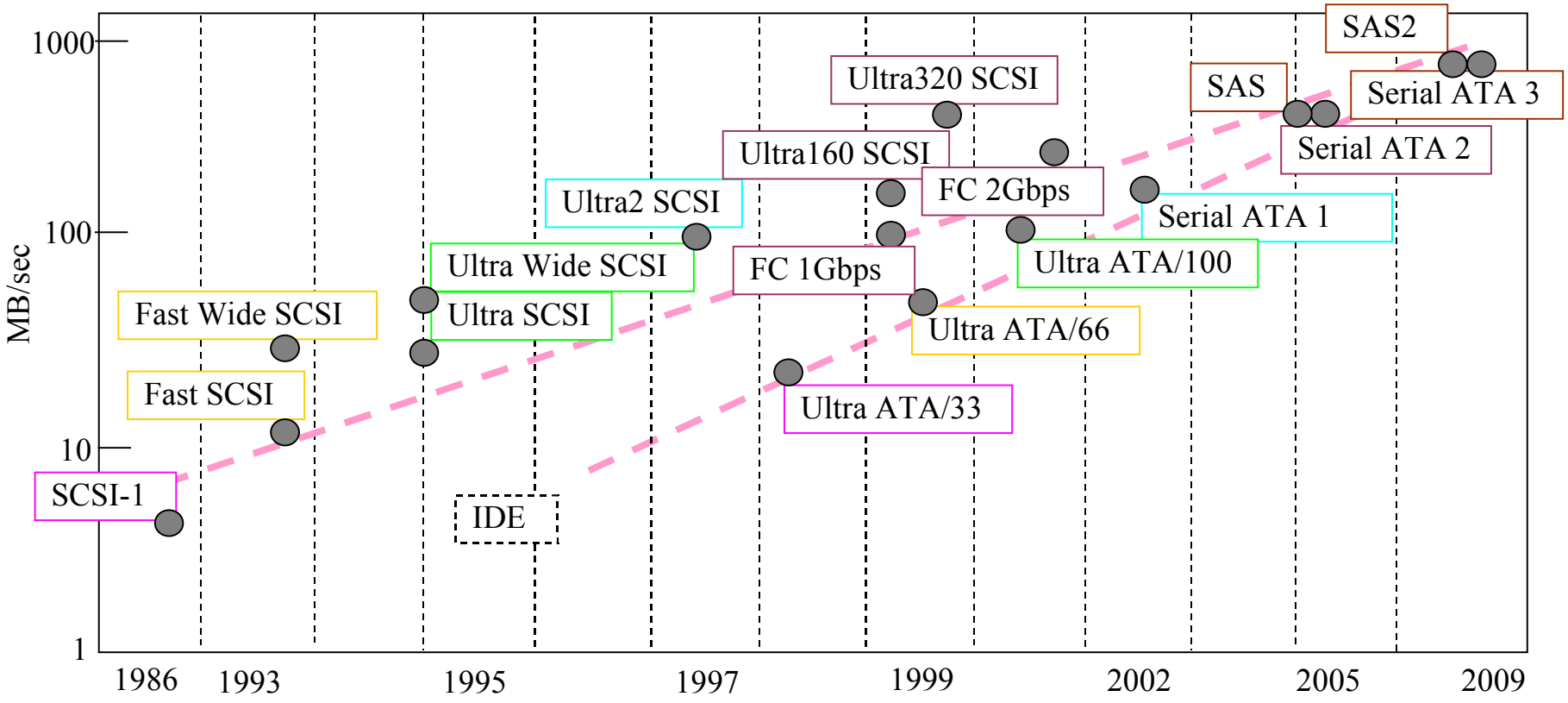
1. インターフェース

- SATA、SAS、FC、iSCSIの違い・特長
  - (前項の通り)SCSIコマンドを使用しているところは共通



1. インターフェース

- HDDインターフェースの歴史

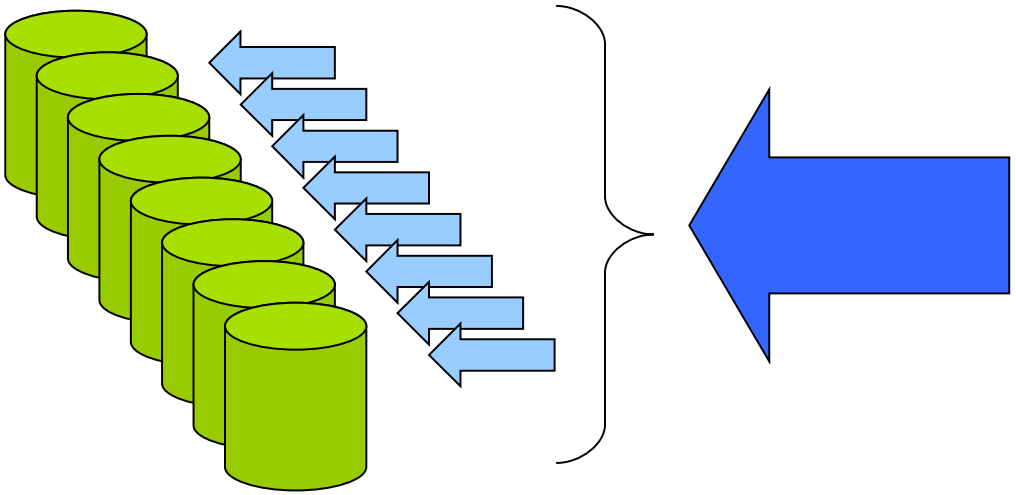


2. RAID

Redundant Array of Independent Disksの略。ハードディスクドライブを複数並列に接続し、容量と速度を向上を図る目的で生まれた技術。

ストライピングとデータに冗長性を持たせた。

今日では外部ストレージとしてRAID機能以外に種々の機能が付加されている。

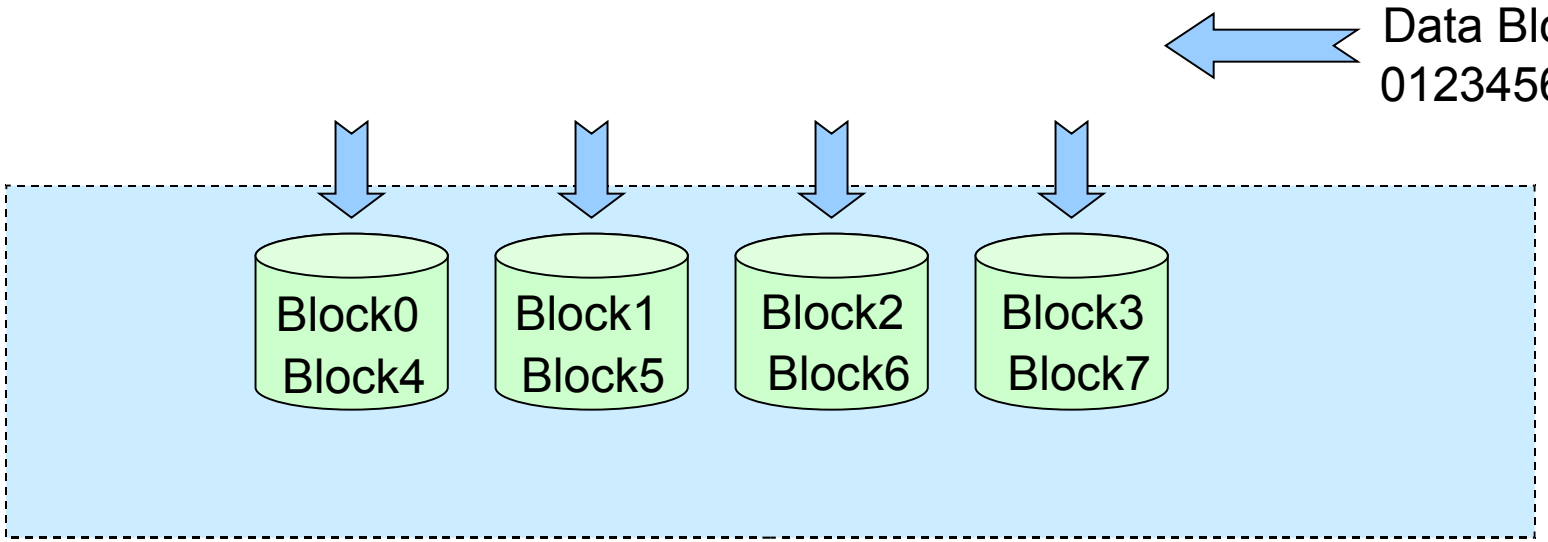


2. RAID

RAID レベル	一般 名称	データ ディスク	説　明	データ 信頼性	データ 転送量	I／O レート
0	ディスク ストライピング	N	データはアレイ内の複数のディスクに 分配される。	△	○	◎
1	ミラーリング	2N 3N等	N台それぞれのディスクに全てのデータ を複写する。	◎	○	○
2		N+m	データはハミングコードで保護されている。 冗長の情報はm台のディスクに分配。	○	◎	○
3	パリティ付きの パラレル転送ディスク	N＋1	データは複数のディスクに下位区分され分 配される。冗長情報は専用ディスク。	○	◎	○
4		N＋1	データは複数のディスクに分配される。冗 長情報は専用ディスク。	○	△	△
5		N＋1	データは複数のディスクに分配される。冗 長情報はアレイ内のディスクにばらまく。	○	○	○
6		N＋2	RAIDレベル5に似ているが、独立して計 算される冗長情報が付加されている。	◎	○	○

2. RAID

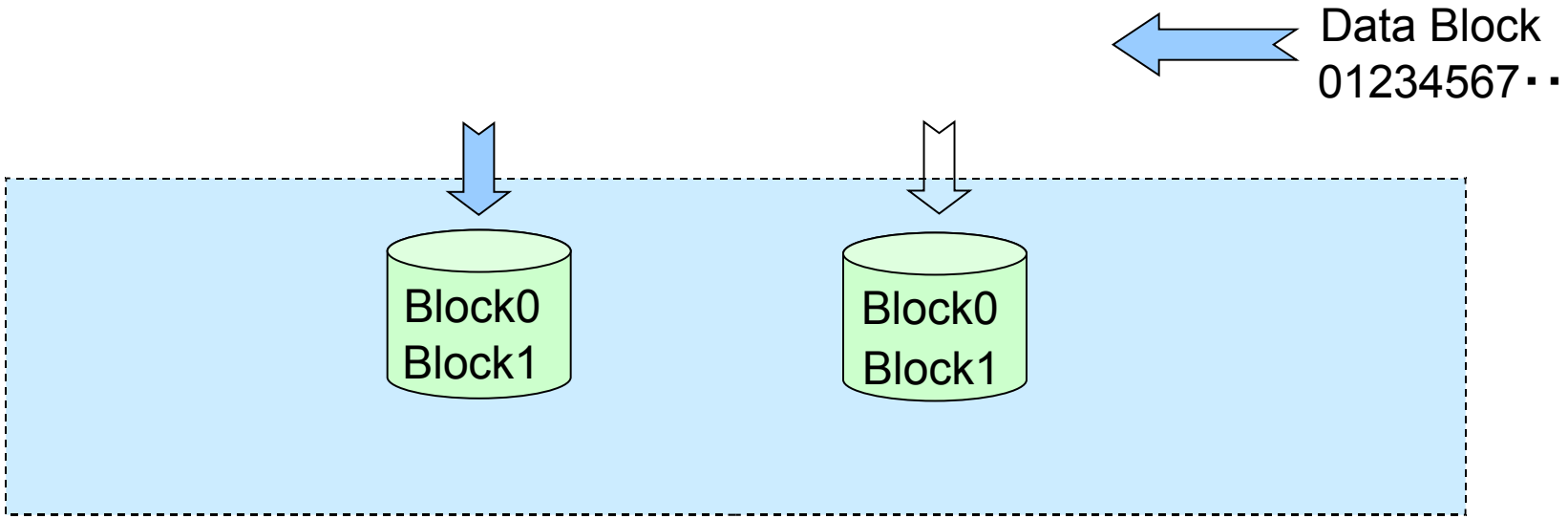
RAIDレベル0：ストライピング



・データはブロックに分割され、各ドライブに分散される。
・冗長性への考慮は無い。
・高速な I／O が可能。

2. RAID

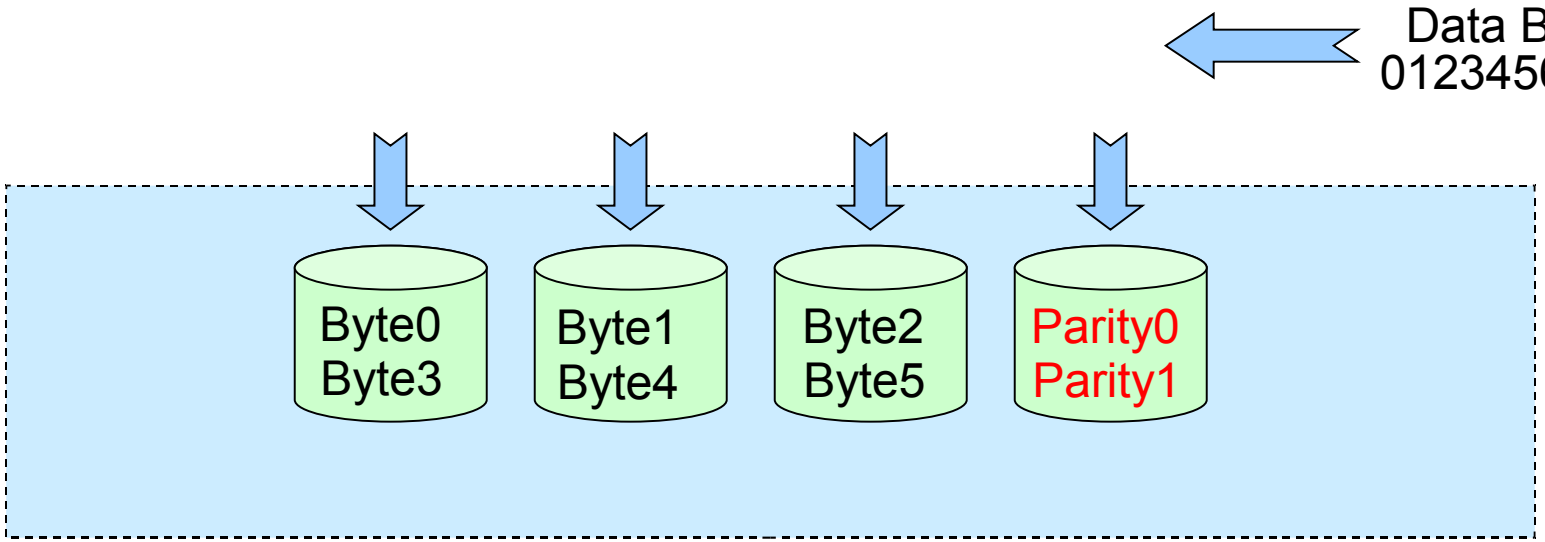
RAIDレベル1:ミラーリング



・データは各ドライブに同じ内容が記録される。
・実容量は物理容量の半分。
・冗長性あり。

2. RAID

RAIDレベル3

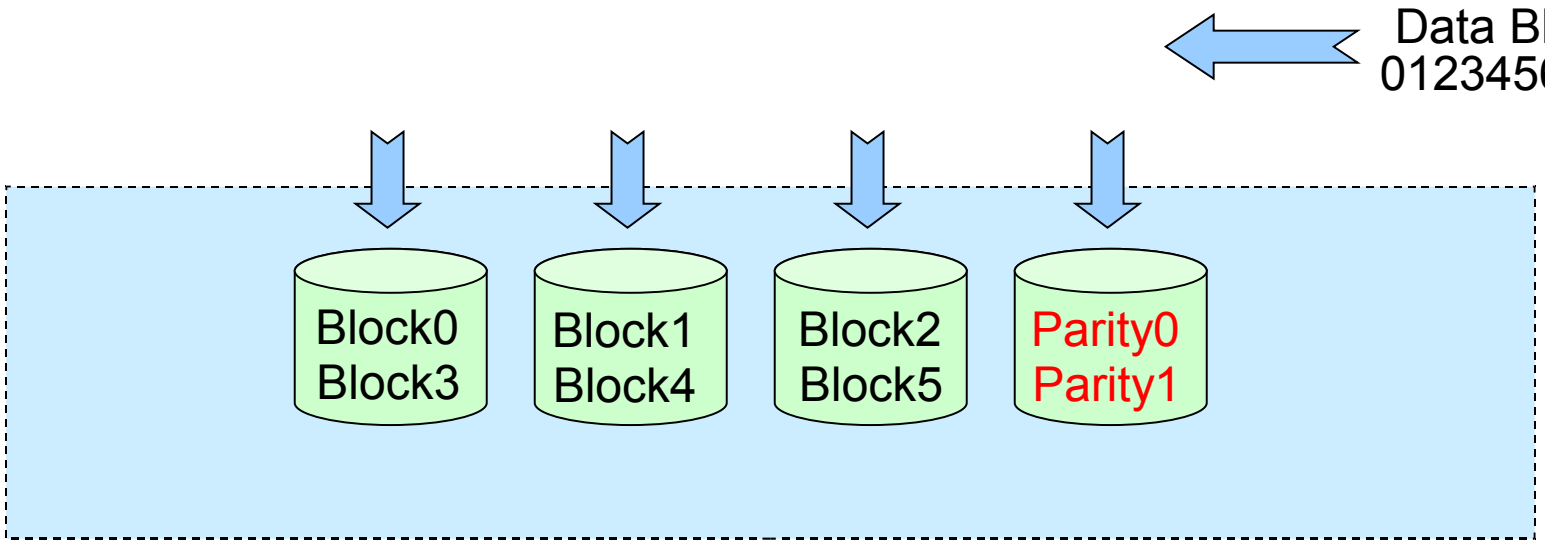


・データはバイトに分割され、各ドライブに分散される。パリティは特定のドライブに置かれる。

・大容量のデータを一度にI／Oする場合に適している。冗長性あり。

2. RAID

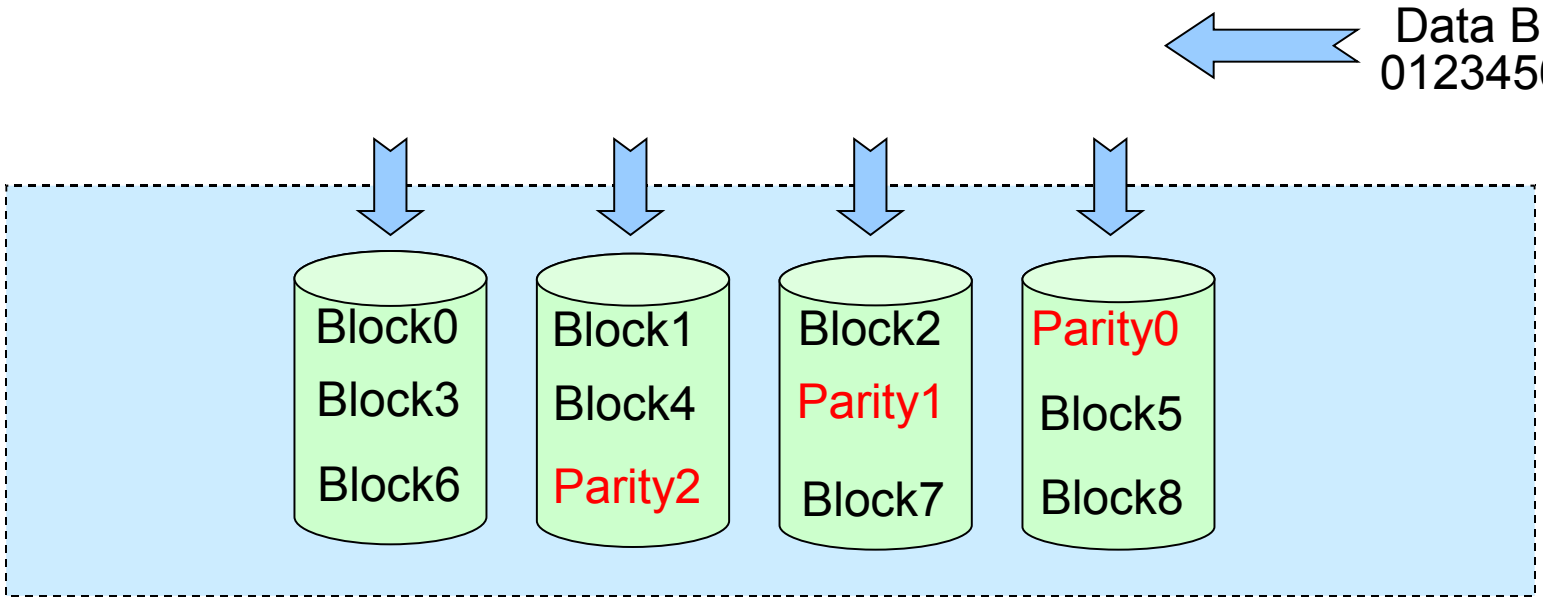
RAIDレベル4



・データはブロックに分割され、各ドライブに分散される。パリティは特定のドライブに置かれる。

2. RAID

RAIDレベル5：パリティ付きストライピング

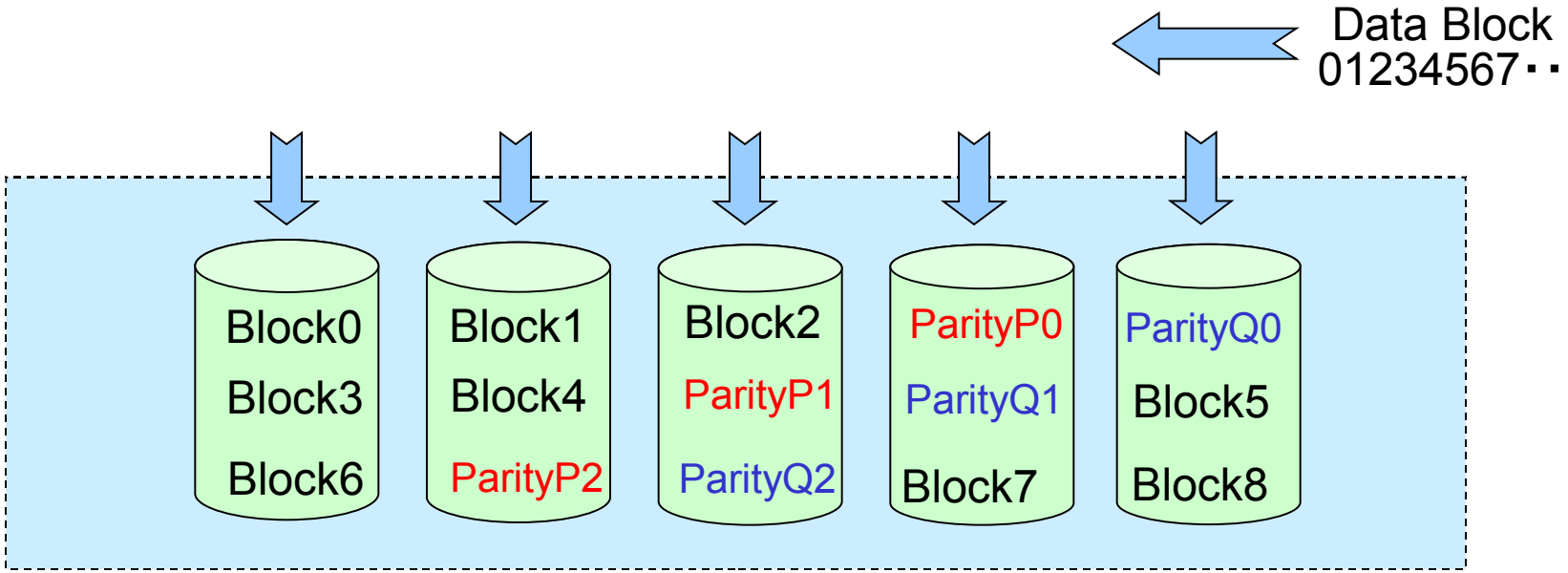


・データはブロックに分割され、各ドライブに分散される。パリティも各ドライブに分散して置かれる。

・小さなデータを頻繁に I／Oする場合に適している。

2. RAID

RAIDレベル6：パリティ付きストライピング

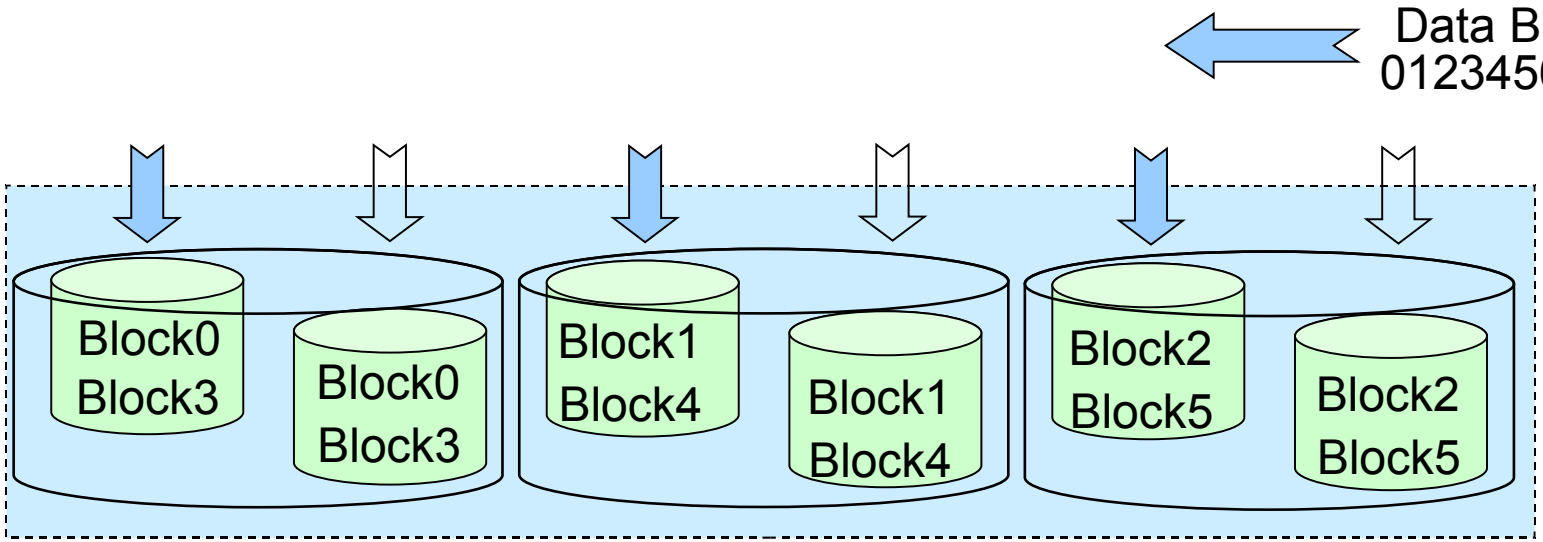


・データはブロックに分割され、各ドライブに分散される。パリティも各ドライブに分散して置かれる。

・ PQ方式、2次元パリティ、対角パリティがある。

2. RAID

RAIDレベル10：ストライピング・ミラー

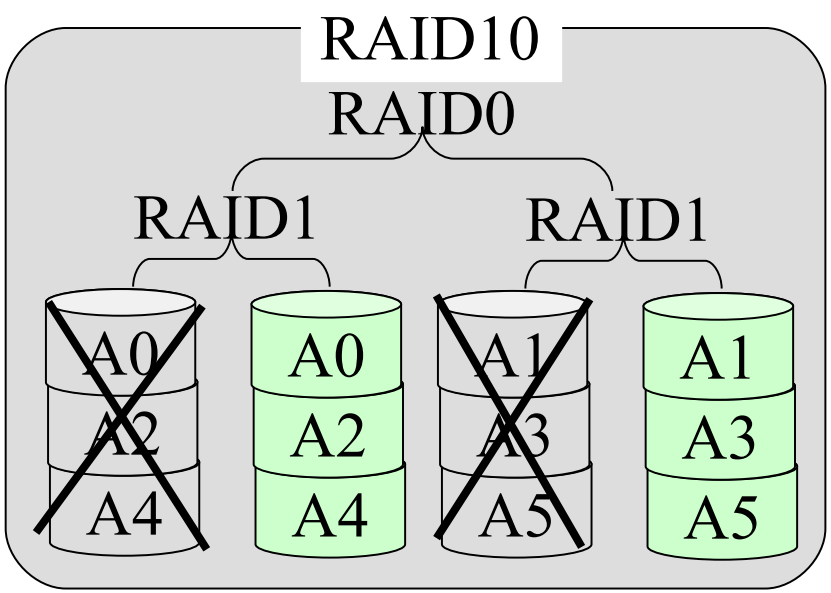


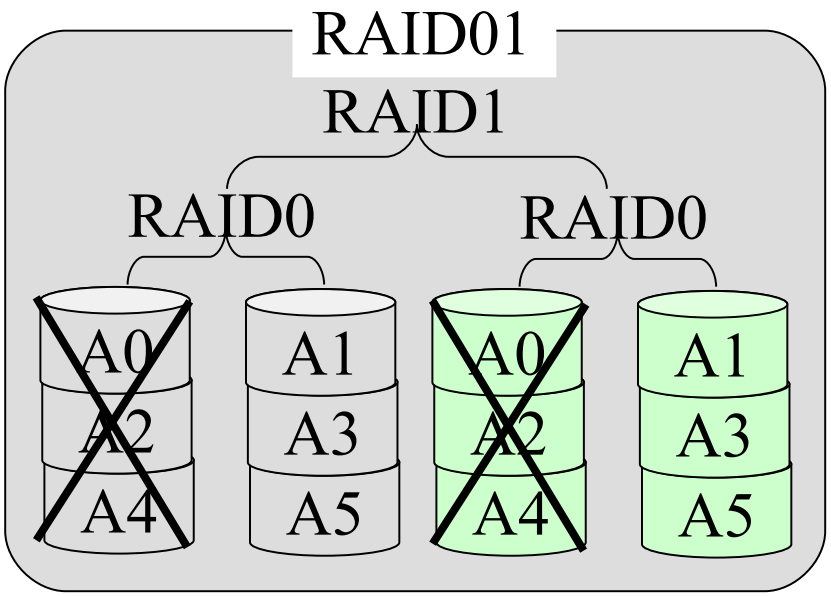
・データは分割され、各ドライブに分散され、別ドライブに同じデータが格納される。

・実容量は物理容量の半分。高速な I／O が可能。冗長性あり。

2. RAID

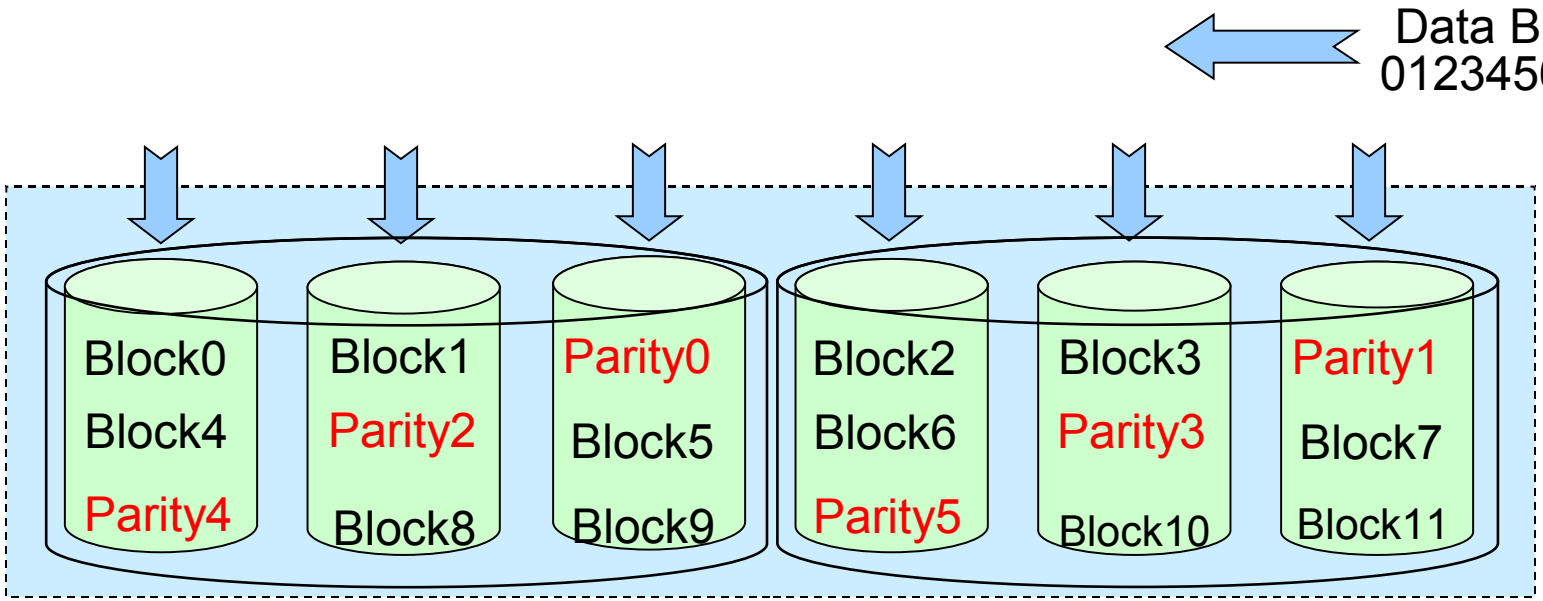
- RAID10(1+0)とRAID01(0+1)の違い
  - RAID10はミラーリングしたアレイをストライピング、RAID01はストライピングしたアレイをミラーリング
  - HDD故障時のデータ保全性が異なります





2. RAID

RAIDレベル50

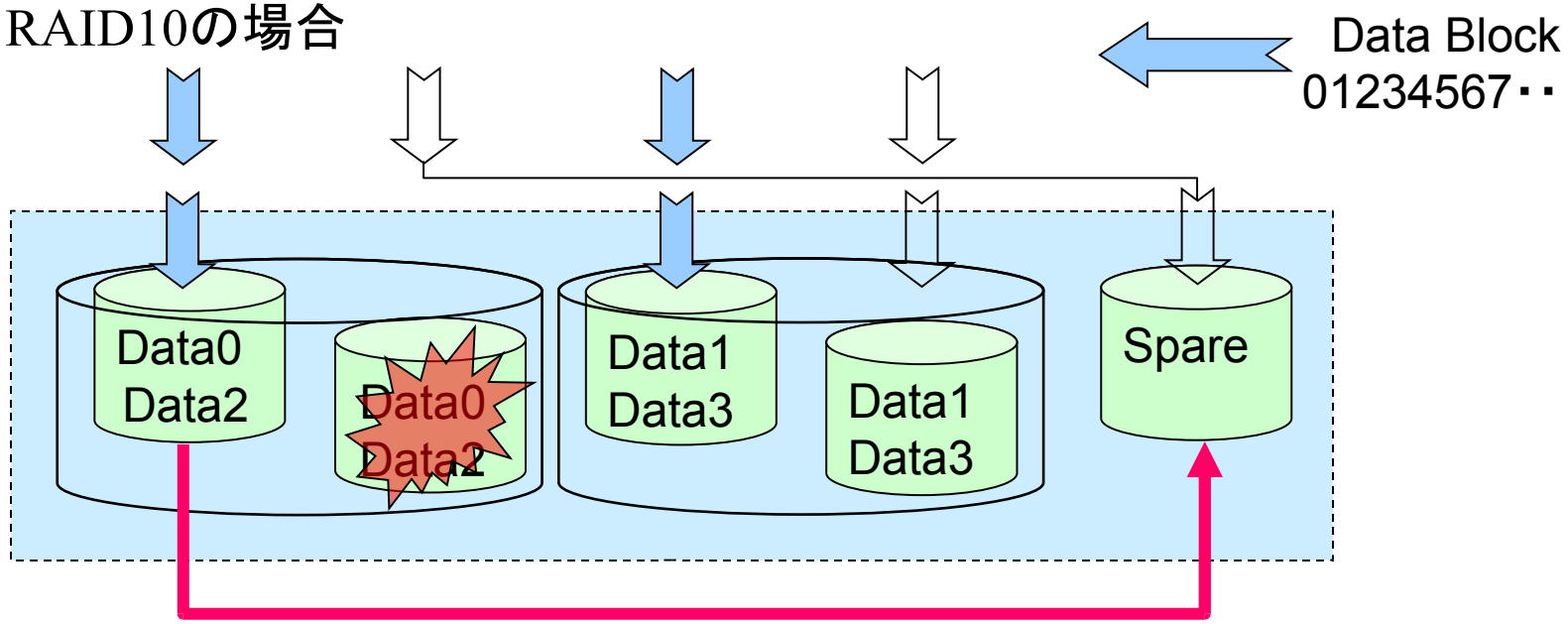


・RAIDレベル5をストライピングしたもの。信頼性を維持しつつI／Oを向上させることが可能。

2. RAID

- ホットスペアの動作

RAID10の場合

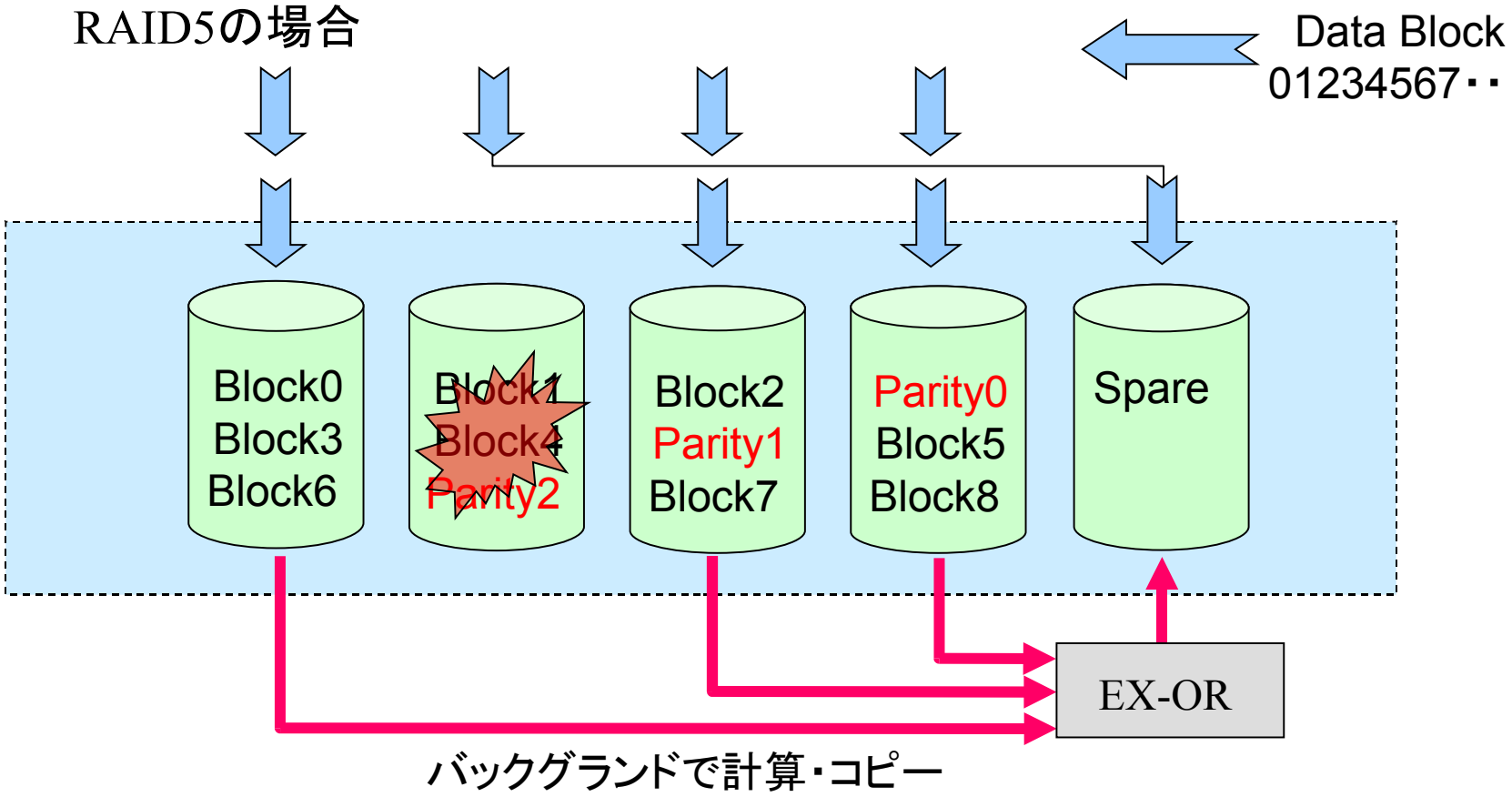


復元時間は機種により異なるが、1GBあたり1分程度

※故障ディスク交換後、書き戻しを行うものや、交換ディスクをスペアディスクとするものなどいくつか種類がある

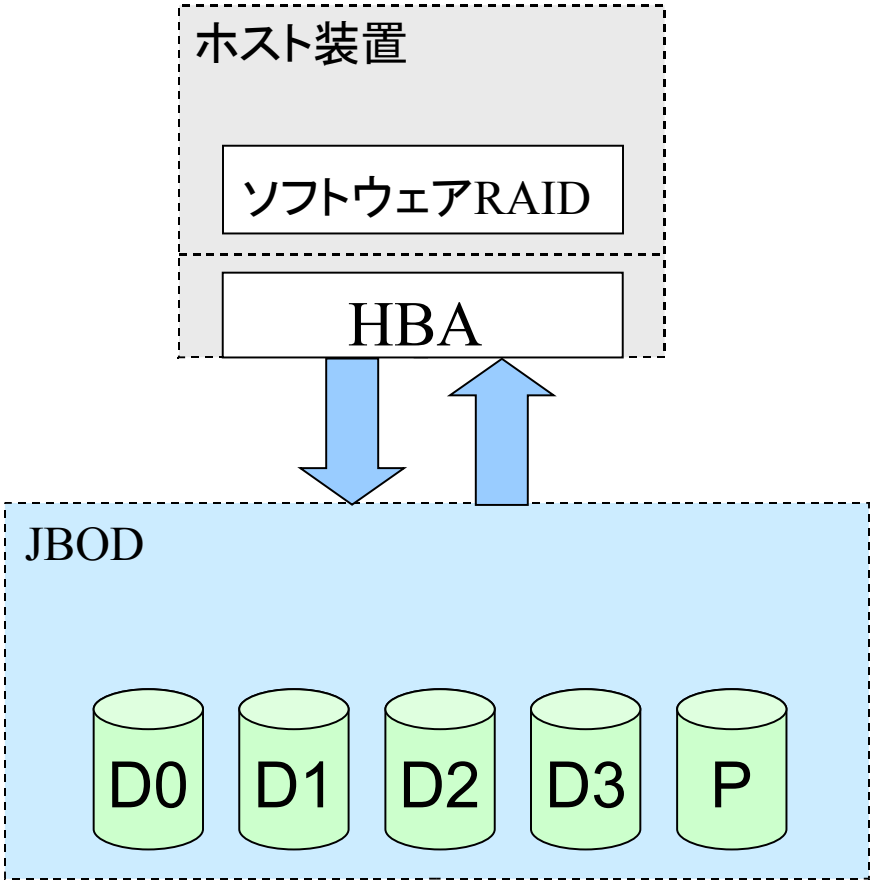
2. RAID

- ホットスペアの動作



2. RAID

- ソフトウェアRAID

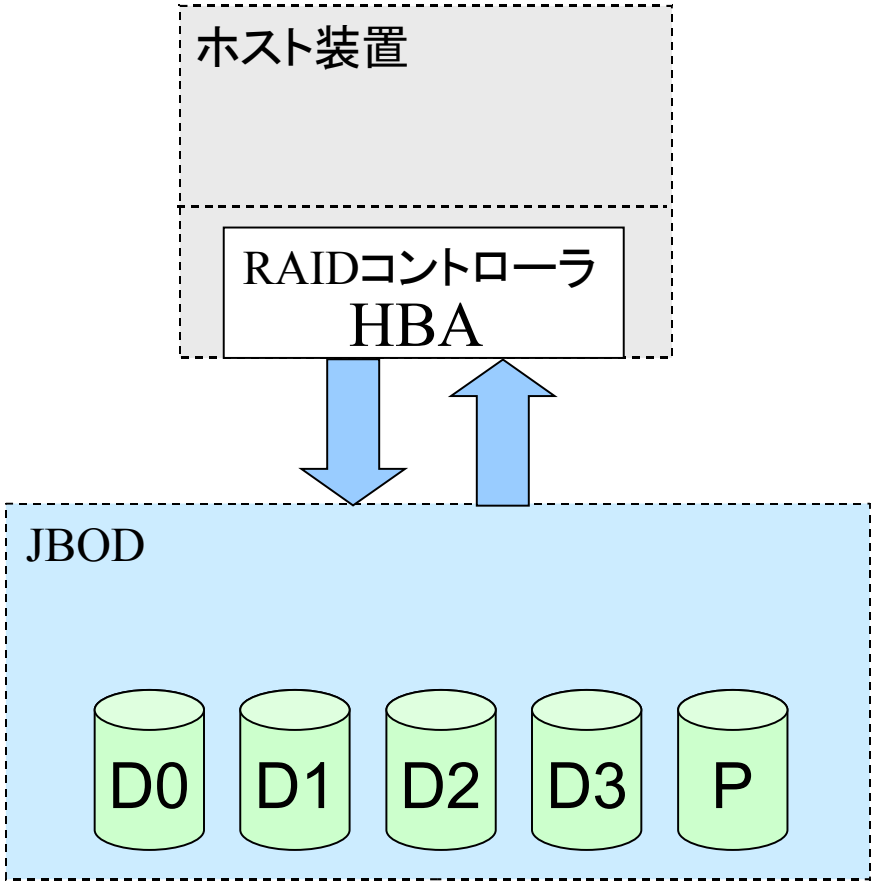


• HBAを内蔵するホストのCPUによりデータのストライピングパリティ計算、ミラー処理、各ドライブとのアクセスを行う。 HBA，JBODシステムがあればできるので比較的安価に構成できるが、CPU負荷がかかる。また、一般的には障害時の復旧手順は複雑。

• クラスターやSAN等の柔軟な構成には対応できない。

2. RAID

- RAIDコントローラ



・HBAにRAID機能を持たせる。ハードウエアにてRAID機能を持っている。比較的低コストであるが、RAIDの制御がOSやホストとの依存性が高いので整合性や障害時の切り分けがやや困難。

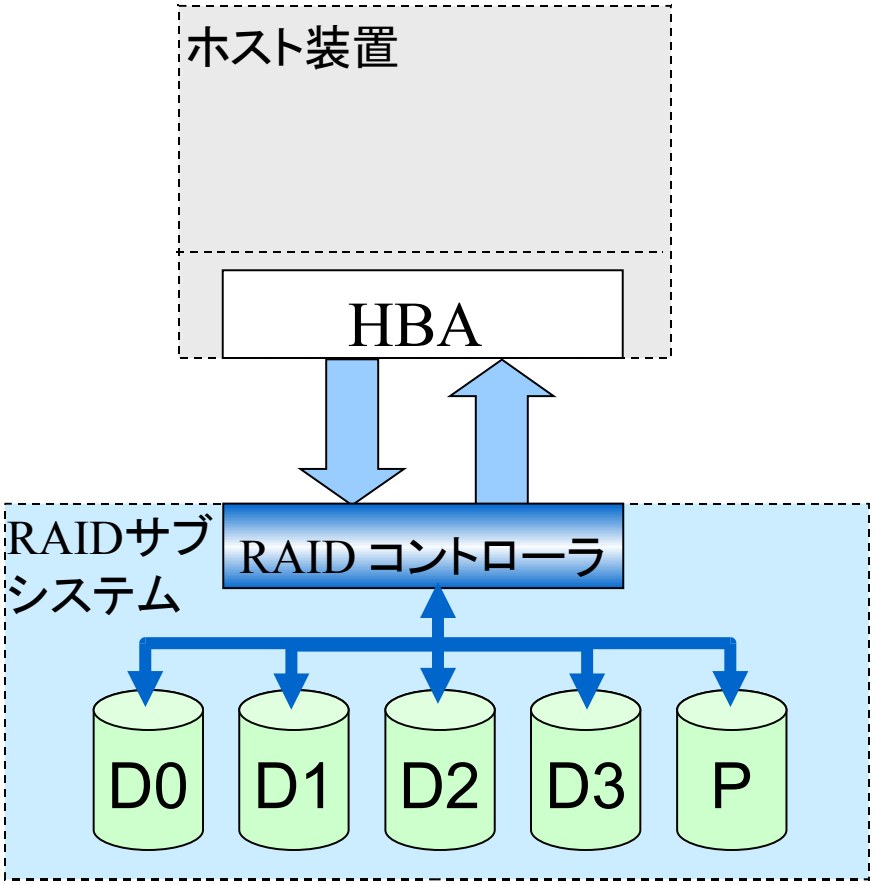
・クラスターやSAN等の柔軟な構成には対応できない。

・IAサーバ等で良く用いられている。

・ホスト等をグレードアップする時はストレージ部分も全部入れ替える必要がある。

2. RAID

- RAIDサブシステム



・ソフトウエアRAID、RAIDコントローラHBAよりも高価。最も柔軟性がある。RAID機能は専用ハードウエアで行われ高速でありホストに負担をかけない。ホストとはインターフェースを介して接続されており、通常は特別なデバイスドライバも必要ないので、障害時の切り分けも容易である。ホストが更新されてもRAID装置は継続して使用できる。

・SANを構成する場合は、RAIDサブシステム型が必要になる。

3．SSD（Solid State Drive：半導体ディスク）

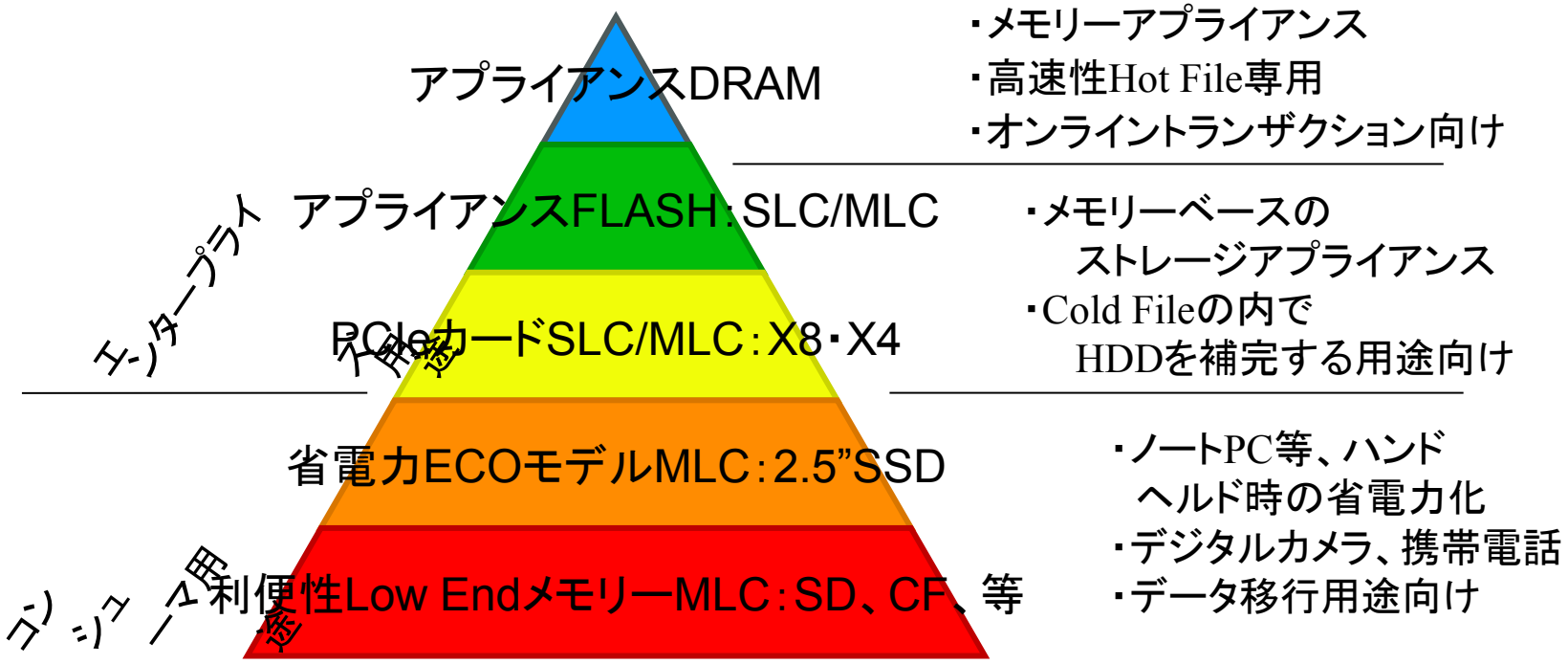
- SSDは、メモリを記録媒体とするドライブ装置。ハードディスクドライブ(HDD)と同じ接続インタフェースを備え、ハードディスクの代替として利用できる他、ハードディスクドライブとは異なるインターフェースを持つ製品もある。

- SSDは、ハードディスクのようにメカ機構を持たないため、ヘッドを移動させる時間(シーク時間)や、対象セクタがヘッド位置まで回転してくるまでの待ち時間(回転待ち時間)がなく、高速に読み書きできる。また、機械的部分がないため消費電力も少なく(IO比で10%以下)、衝撃にも強い。
- 書き換え回数は、HDDに対して少なく、多いもので10万回程度、少ないものだと5000回程度となっている。

- 最近では、Flashを搭載したSSDのことを指すことが多い。

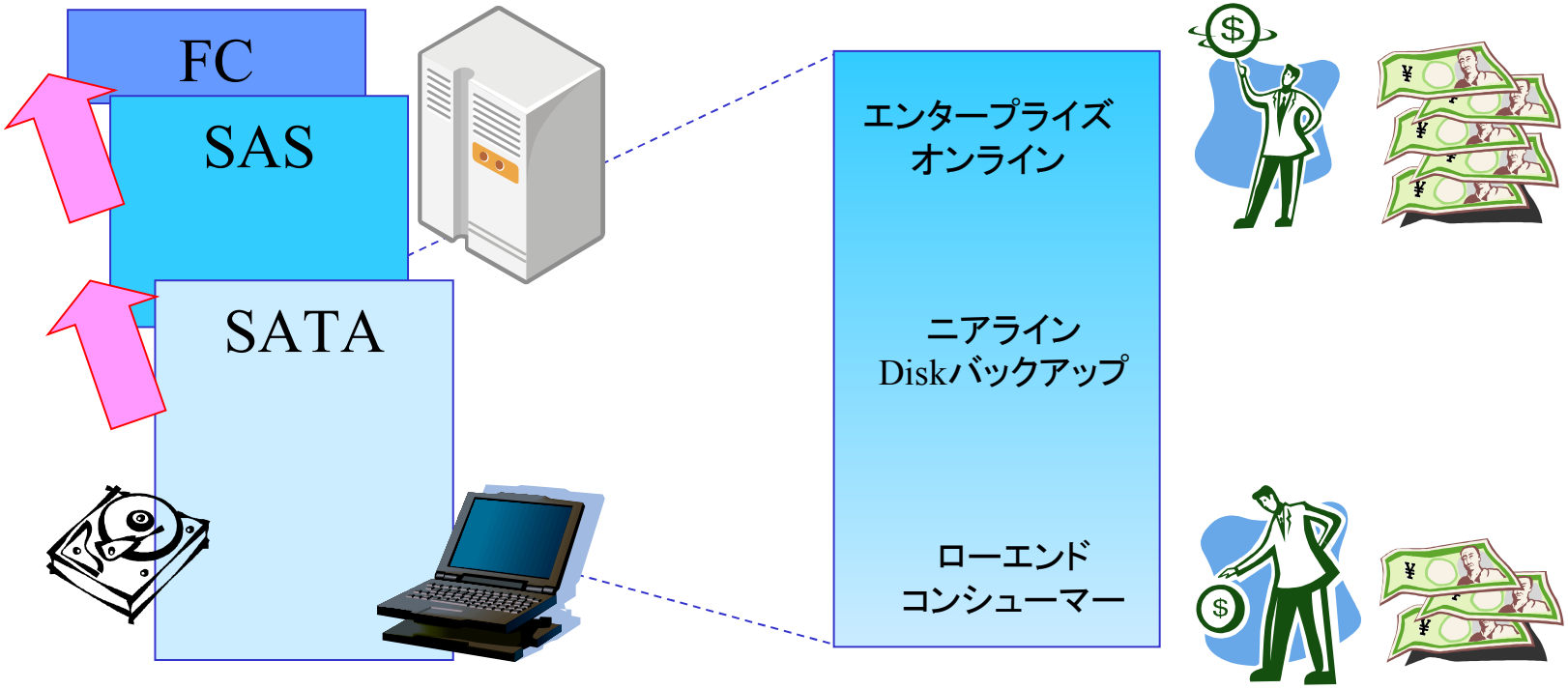
3．SSD（Solid State Drive：半導体ディスク）

- SSDの種類/階層
  - HDDと同様に、SSDにも、用途や特長によって分類がある。



3．SSD（Solid State Drive：半導体ディスク）

- HDDの種類/階層（参考）



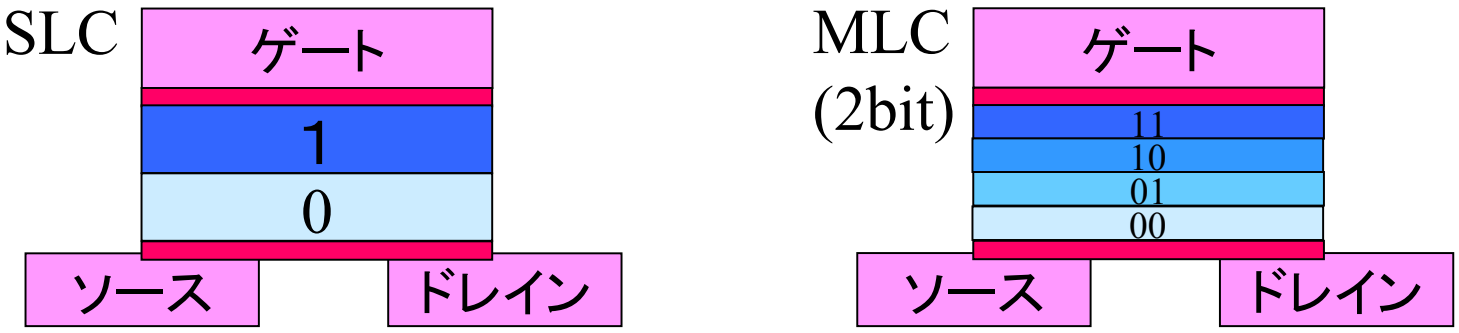
3．SSD（Solid State Drive：半導体ディスク）

- 半導体の種類による違い
  - DRAM/SRAM
    - リード／ライト高速だが、高価
    - バックアップ機構必要
  - Flashメモリ
    - 比較的安価になってきており、ローエンドについては、携帯やデジタルカメラ、NetBookなどで使われてきている
    - また、PC用途やエンタープライズ用途でも普及してきている
    - 一般的にランダムリードは高速だが、シーケンシャルリードは同等、ライトは同等か低速
    - SLC(Single Level Cell)とMLC(Multi Level Cell)(2bit、3bit以上)があり、それぞれ特長がある
    - ウェアレベリング(平均化処理)により書き換え回数が少ない部分を補っている

3．SSD（Solid State Drive：半導体ディスク）

- SLCとMLC

  SLCは1つの記録素子に1ビットのデータを記録するが、MLCは1つの記録素子に2ビット以上のデータを記録する。
  SLCタイプはその書き込み速度と耐性により、サーバ向けや耐久性が求められる分野で使用され、書き込み速度が速い、低消費電力、書き込み耐性が高いなどの特長がある。
  一方、MLCタイプはSLCタイプと比べて書き込み耐性と速度で劣るものの、値段が安く大容量化しやすいという利点がある。



3．SSD（Solid State Drive：半導体ディスク）

- ウェアレベリング

  特定のブロックだけに集中して書き込まないように使用するブロックを分散化させ、特定ブロックの劣化が進んで、寿命が短くなることを抑制する技術。一般的にSLCで～10万回、MLCで～5000回程度と言われている。

- 信頼性向上：アクセス低減

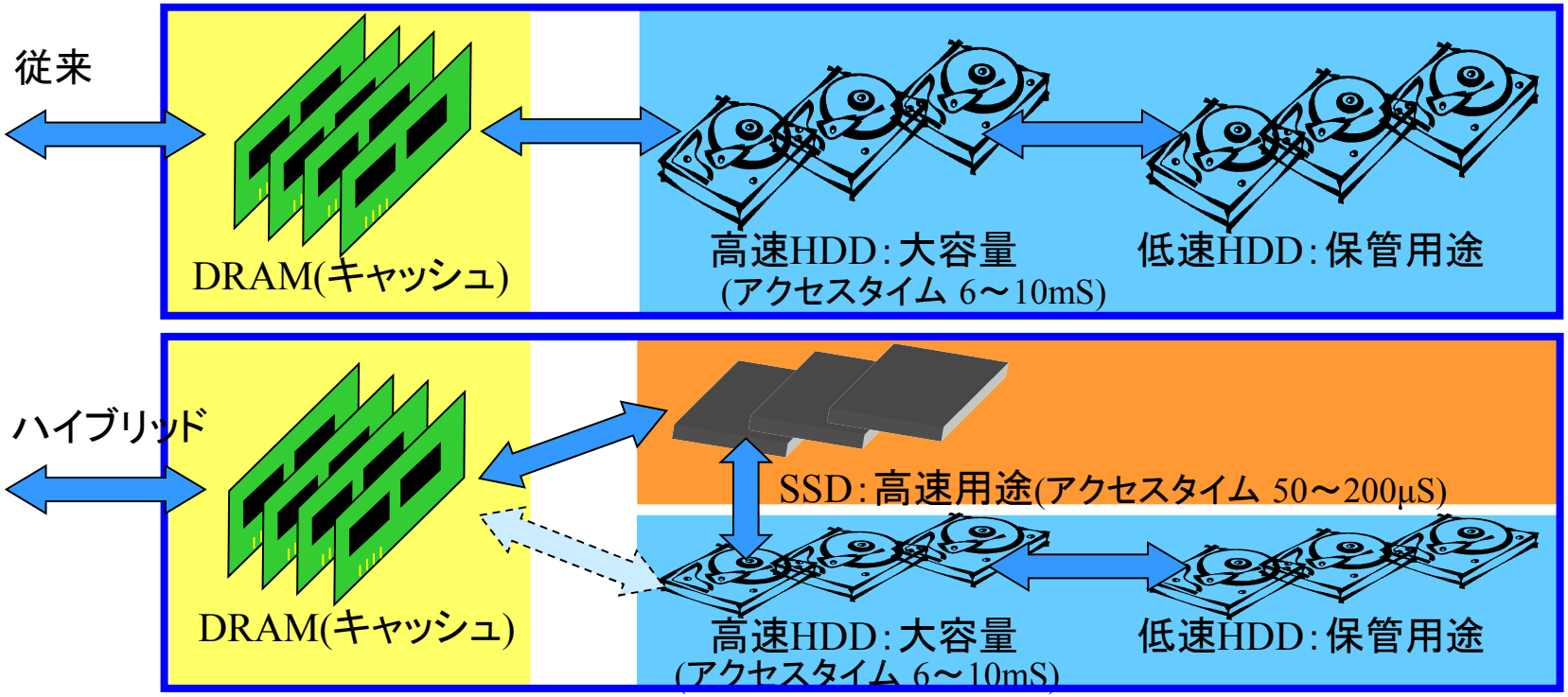
  ウェアレベリング等を行うために、キャッシュメモリを持っているが、このメモリをDRAMのような揮発メモリから大容量の不揮発メモリに変えることで、Flashメモリに対するアクセスを低減する仕組みがでてきている。

- 信頼性向上：予備領域

  基板上のメモリは容易に交換できないため、最初から予備領域を持ったものもでてきている。

3．SSD（Solid State Drive：半導体ディスク）

- HDDとSSDの違い（ハイブリッド利用）
  - HDDとSSDは、用途に応じて使い分けがされている。
    (例：HDDは大容量、保管用途、SSDは高速用途など)



3．SSD（Solid State Drive：半導体ディスク）

- エンタープライズ製品とコンシューマ製品の違い
  - エンタープライズ製品は3年保証を完備している。
    (コンシューマ製品は1年以下)
  - エンタープライズ製品はガベージマネージメントをバックグラウンドで実施しているため、コンシューマ製品で発生するプチフリーズ現象を防ぐことができる。

ご静聴ありがとうございました

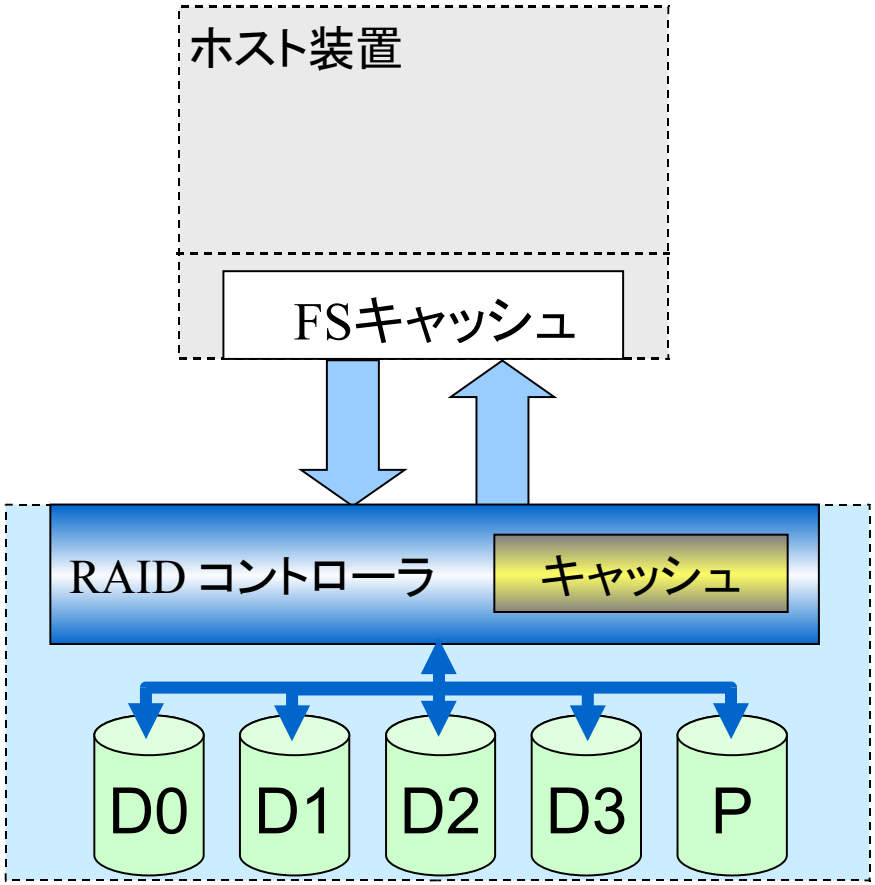
＜休憩＞

4. デバイスの信頼性と冗長

キャッシュメモリとは？

・ここで言うキャッシュとはホスト装置内にあるファイルシステムのキャッシュではなく、ディスクアレイ（RAIDサブシステム）内にあるキャッシュのことである。

・ホスト装置からのリード／ライトアクセスを一時的に受けておくことにより、ホスト装置とディスクアレイ間のI/Oを向上させている。ただし、ライトの場合、キャッシュがデータを受け取ったとき、ホスト装置からは書き込みが完了しているのに対し、HDDへは書かれてない場合がある。従って、キャッシュのデータは故障に対して保全する必要がある。

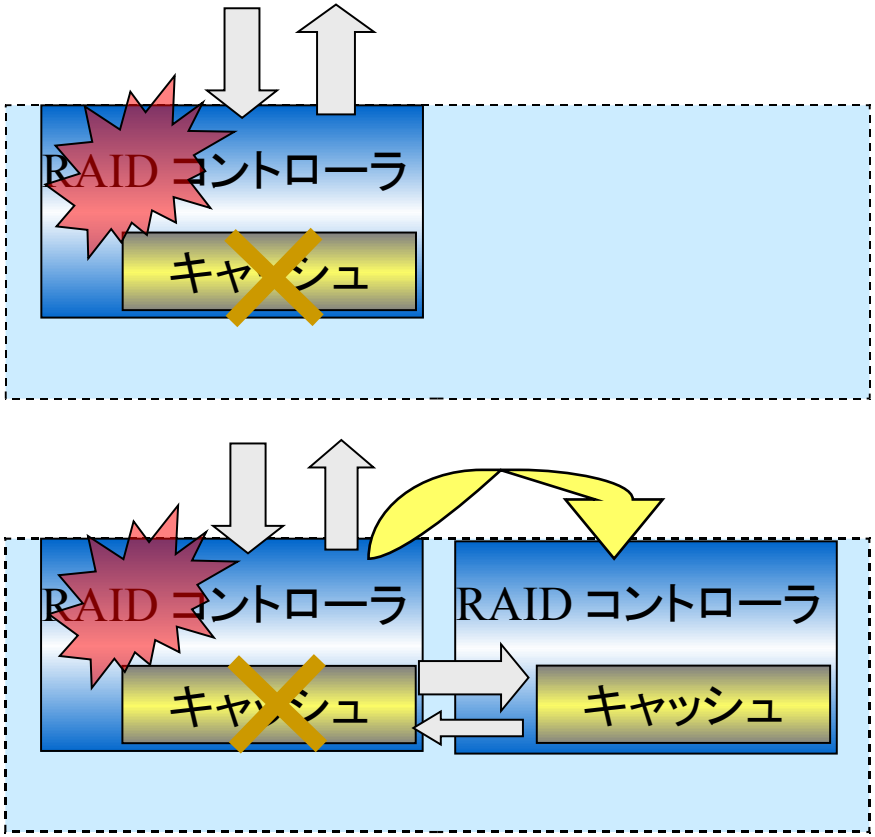


4. デバイスの信頼性と冗長

RAIDコントローラー障害時のデータ保全

・キャッシュ・メモリやRAIDコントローラの故障の時にはキャッシュの内容が消えるため、データの整合性が保証されません。

・対策として、RAIDコントローラを二重化し、キャッシュのデータをミラーリング(同じライト・データを両方のストレージ・プロセッサに存在)し信頼性の向上を図ります。



4. デバイスの信頼性と冗長

停電時のデータ保全

①RAIDサブシステム全体をUPSで電源供給する。

内部対策がなされていない為、UPSの故障や電源ケーブルの抜け、内部電源の故障時にはデータ損失につながります。

②キャッシュメモリーのみを電池で保持する。

最もポピュラーな方式。注意点としては、電池の保持時間内に必ず電源は復旧する必要があります。長時間の停電時には注意が必要です。キャッシュにデータが格納されている状態でコントローラーを交換するとデータ損失につながります。電池の寿命に注意し、寿命が来る前に交換する必要があります。

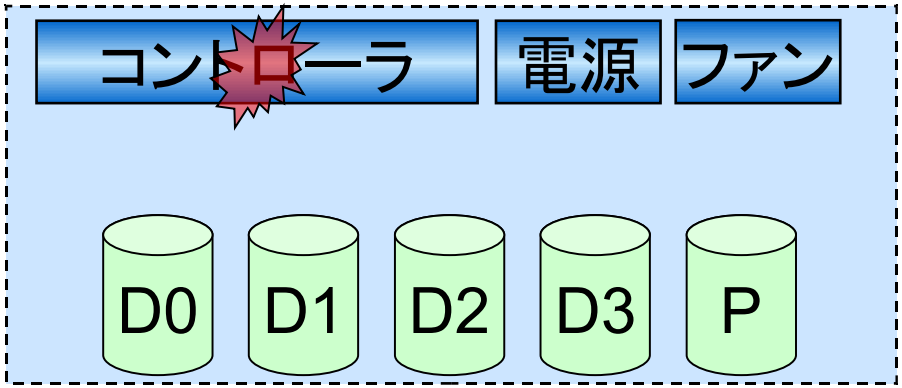
③ディスクに退避する。

停電時には、キャッシュ・メモリからディスク領域に退避します。装置内蔵のバッテリー・バックアップ・ユニットが必要になります。データをディスク・ドライブに退避するので、停電時間を考慮する必要がありません。コントローラが交換されてもデータは保全されます。

4. デバイスの信頼性と冗長

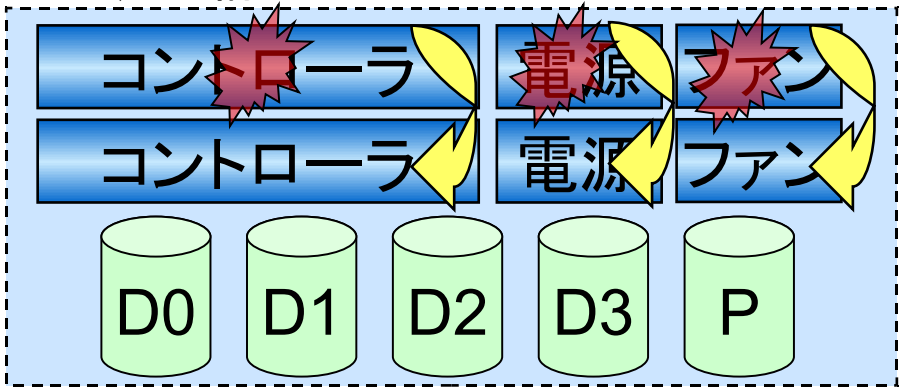
RAIDコントローラや電源、その他構成要素の単一故障によってデバイス全体が障害となってはならない。

・シングル構成



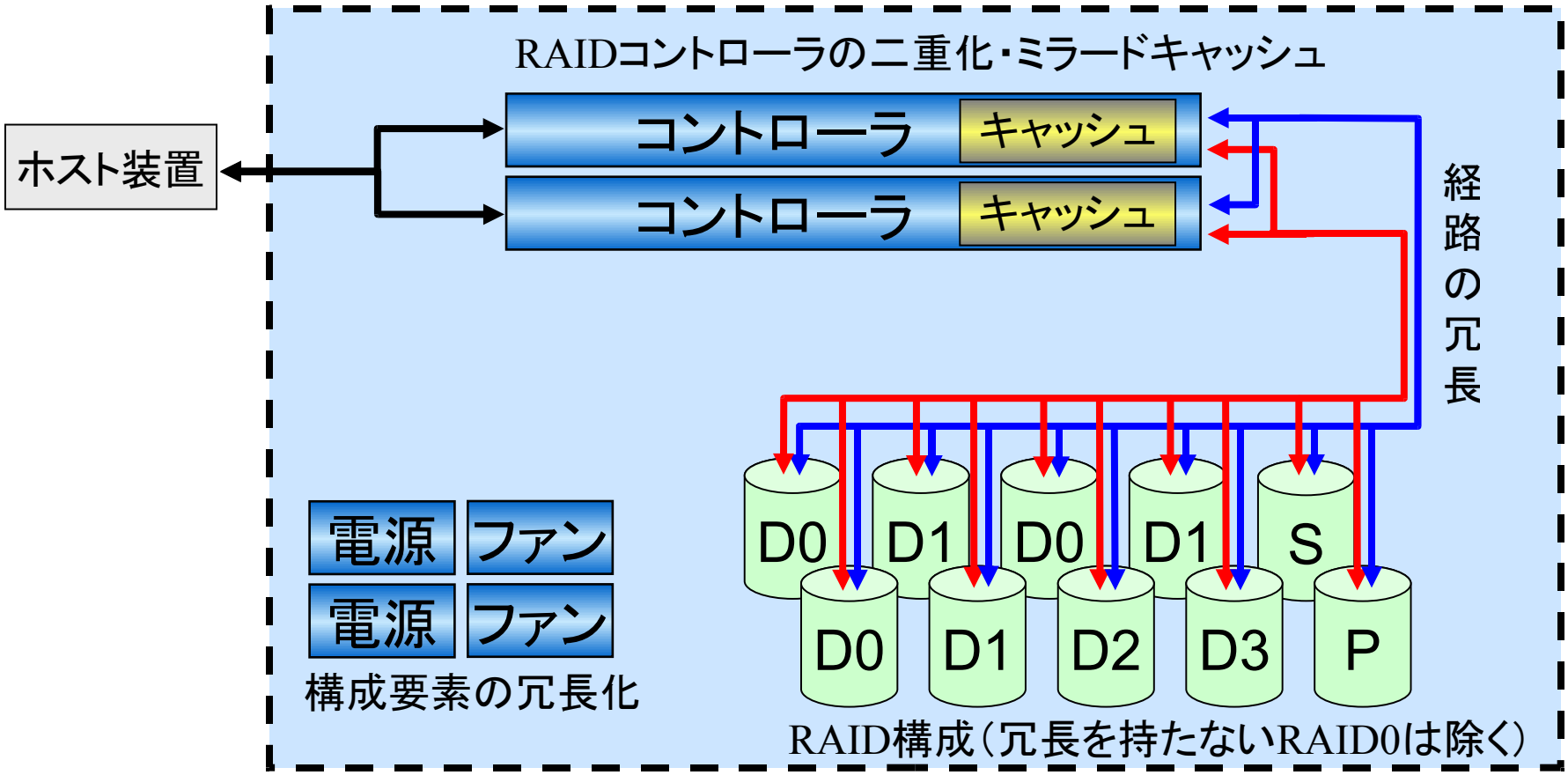
単一の故障で装置が障害となる

・冗長化構成

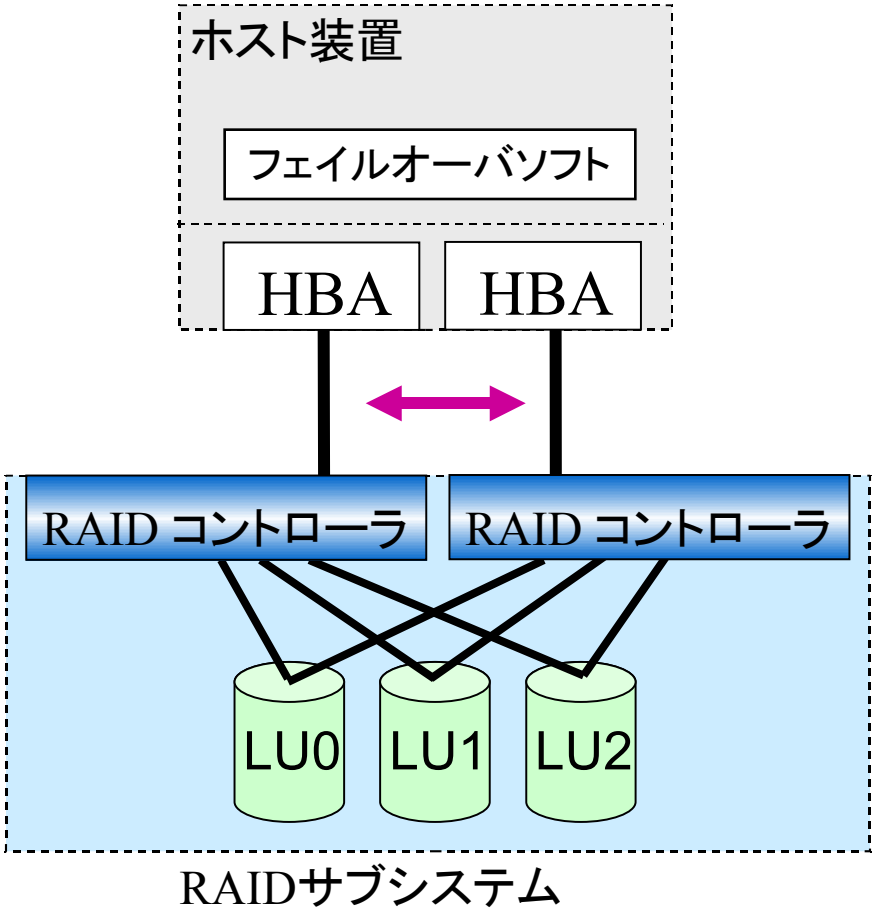


単一の故障でも装置としては障害とならない

4. デバイスの信頼性と冗長

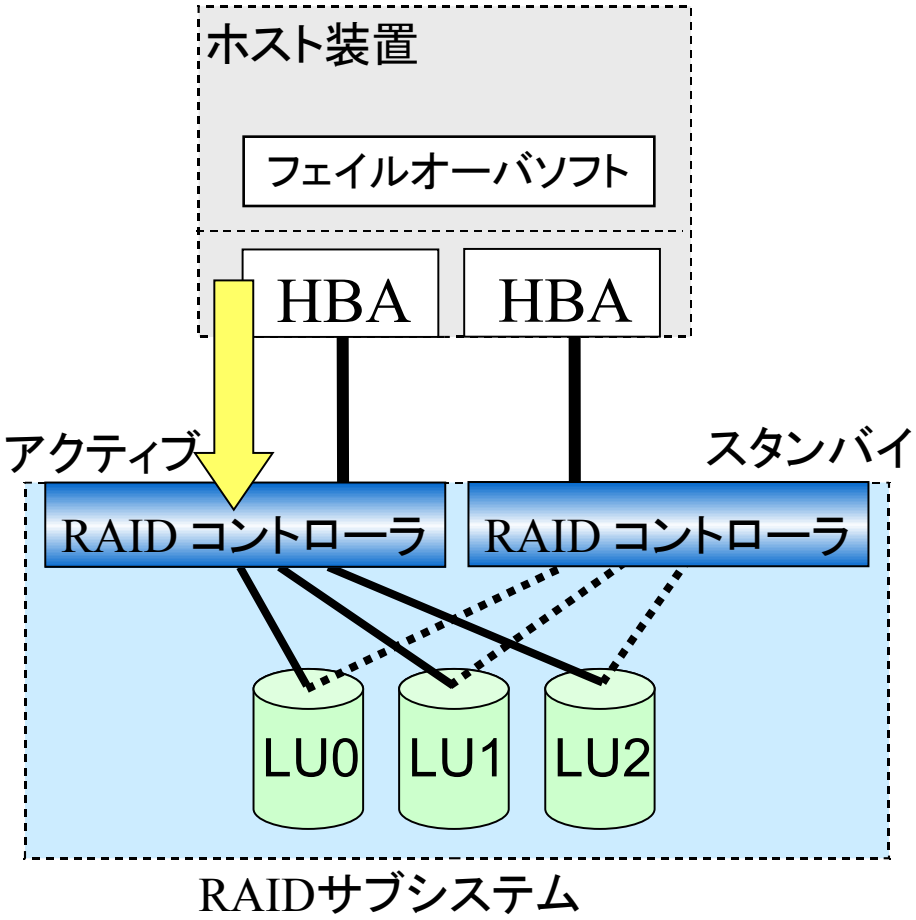


4. デバイスの信頼性と冗長



- ホスト間のバスを二重化することにより、HBA、バス、RAIDコントローラの故障時にも動作継続が可能になります。
- フェイルオーバーソフトは、RAIDサブシステムとの組み合わせで動作します。通常、RAIDサブシステムのベンダーがフェイルオーバソフトも提供します。
- フェイルオーバーソフトは、OSのデバイスドライバレベルと密接な関係で動作します。
- 他のボリュームマネージャーソフトとの相性も確認が必要です。

4. デバイスの信頼性と冗長

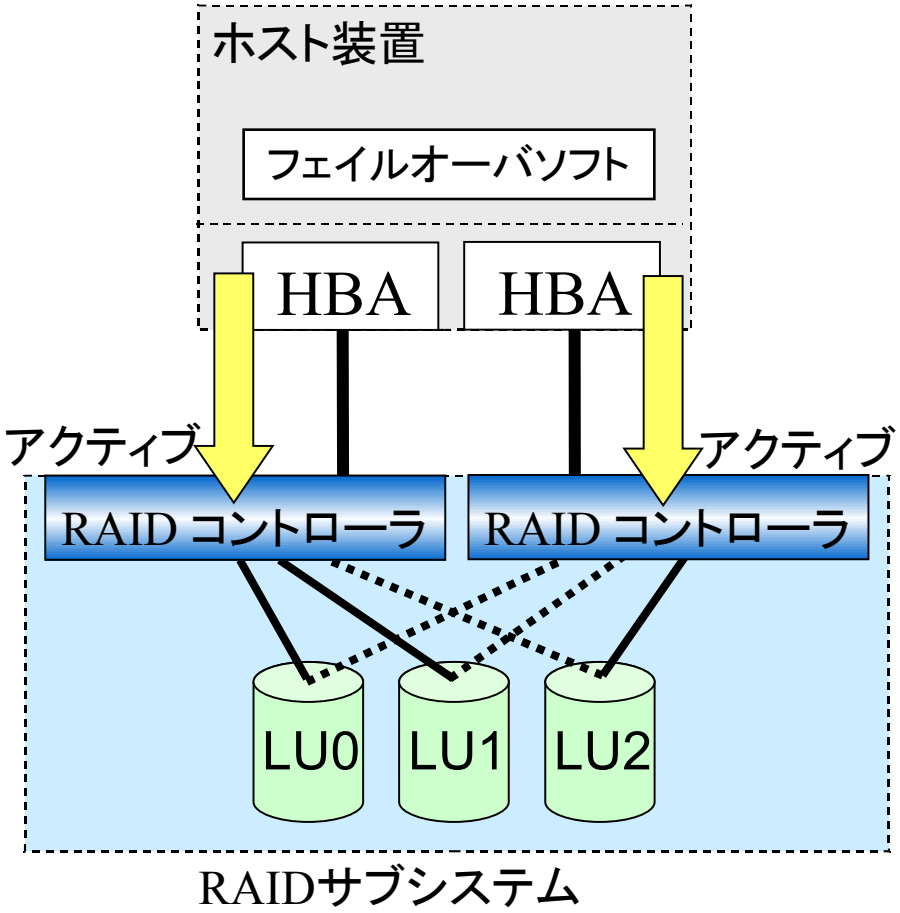


アクティブスタンバイ型

・ 二つあるパスのうち、一つのパスは稼動（アクティブ）状態で、残りのパスは待機（スタンバイ）状態の形態を、アクティブスタンバイ型と言います。

・ RAIDサブシステムによっては、アクティブ・スタンバイ型しかサポートされていない場合もあります。

4. デバイスの信頼性と冗長



アクティブアクティブ型

- 両方のパスが稼動状態の形態をアクティブ・アクティブ型といいます。
- アクティブスタンバイ型と比較し、パスの資源を有効に使用でき、負荷分散が可能です。