JAVCOM技術研究委員会1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1回～

平成21年1月27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. HDD
2. インターフェース
3. RAID
4. デバイスの信頼性と冗長
5. SSD(半導体ディスク)
6. DAS・NAS・SAN

＜休憩10分＞

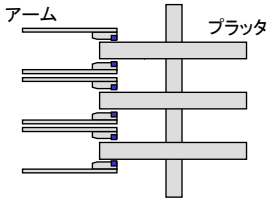
Ⅱ. 映像分野におけるストレージの使い方（Q&A方式）

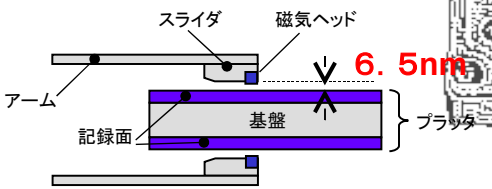
Ⅲ. ストレージ応用及び最新の技術動向（Q&A方式）

Japan Data Storage Forum

1. HDD

- HDDの構造





アーム

プラッタ(磁気ディスク)

インターフェース

裏側

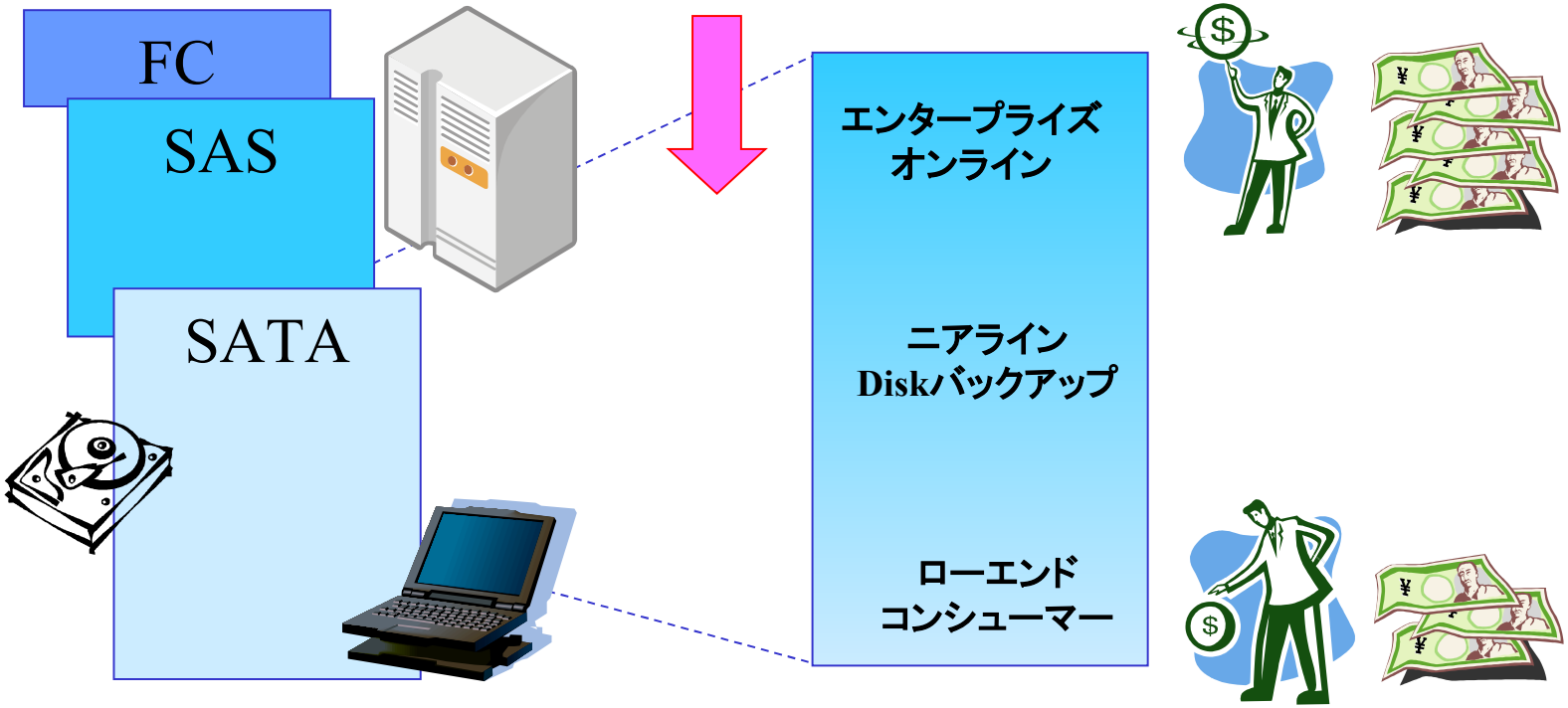
コントローラ基板

磁気ヘッド

モーター

1. HDD

- HDDの用途と違い
- 高いHDDは壊れにくいか

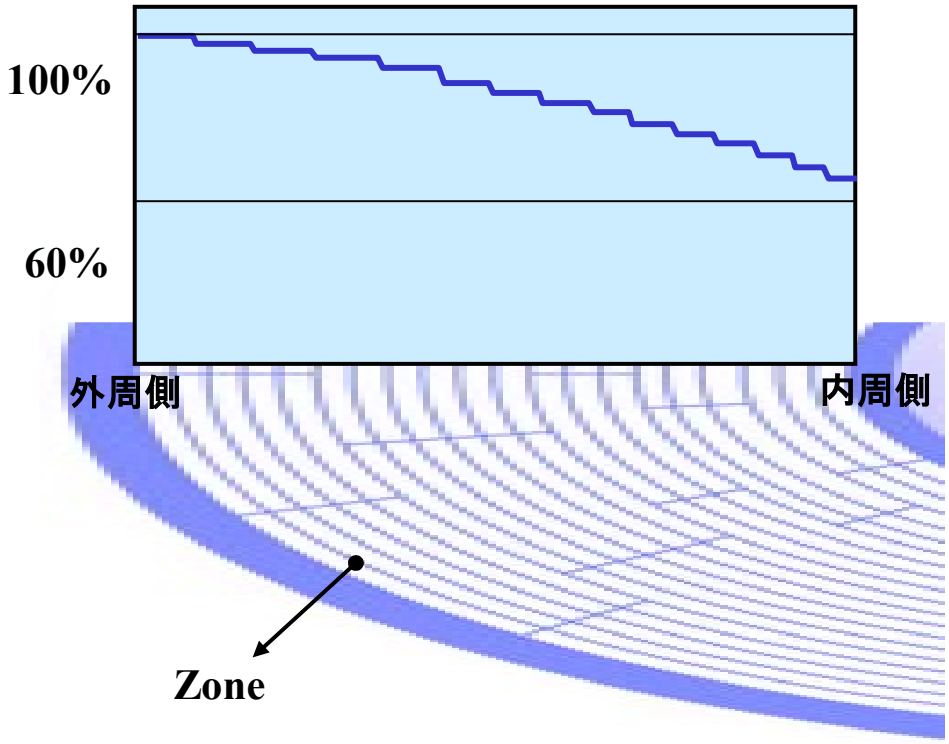


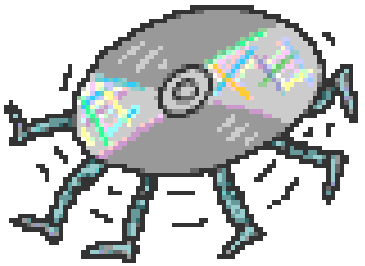
1. HDD

- 回転速度の違いと性能差

  アクセス時間=シーク時間＋平均回転待ち時間＋データ転送時間

- 外周と内周の性能差





1. HDD

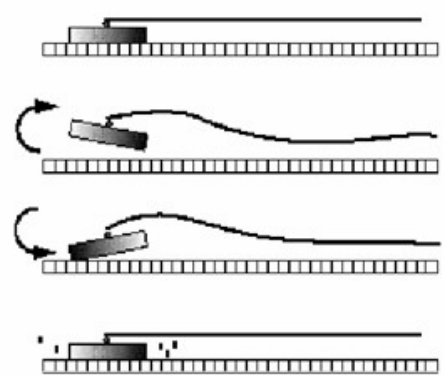
- 熱、振動、塵埃に弱い

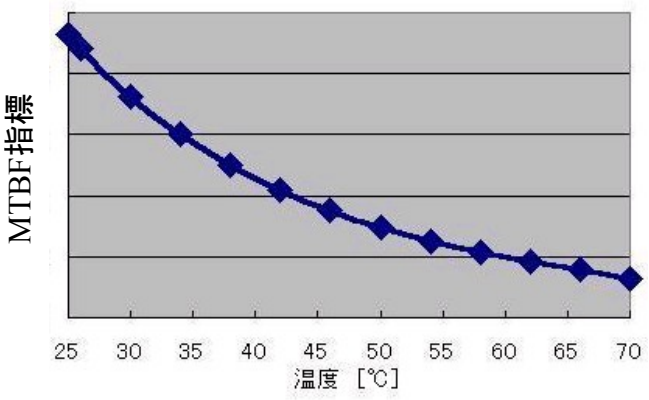
・磁気ディスクは特定の振動でヘッドクラッシュやオフトラックを起こす
（ある程度の振動は、オフトラック防止機能で防ぐことができます）

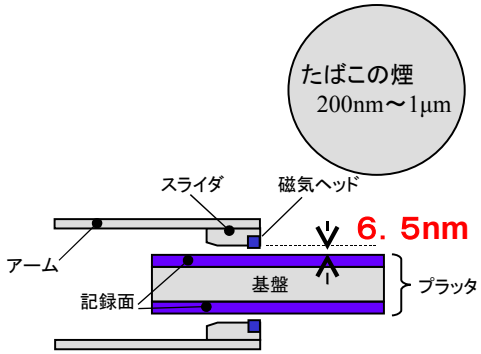
オフトラック防止機能：ヘッド位置がずれた場合、READ／WRITE作業を一時中断する機能

・一般的に、10℃上昇で寿命は半分となる(アレニウスの法則)

・塵埃がプラッタを傷つける可能性がある

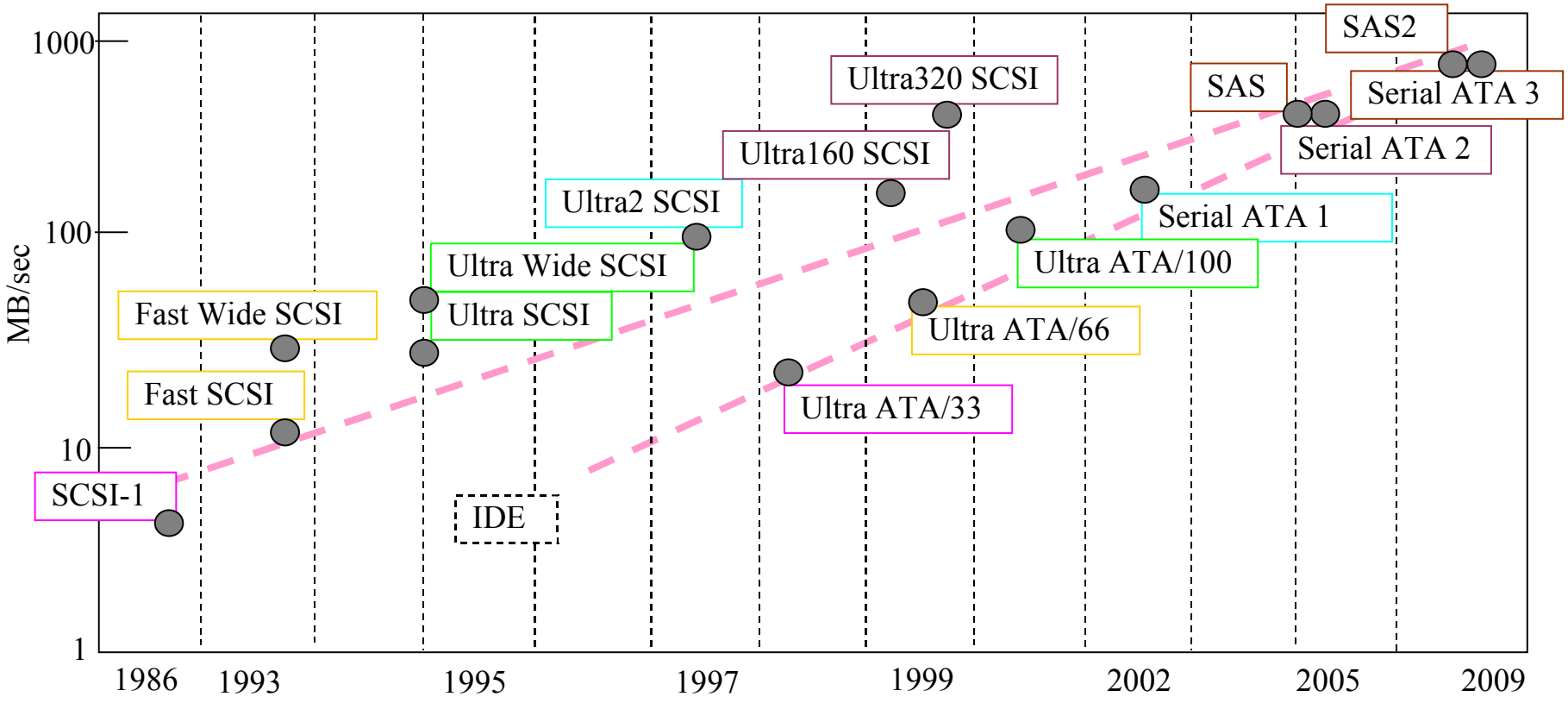






1. HDD

- HDDインターフェースの歴史



1. HDD

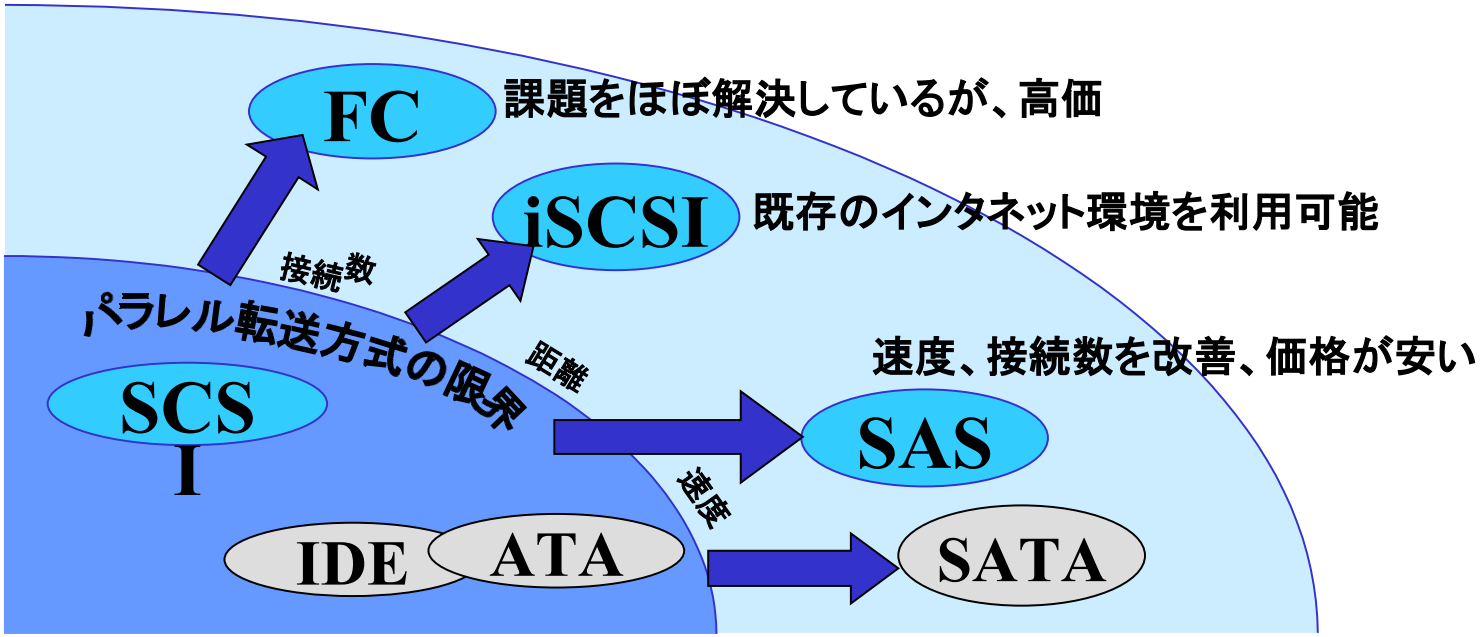
- SMART(Self-Monitoring, Analysis and Reporting Technology)
  HDD故障の早期発見、故障の予測を目的として、さまざまな検査項目(エラー回数や温度など)を診断、収集する仕組み。

  例えば・・・こんなことがわかります。

  ・ハードディスクが回転を開始してから規定の回転数に達するまでにかかった平均時間
  ・工場出荷状態からのハードディスクの通電時間の合計
  ・ハードディスクの電源をON/OFFした回数
  ・オフトラックの数。数値が0でなければバックアップを取る
  ・ロード/アンロード機構によって磁気ヘッドが磁気ディスク表面から退避場所に退避し、その後再び磁気ディスク表面に戻った回数の合計
  ・ハードディスクの現在の温度。一般的に動作が保障されている最高温度は55℃である
  ・磁気ヘッドの浮上高
  ・データの書き込み中に加わった大きな衝撃を表す
  ・電源を抜くなどしてハードディスクが強制的に停止し、磁気ヘッドが緊急退避した回数

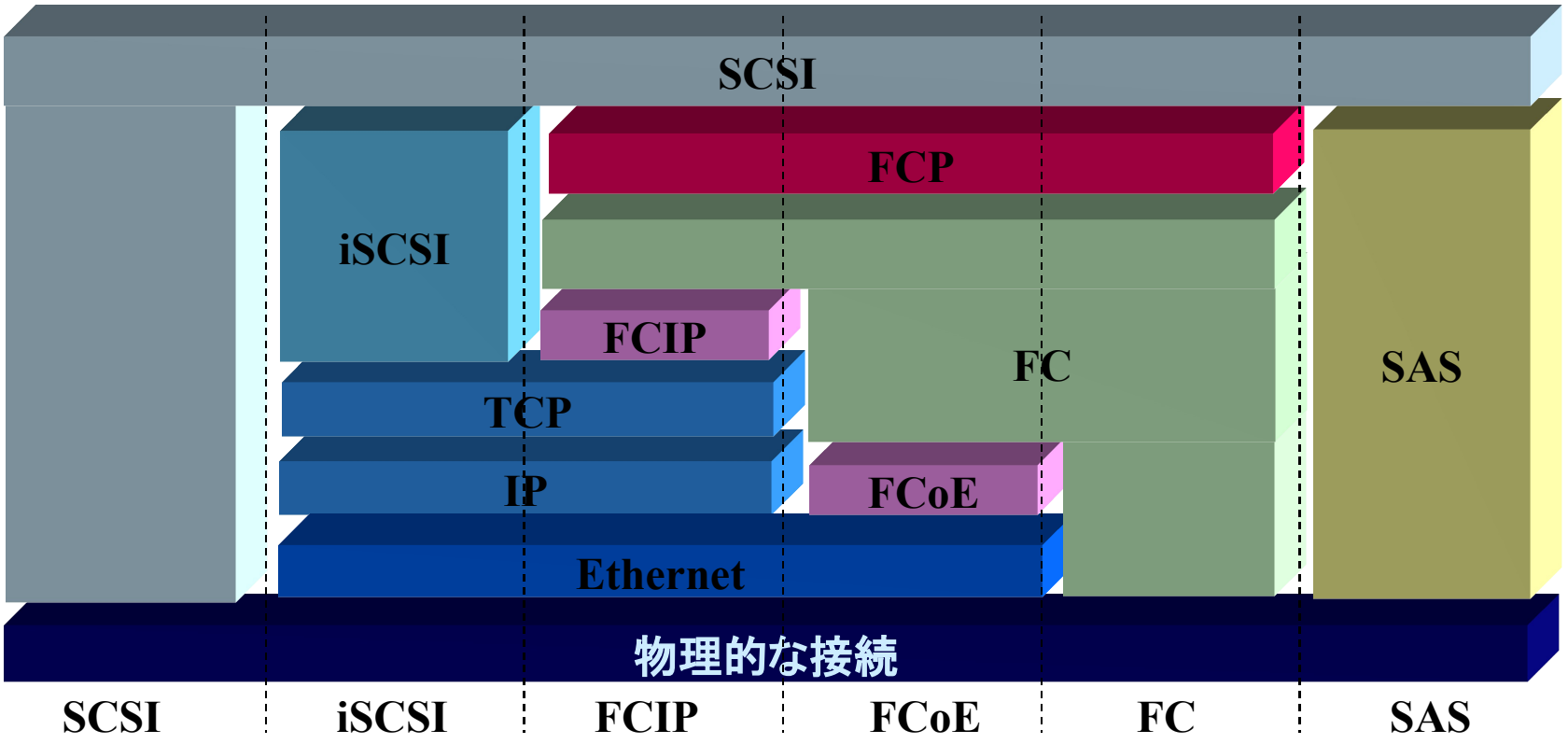
2. インターフェース

- 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**



2. 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**



2. インターフェース

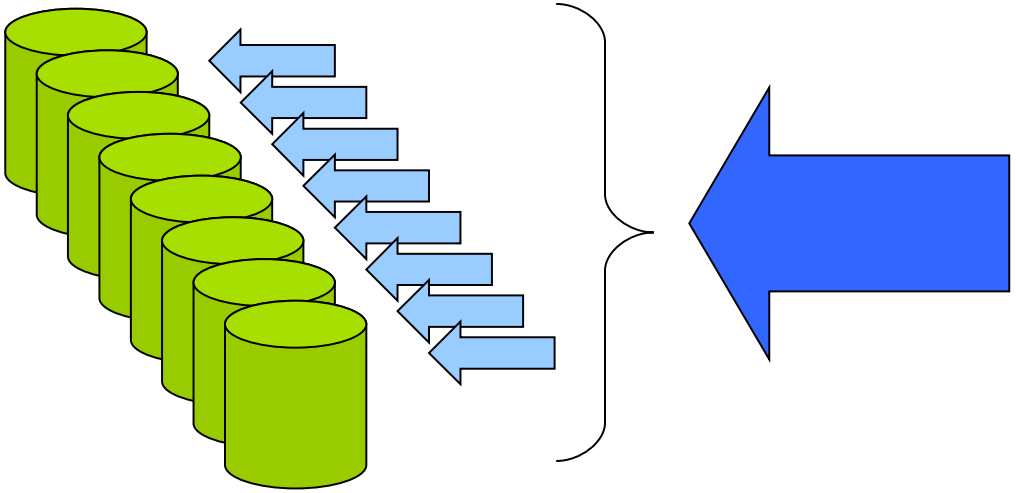
- HBA、SAN-Switchラインナップ

3. 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。

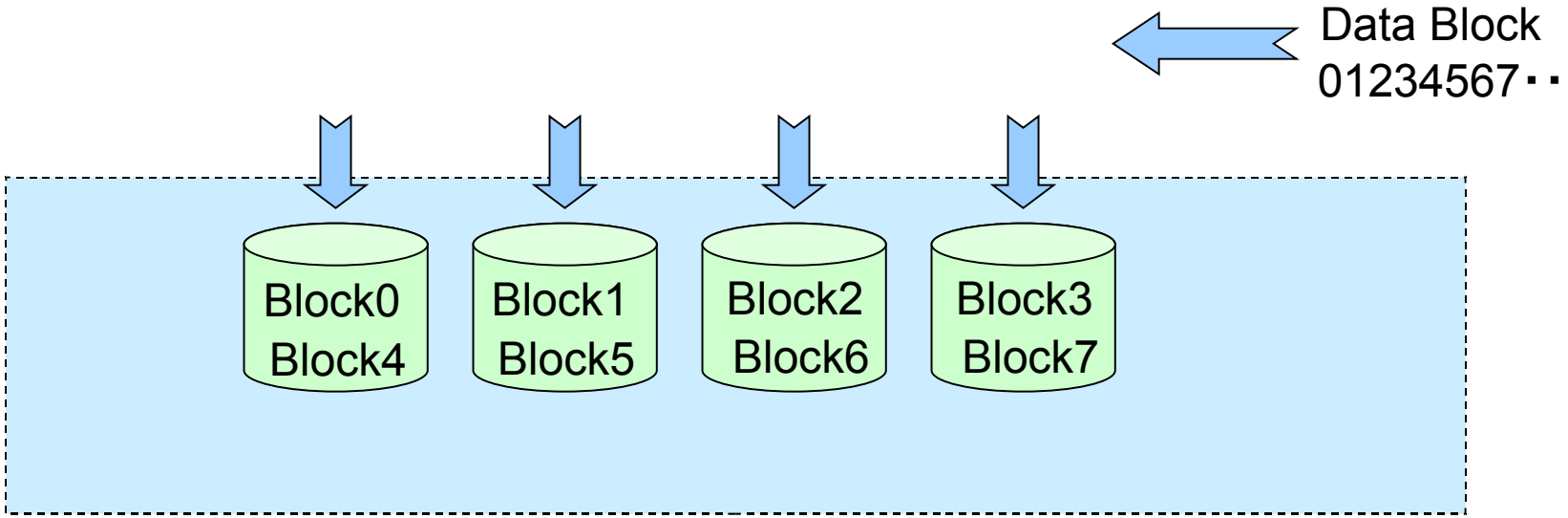


3. RAID

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。	◎	○	○

3. RAID

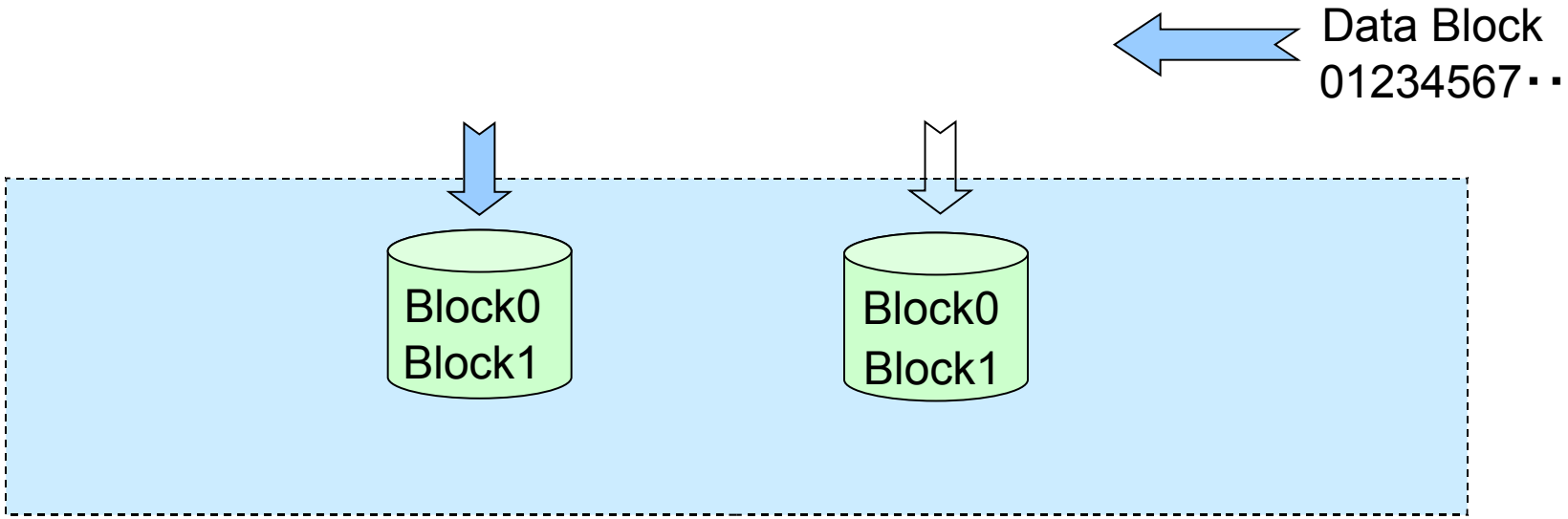
RAIDレベル0：ストライピング



・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。

3. RAID

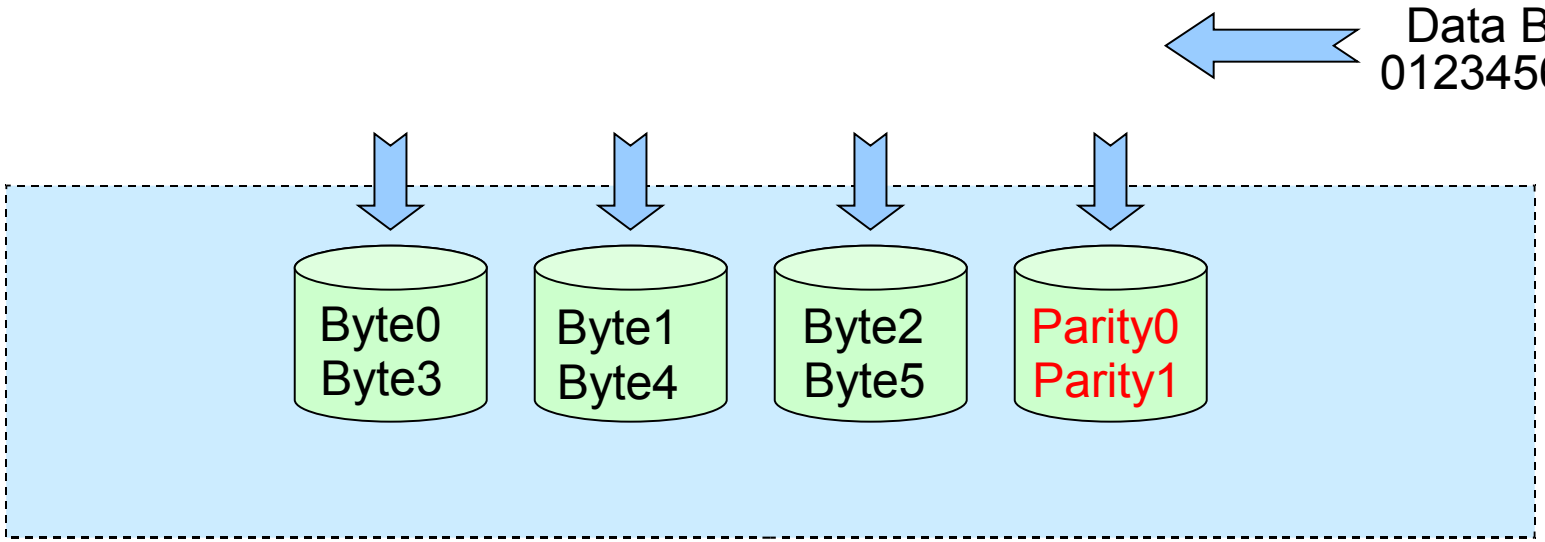
RAIDレベル1:ミラーリング



・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。

3. RAID

RAIDレベル3

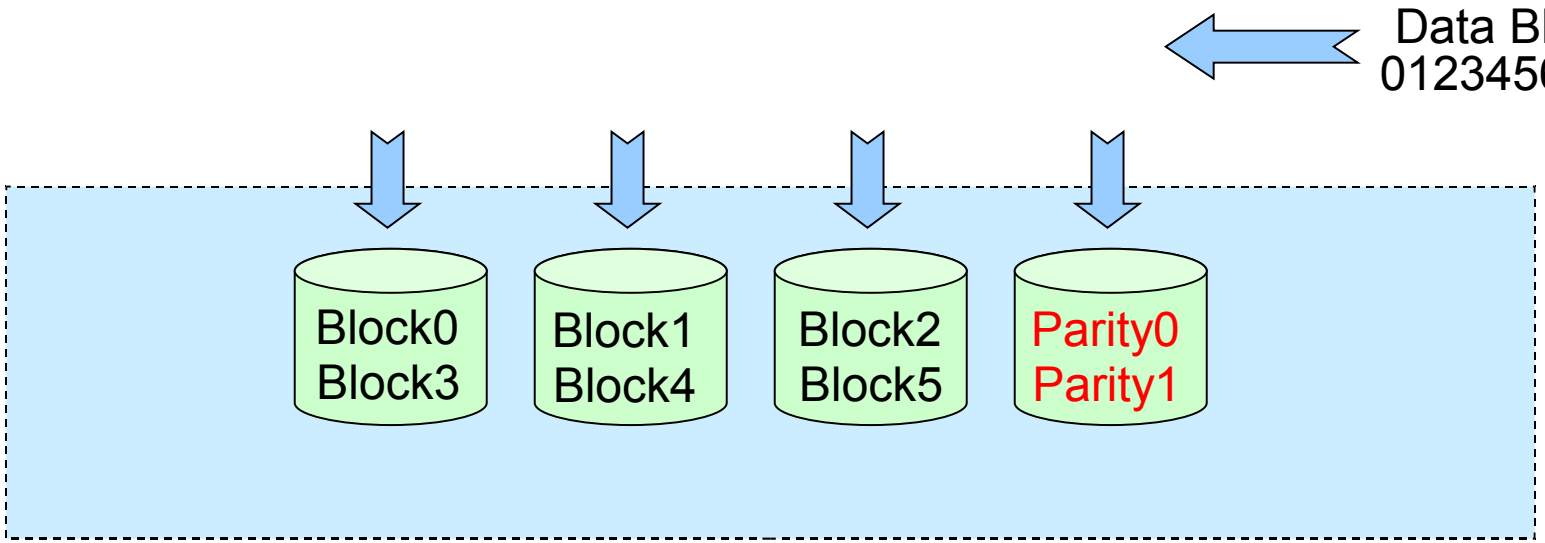


・データはバイトに分割され、各ドライブに分散される。パリティは特定のドライブに置かれる。

・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。

3. RAID

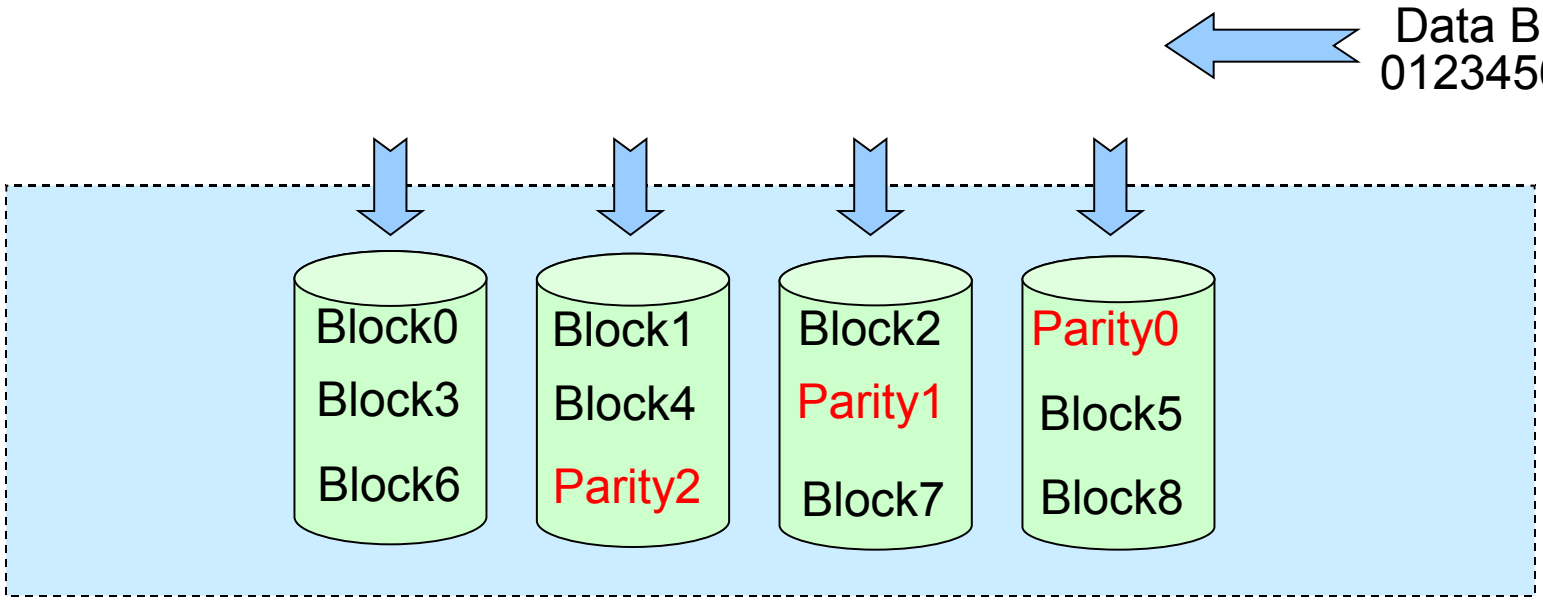
RAIDレベル4



・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。

3. RAID

RAIDレベル5：パリティ付きストライピング

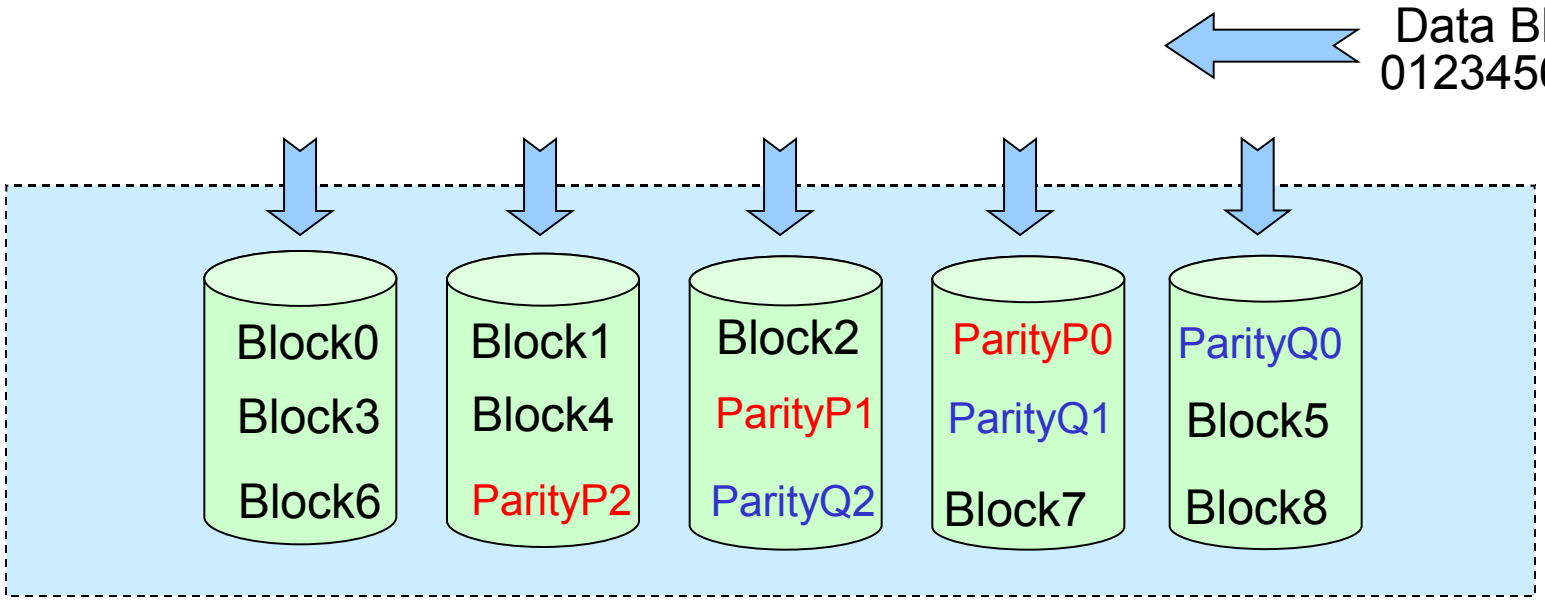


・データはブロックに分割され、各ドライブに分散される。パリティも各ドライブに分散して置かれる。

・小さなデータを頻繁に I／Oする場合に適している。

3. RAID

RAIDレベル6：パリティ付きストライピング

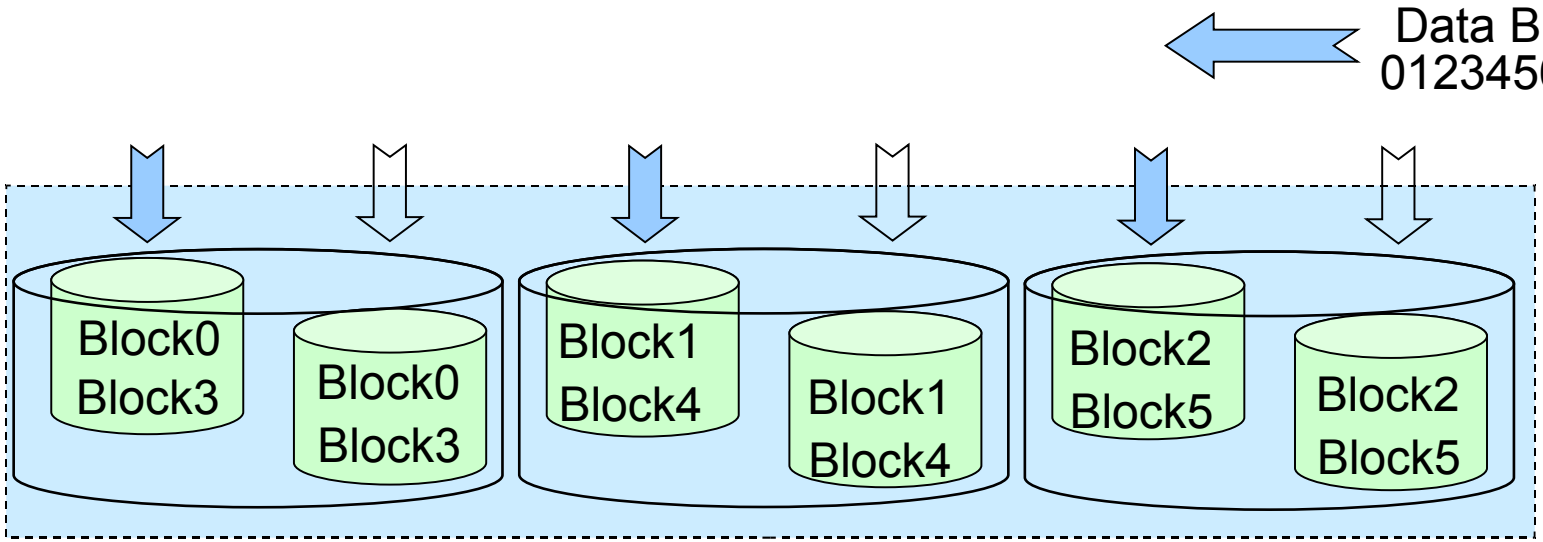


・データはブロックに分割され、各ドライブに分散される。パリティも各ドライブに分散して置かれる。

・ PQ方式、2次元パリティ、対角パリティがある。

3. RAID

RAIDレベル10：ストライピング・ミラー

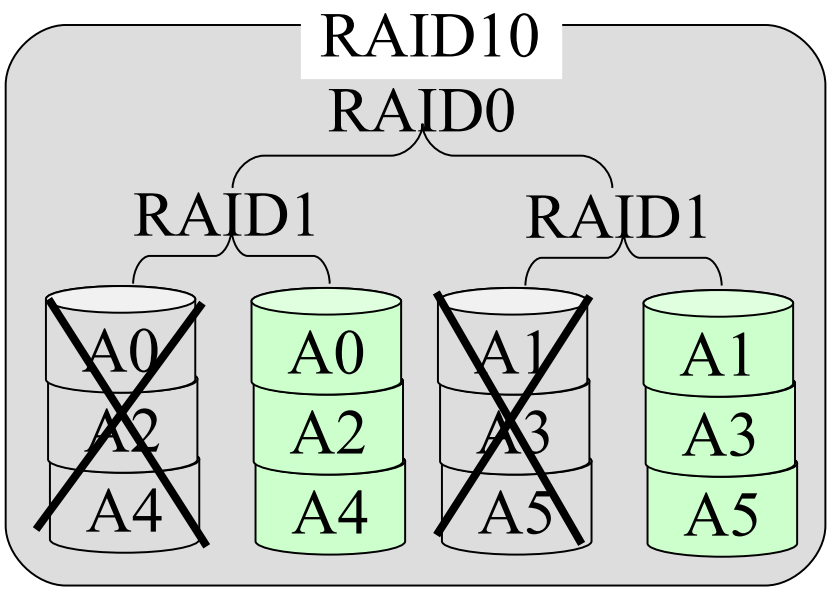


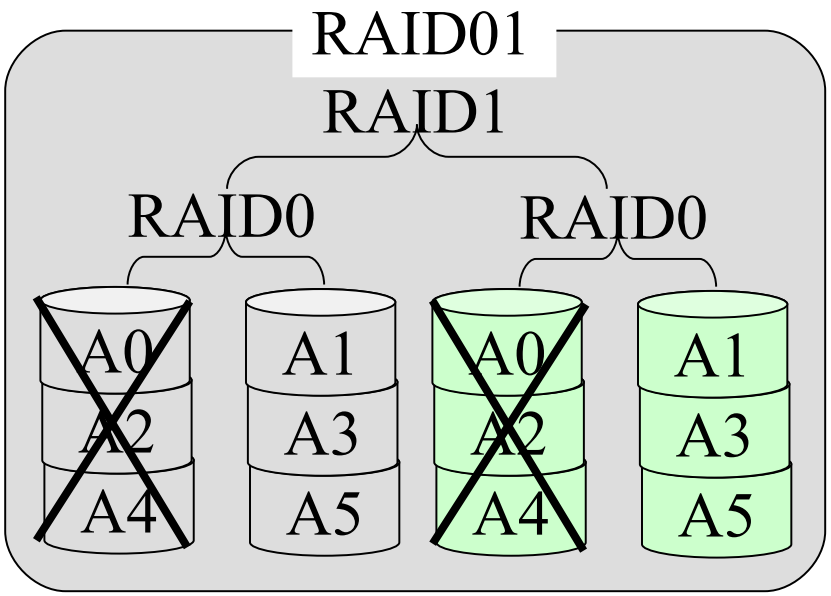
・データは分割され、各ドライブに分散され、別ドライブに同じデータが格納される。

・実容量は物理容量の半分。高速なI／Oが可能。冗長性あり。

3. RAID

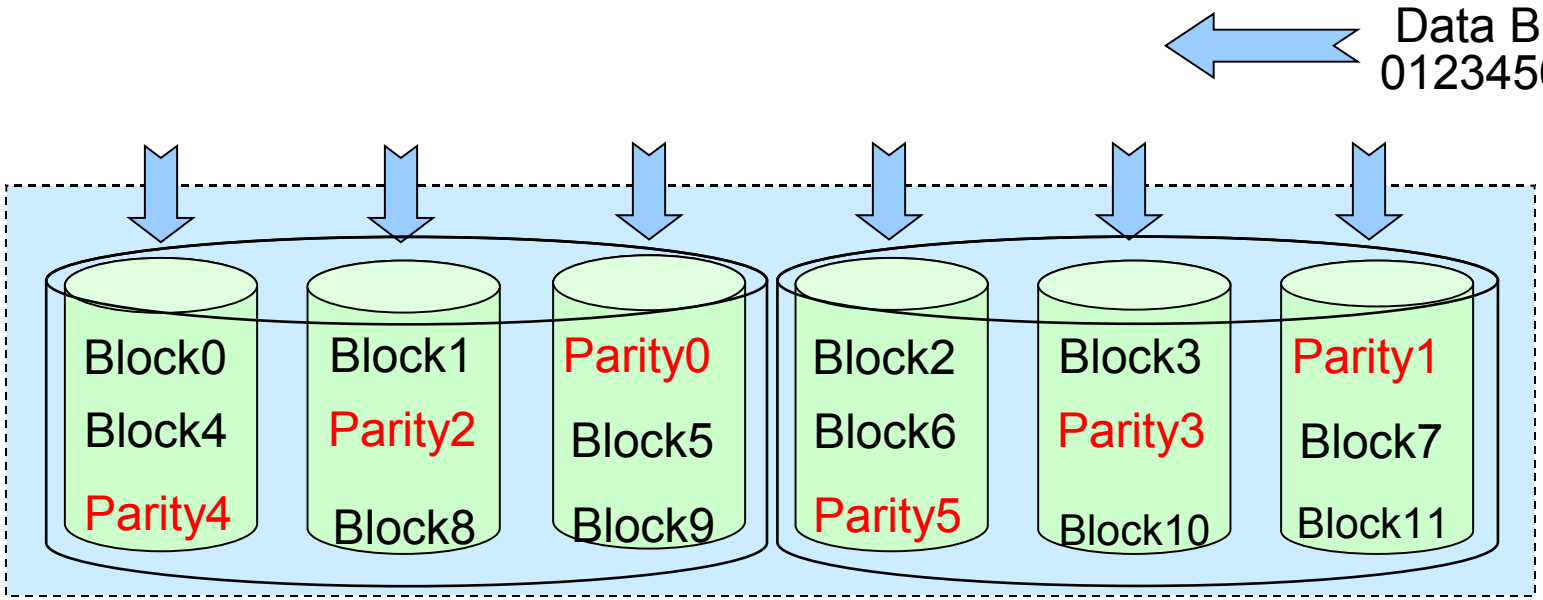
- RAID10(1+0)とRAID01(0+1)の違い
  - RAID10はミラーリングしたアレイをストライピング、RAID01はストライピングしたアレイをミラーリング
  - HDD故障時のデータ保全性が異なります





3. RAID

RAIDレベル50

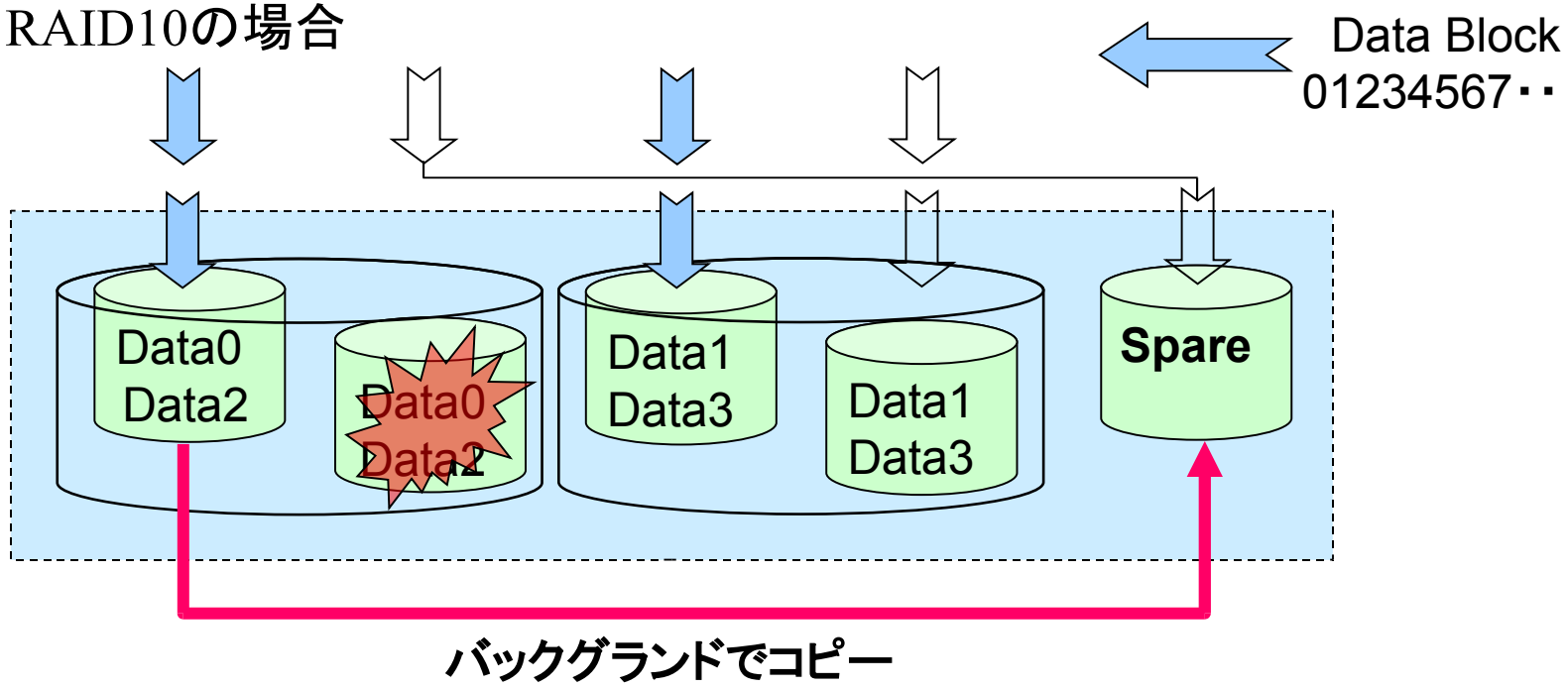


・RAIDレベル5をストライピングしたもの。信頼性を維持しつつI／Oを向上させることが可能。

3. RAID

- ホットスペアの動作

RAID10の場合



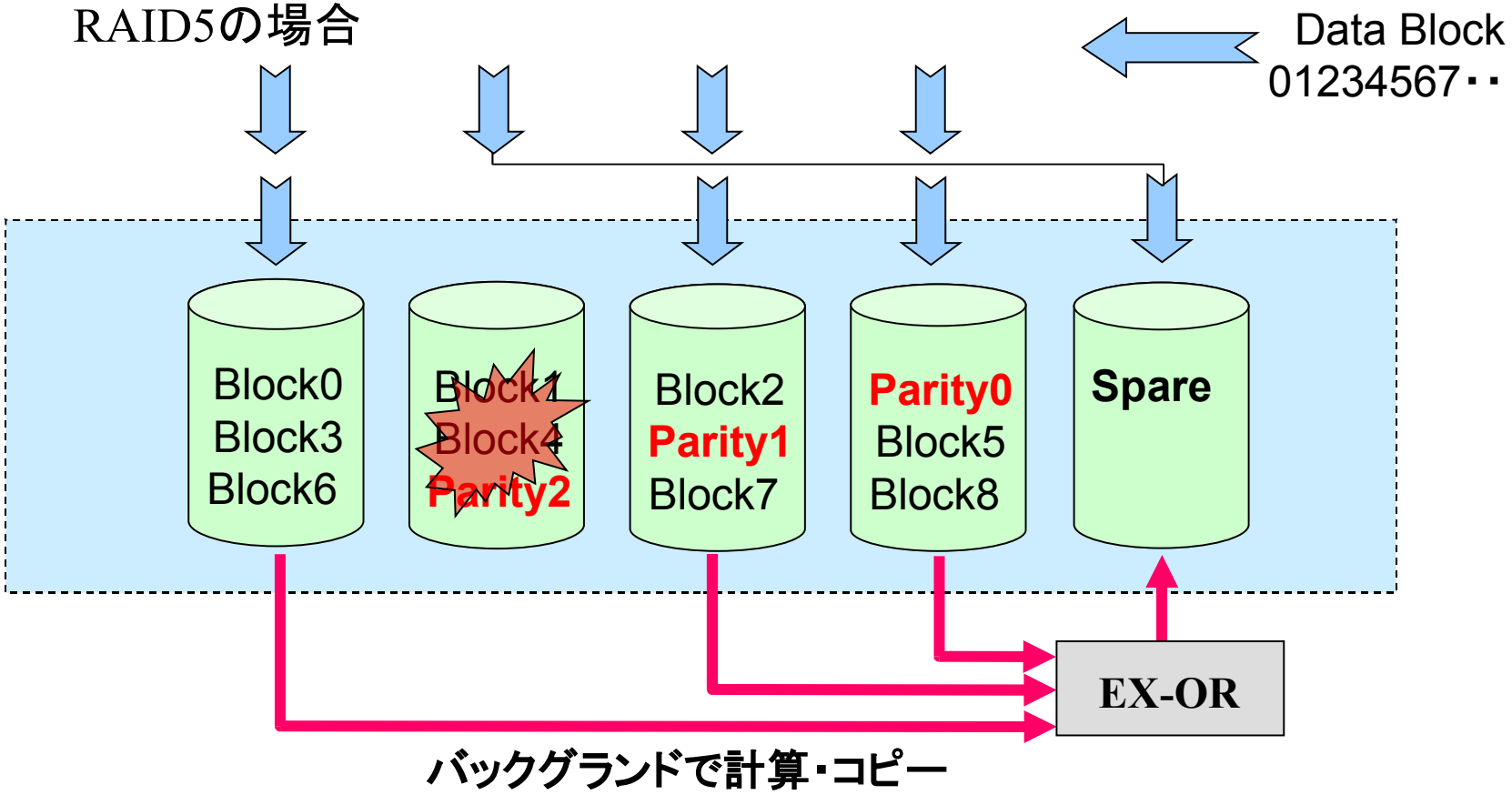
復元時間は機種により異なるが、1GBあたり1分程度

※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

3. RAID

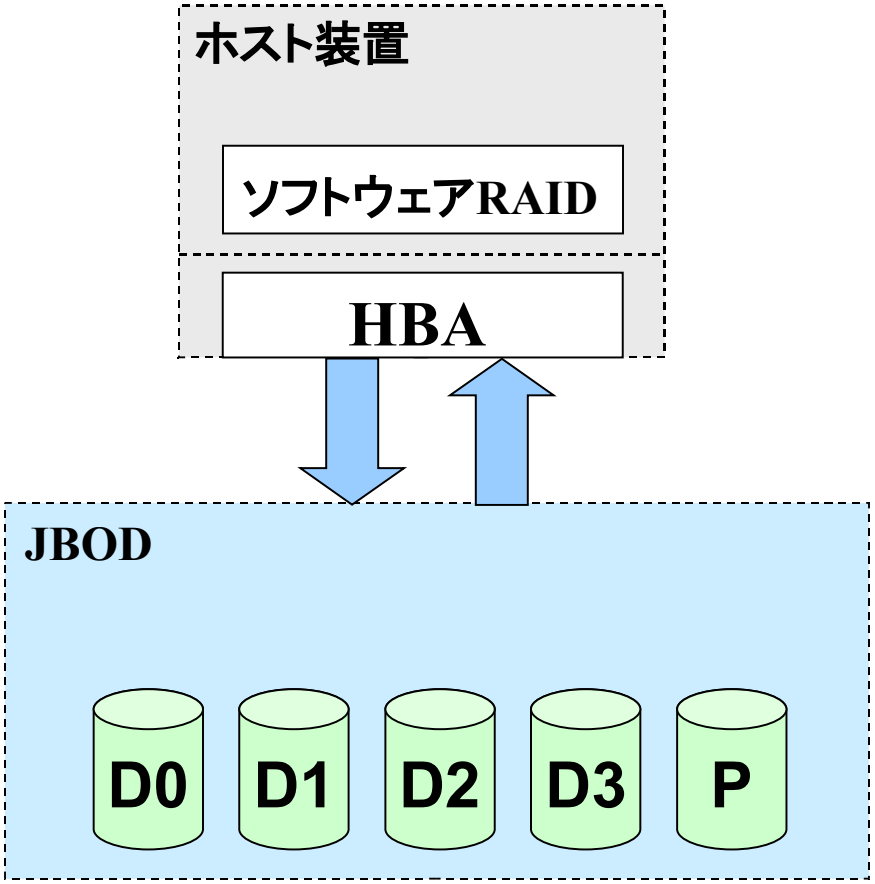
- ホットスペアの動作

RAID5の場合



3. RAID

- ソフトウェアRAID

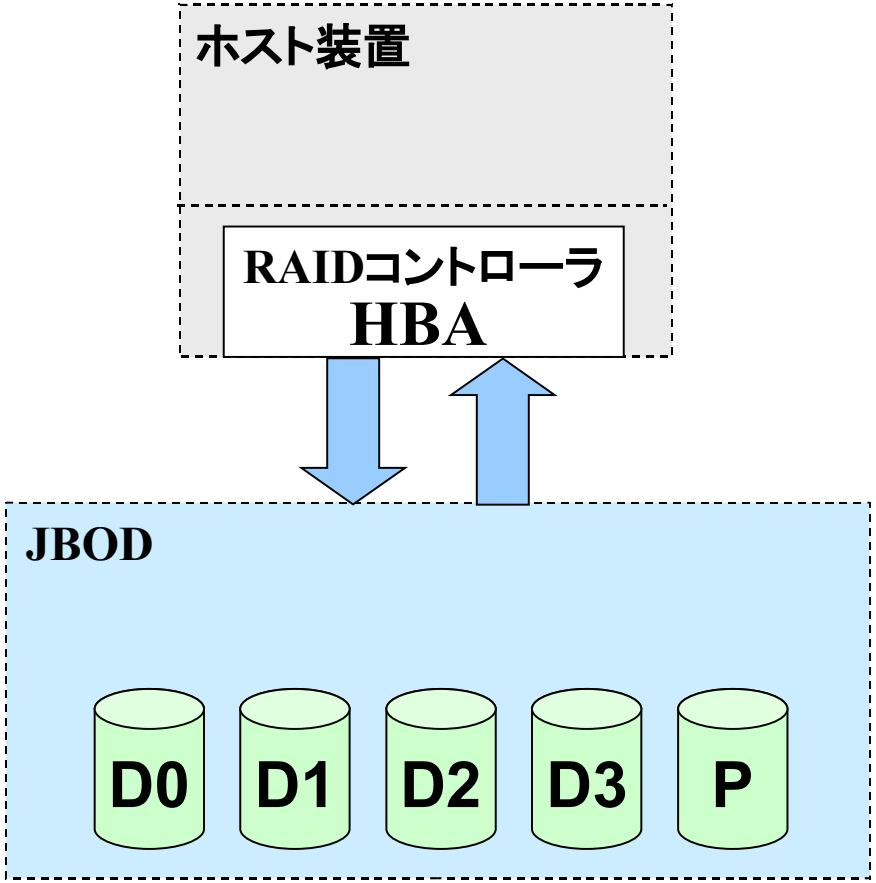


・HBAを内蔵するホストのCPUによりデータのストライピングパリティ計算、ミラー処理、各ドライブとのアクセスを行う。 HBA, JBODシステムがあればできるので比較的安価に構成できるが、CPU負荷がかかる。また、一般的には障害時の復旧手順は複雑。

・クラスターやSAN等の柔軟な構成には対応できない。

3. 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。

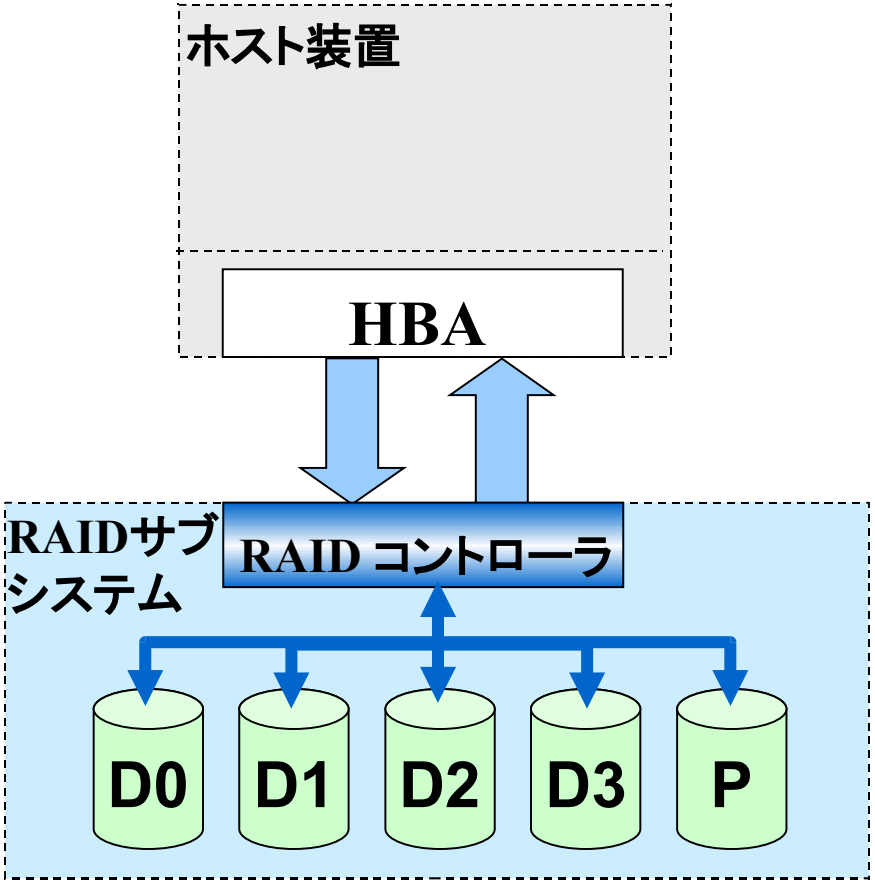
· クラスターやSAN等の柔軟な構成には対応できない。

· 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。

3. 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。

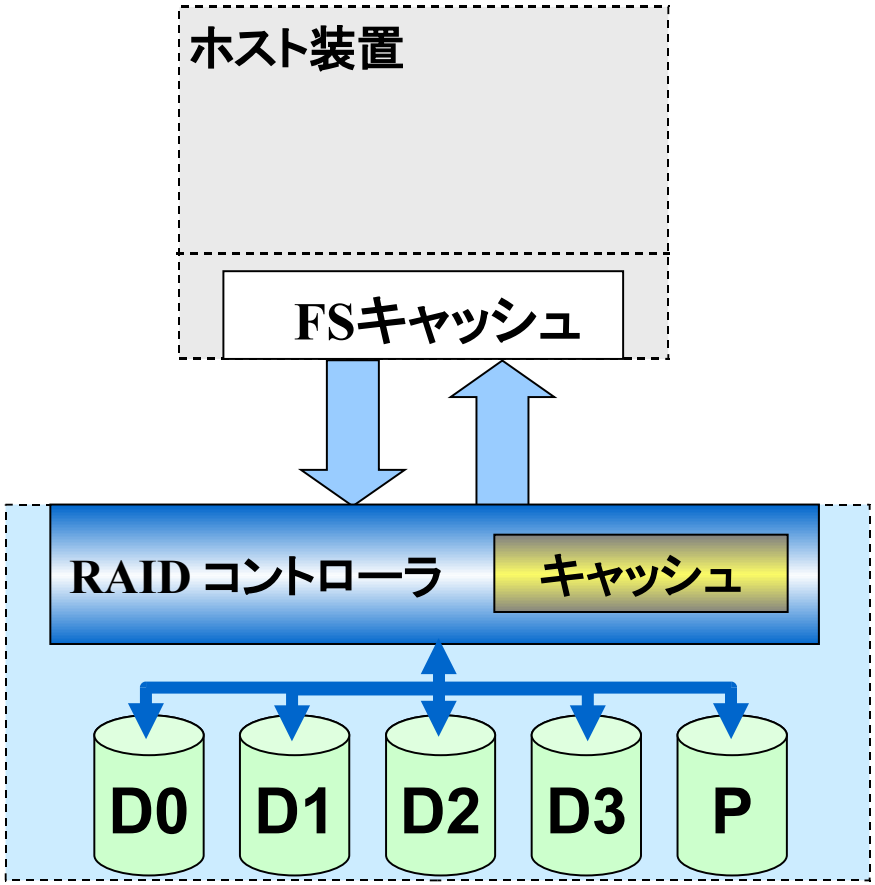
・SANを構成する場合は、RAIDサブシステム型が必要になる。

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。

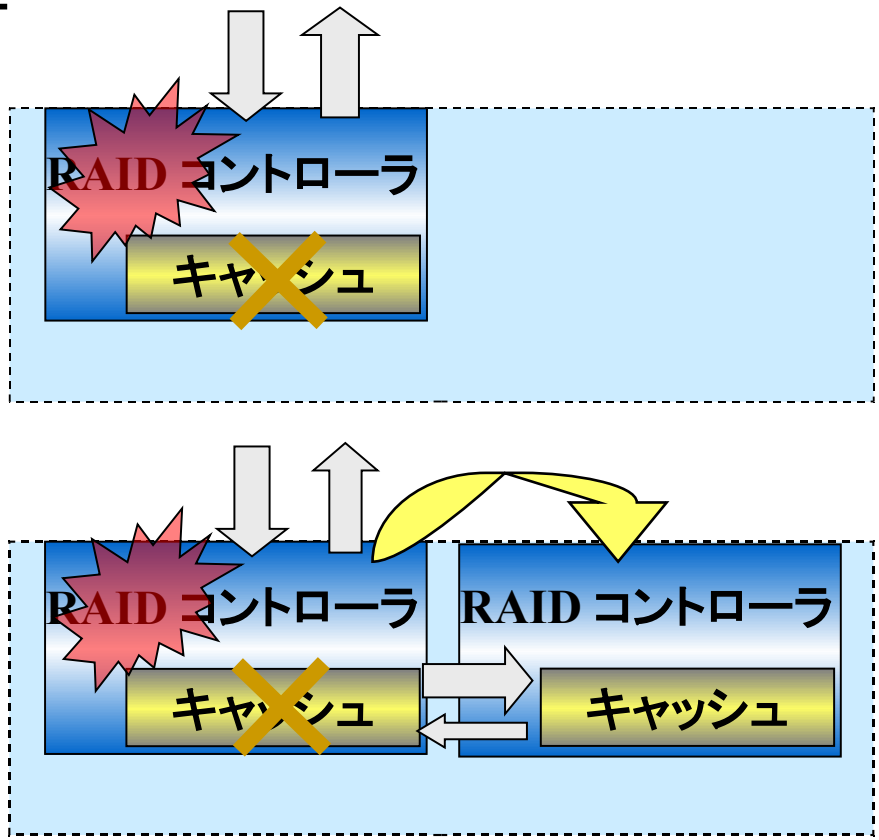


4. デバイスの信頼性と冗長

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。

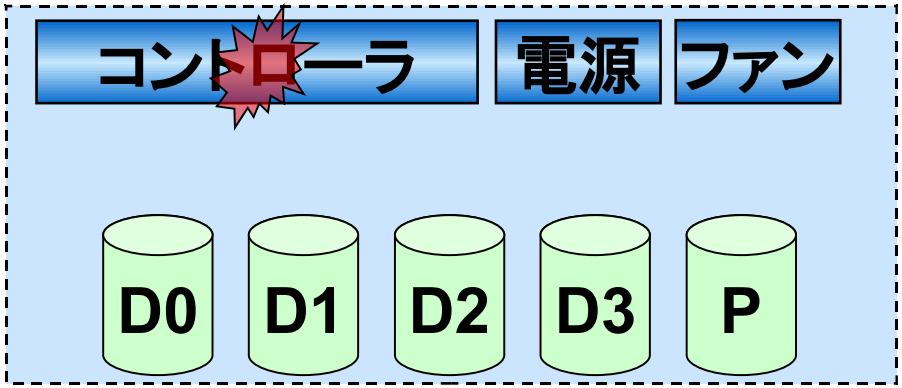
③ディスクに退避する。

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

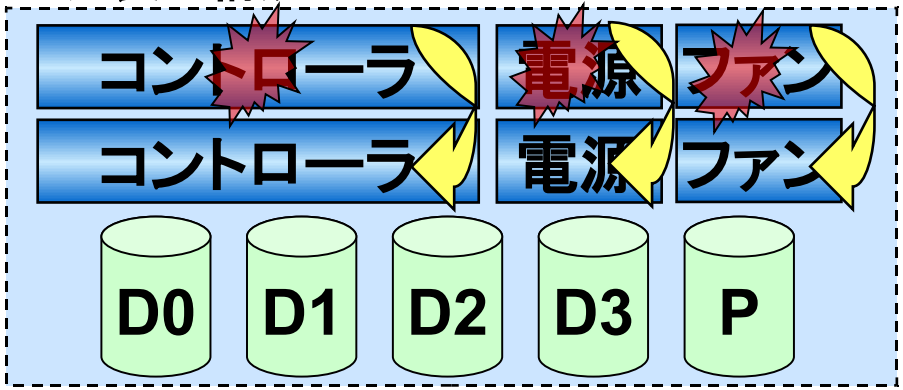
RAIDコントローラや電源、その他構成要素の単一故障によってデバイス全体が障害となってはならない。

・シングル構成



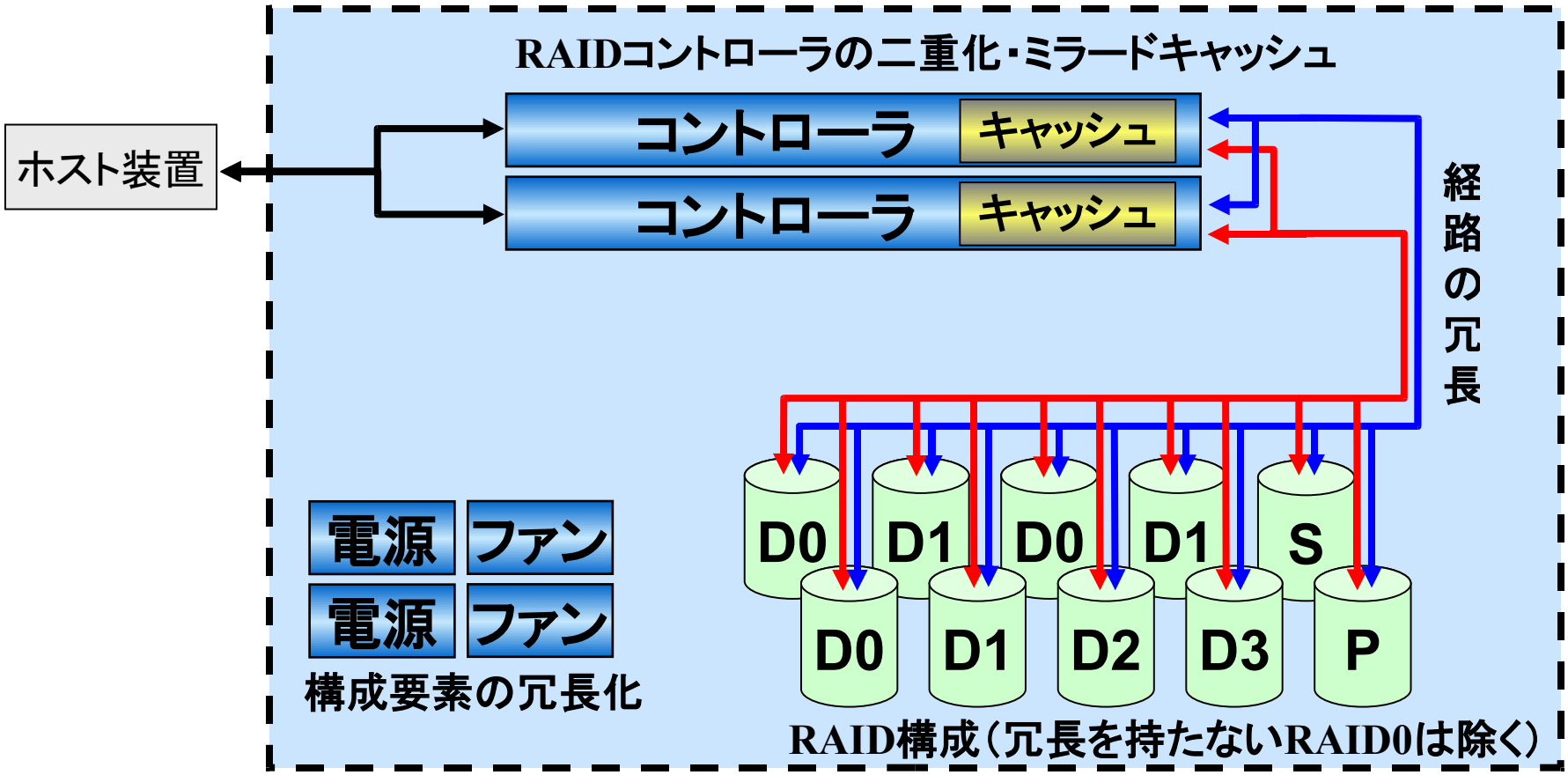
単一の故障で装置が障害となる

・冗長化構成

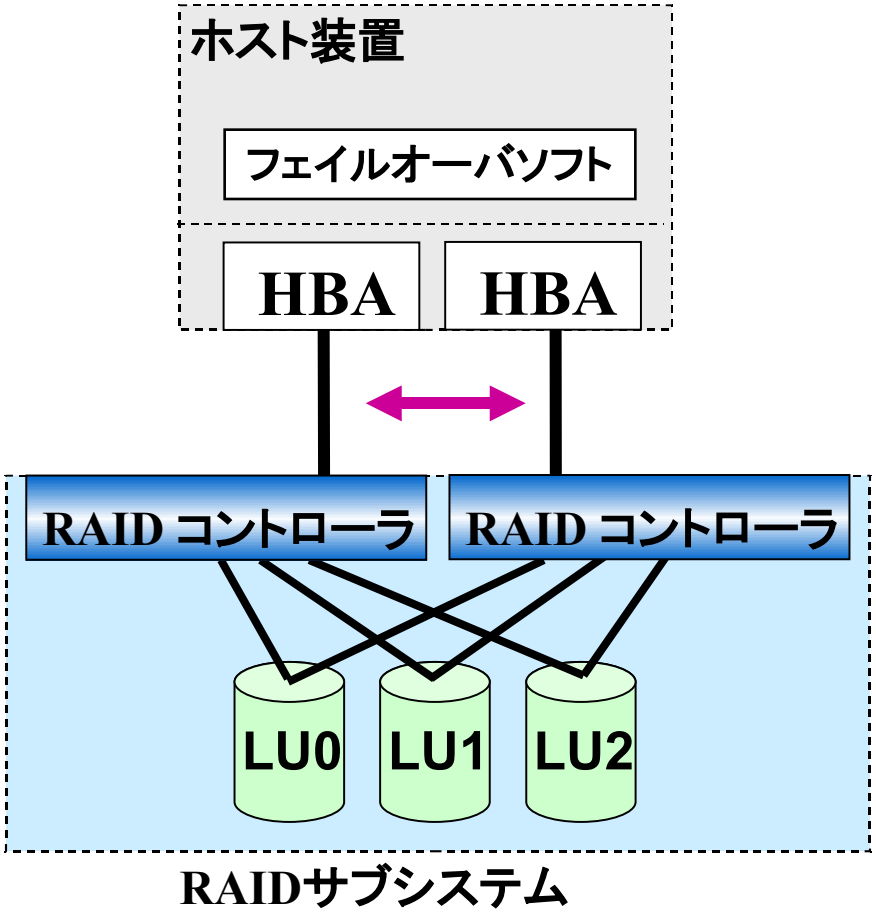


単一の故障でも装置としては障害とならない

4. デバイスの信頼性と冗長

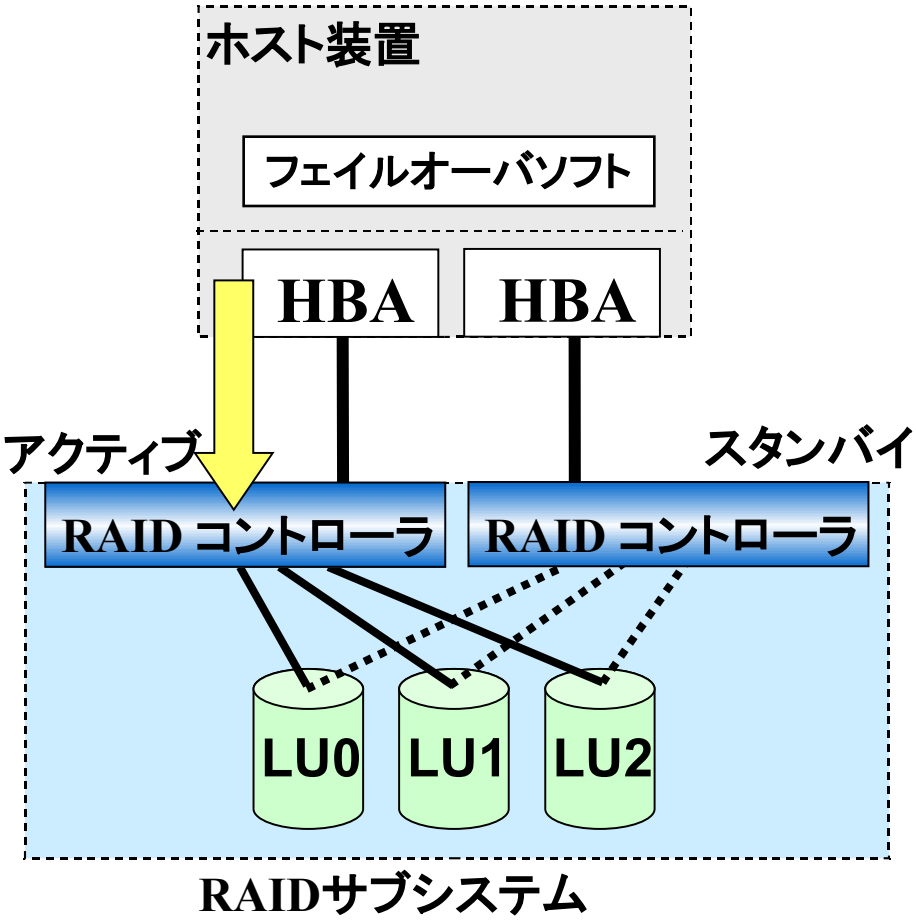


4. デバイスの信頼性と冗長



- 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。

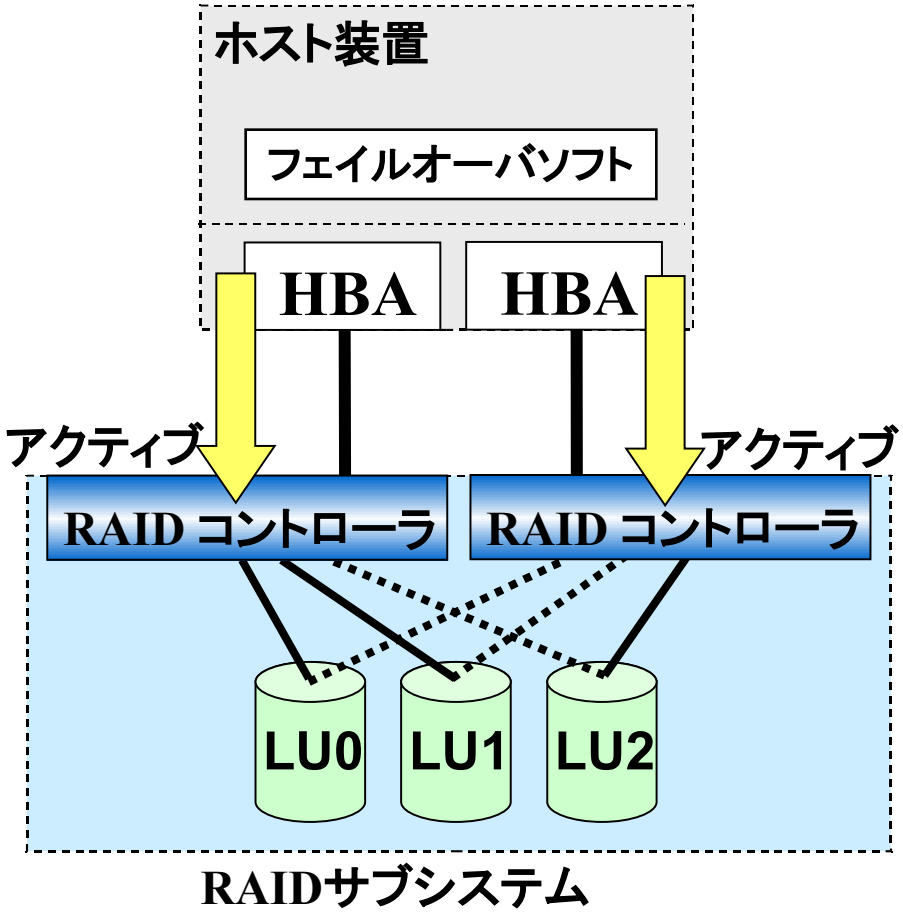
4. デバイスの信頼性と冗長



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。

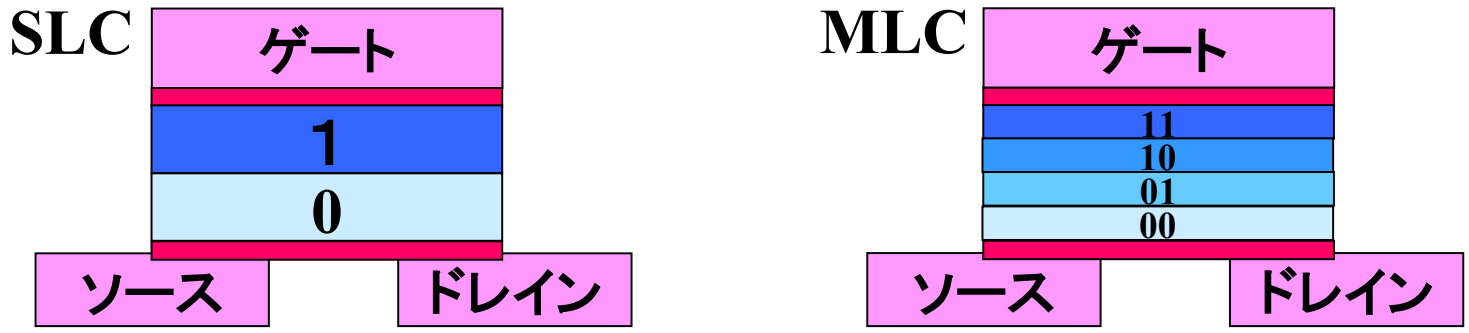
5. SSD (Solid State Drive : 半導体ディスク)

- 半導体の種類による違い
  - Flashメモリ
    - 比較的安価になってきており、個人向けでの使用も見受けられる
    - 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、ライトは低速
    - SLC(Single Level Cell)とMLC(Multi Level Cell)
    - ウェアレベリング(平均化処理)
  - DRAM/SRAM
    - リード／ライト高速だが、高価、バックアップ機構必要

5. 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。



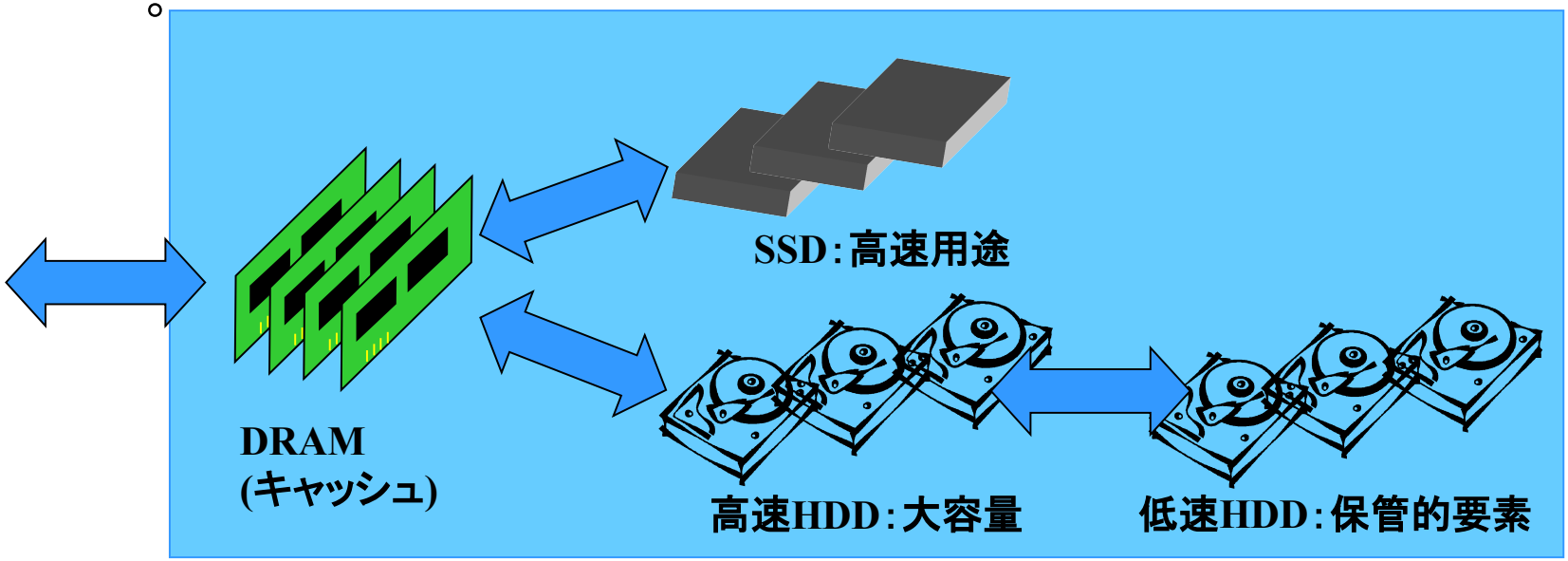
5. SSD（Solid State Drive：半導体ディスク）

- 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～1万回程度と言われている。

5. SSD（Solid State Drive：半導体ディスク）

- ハイブリッド利用
  - 高速性が求められるところにはSSD、容量(高速性が必須ではない）が必要な部分にはHDD（回転数により階層化）を利用することで、サーバからの要求・用途にあった装置がでてきている



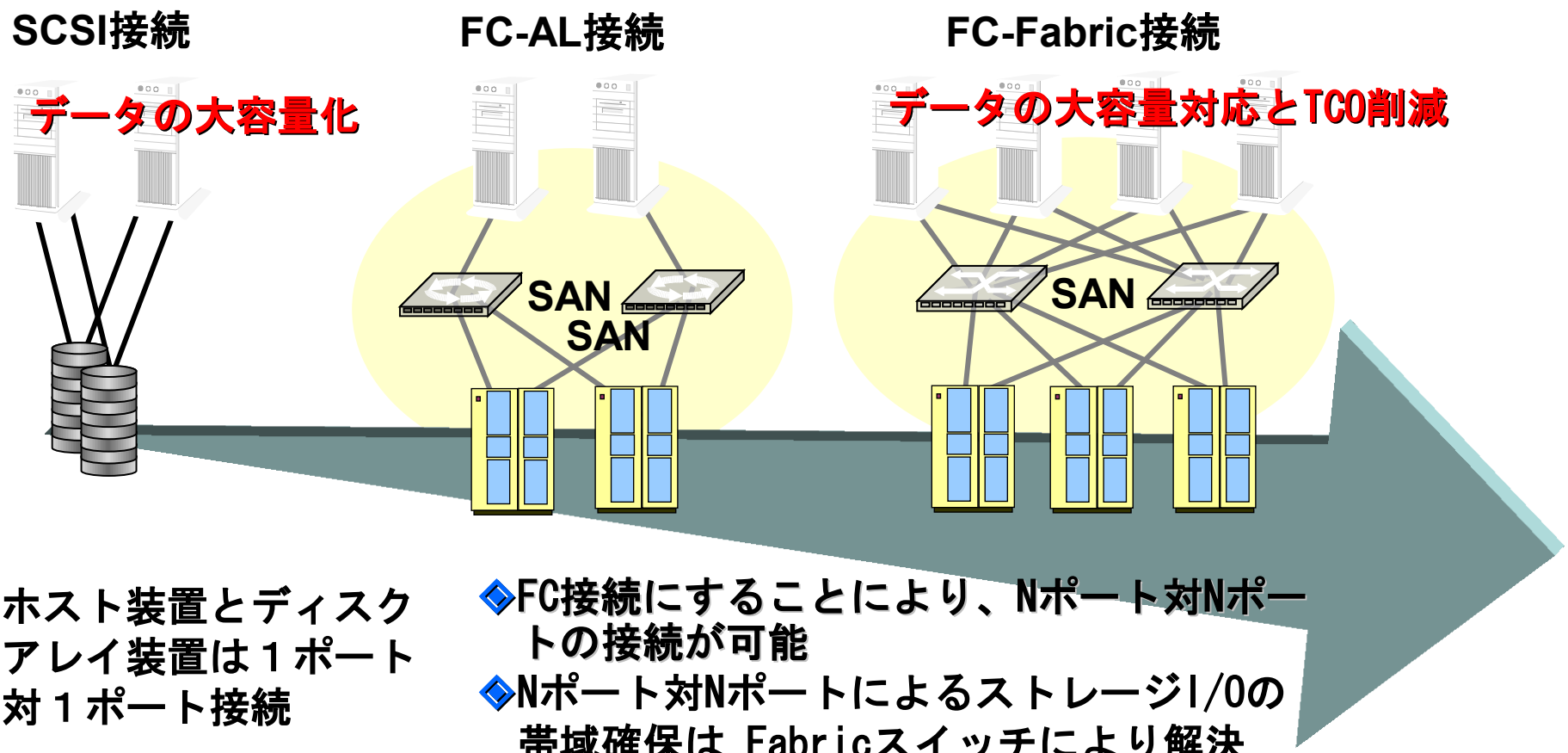
※装置ではなく、単一のHDDでも同じような動きがある。

6. DAS・NAS・SAN

- DAS・NAS・SANの違い・特長

  - DAS(Direct Attached Storage)
    コンピュータと直接接続されるストレージ装置。SCSIなど
  - NAS(Network Attached Storage)
    Network(主にEthernet)に接続される、NFS, CIFSなど、ファイルアクセスベースのプロトコルを用いて接続する様態、装置
  - SAN(Storage Area Network)
    FibreChannelなどチャネル型の接続インタフェースにより、ストレージ装置、コンピュータでネットワークを構築した様態。NASとは違い、ブロックアクセスベースのプロトコルを用いる。

6. DAS・NAS・SAN

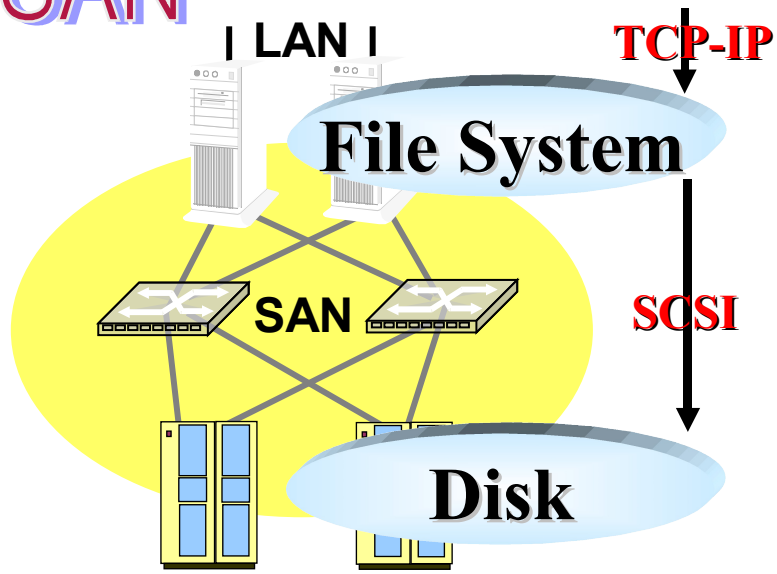


- ホスト装置とディスクアレイ装置は１ポート対１ポート接続

- FC接続にすることにより、Nポート対Nポートの接続が可能
- Nポート対Nポートによるストレージ I/Oの帯域確保は Fabricスイッチにより解決

6. DAS・SAN・NAS

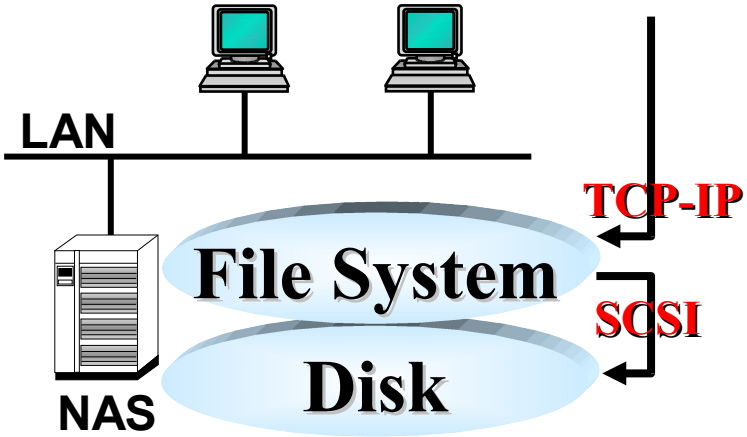
SAN



SCSIプロトコル：ブロックアクセス

デバイスシェアリング

NAS



IPプロトコル：ファイルアクセス

ファイルシェアリング

ご静聴ありがとうございました

＜休憩＞

Ⅱ．映像分野におけるストレージの使い方

Japan Data Storage Forum

（a）容量は？

ワーキングエリア　　TB

保存エリア　　TB

（b）必要とする転送速度は？

	Write	Read
Sequential	MB/S	MB/S
Random	MB/S	MB/S

（c）必要とする IOPS は？

100　　1,000　　10,000　　100,000

Ⅲ. ストレージ応用及び最新の技術動向

- ストレージの設置場所は、オフィス　or　サーバールーム？
- ストレージは保守契約に入っていますか？
- システム増設時のデータ移行はどうしていますか？
- バックアップを取っていますか？
- リストアした経験はありますか？
- バックアップとアーカイブの違い
- データの保存期間・保存場所は？
- 遠隔地へのデータ転送は？
- DR（**Disaster　Recovery**）、BCP（**Business Continuity Plan**）を考えていますか？