TotalStorage™

インターネット技術者のための ストレージ・ネット・ワーキング入門

2003年12月

佐野正和　　日本アイ・ビー・エム株式会社
sanomasa@jp.ibm.com

目次

- 第一章
  - 基礎的なことをそこそこ知っておきましょう
- 第二章
  - ファイバーチャネルSANからIPを活用するハイブリッドSANへ
- 第三章
  - 仮想化技術とストレージ・ネットワーク

日本アイ・ビー・エム株式会社

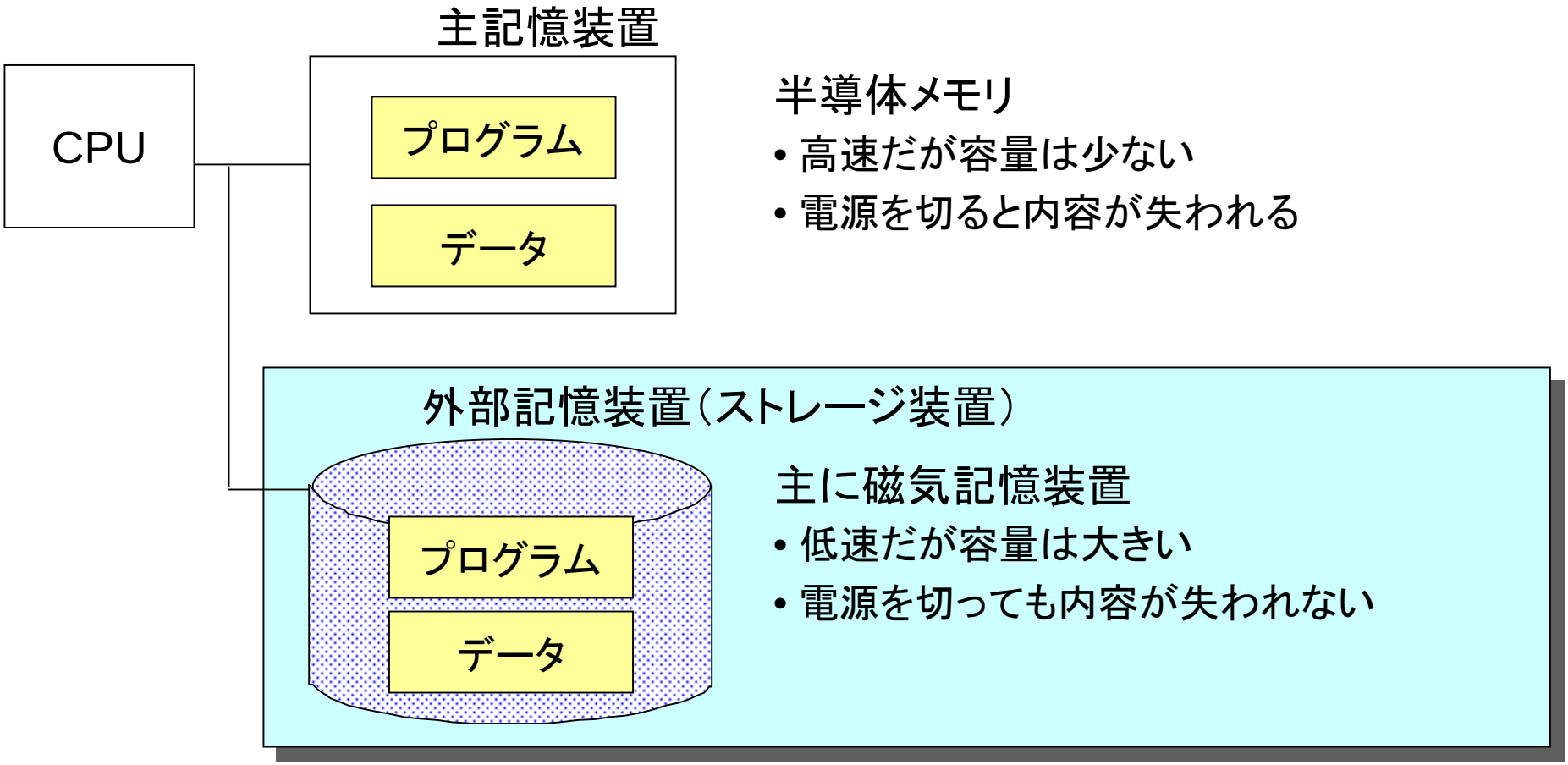
第一章

基礎的なことを
そこそこ知っておきましょう

日本アイ･ビー･エム株式会社

ストレージとは

コンピュータ・システムの構成



磁気記憶装置

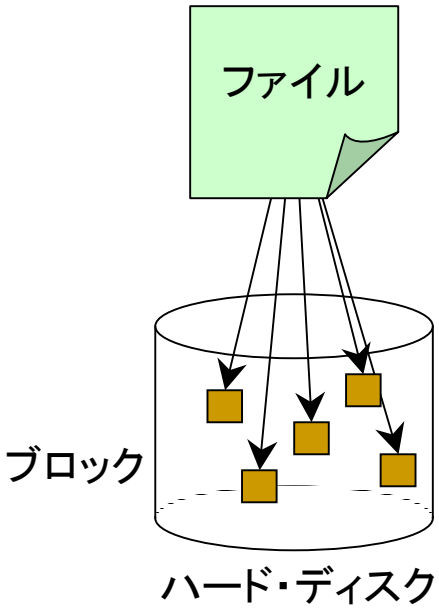
- 特長
  - 不揮発性
    - 磁気を利用して情報を保持するので永続性がある
  - 読み書き可能
    - 外部から磁界を与える(電磁石)ことで保持する内容を変更
    - 保持している内容(磁界)を電気信号として取り出す

- アクセス方式
  - シーケンシャル(順次)・アクセス・メディア
    - 磁気テープ
    - データがテープの先頭から末尾まで一方向に格納
  - ランダム・アクセス・メディア
    - FDDやHDD
    - ディスク表面が2次元であることを生かしてデータを格納
      - トラック、セクター
      - 複数のディスク(プラッタ)を用いることがほとんどなので実際は3次元に配置される

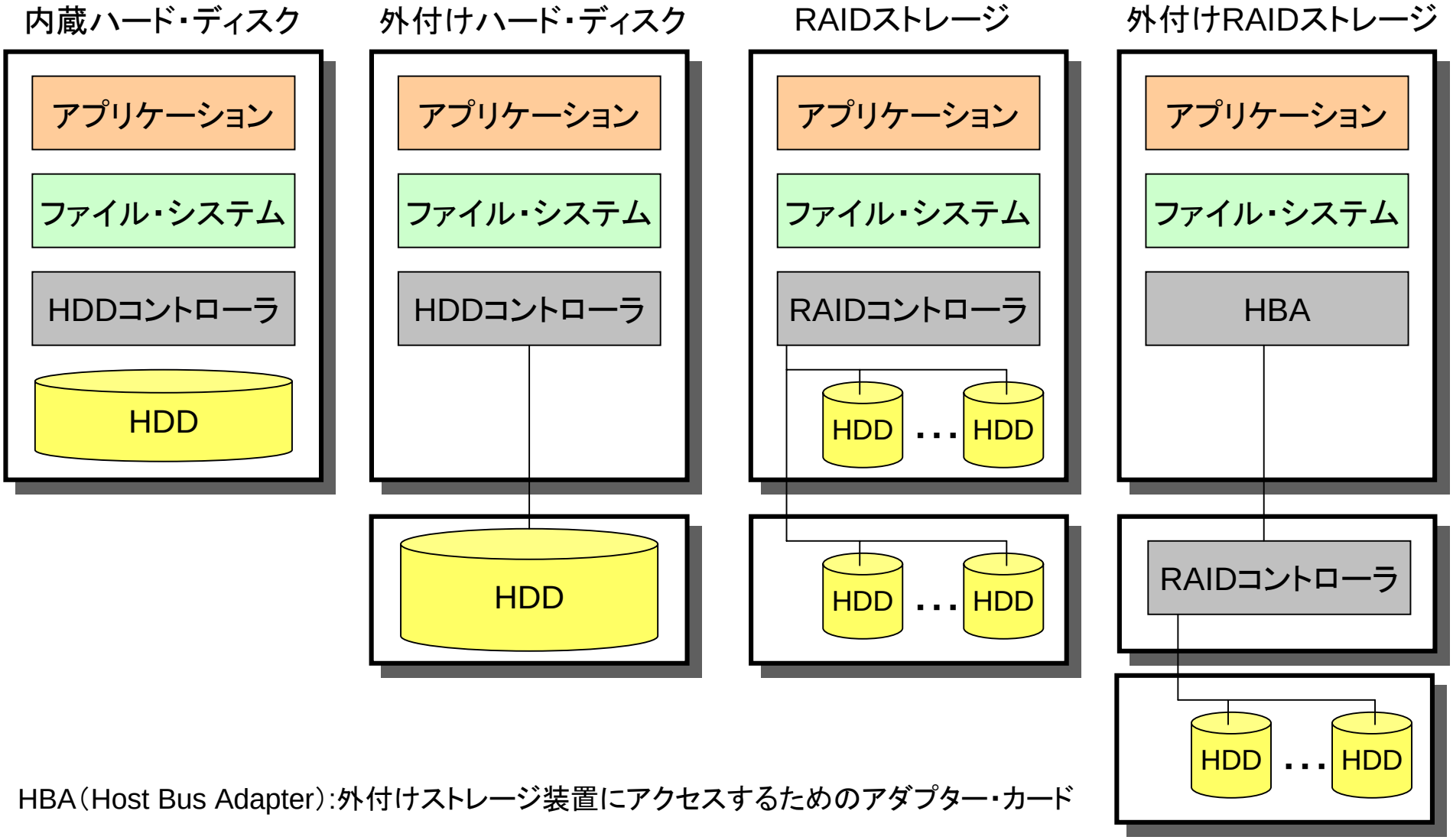
ディスク装置とファイル・システム

- ディスク装置への入出力
  - ディスク記憶装置はクラスタと呼ばれる領域で管理される
    - クラスタの指定方法はディスクインタフェースなどに依存する
  - 指定したクラスタへのデータの書き込み
  - 指定したクラスタからのデータの読み出し

- ファイル・システムへの入出力
  - ファイル・システムはOSなどに依存し様々な方式が存在
  - ファイルは複数のクラスタ（ブロック、セクター）から構成される
  - ファイル名やクラスタ構成を保持するアロケーション・テーブル（i-node）を管理する



サーバーとHDDの分離

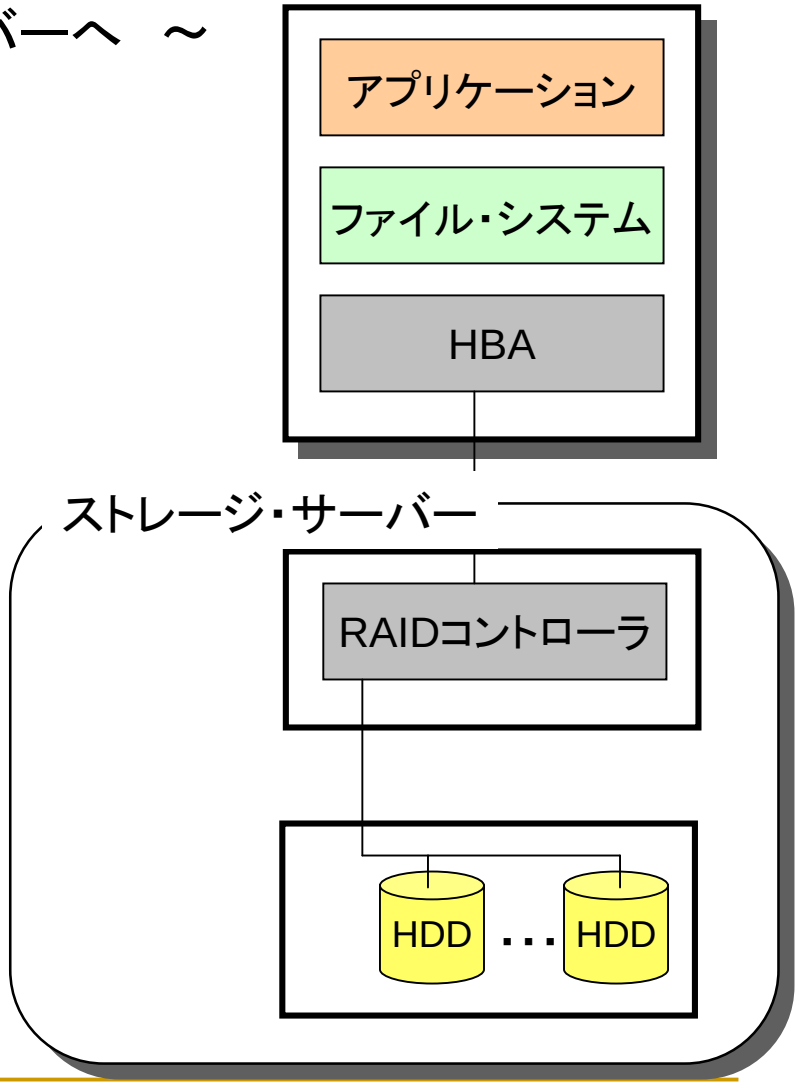


HBA（Host Bus Adapter）:外付けストレージ装置にアクセスするためのアダプター・カード

外付けストレージのインテリジェンス化

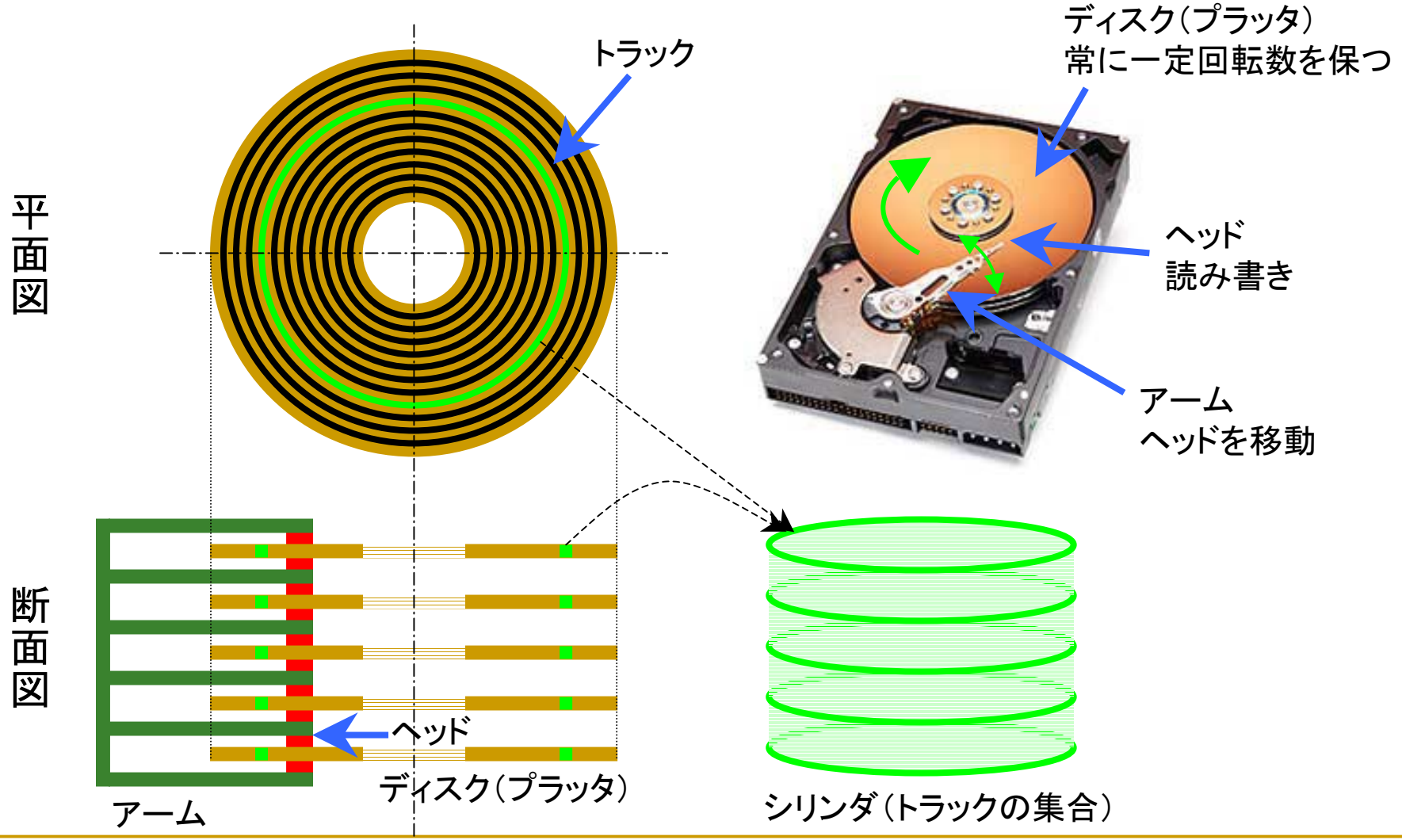
～　ストレージ・デバイスからストレージ・サーバーへ　～

- 高性能化
  - 高度なRAID制御
  - 高性能コントローラ
  - ハードウェアパリティ計算
  - 大容量キャッシュメモリ搭載
- 高可用性
  - コントローラの2重化
  - キャッシュの保護
    - バッテリ・バックアップ
    - キャッシュのミラーリング
- 接続性
  - サーバーとのさまざまな接続形態サポート
    - SCSI、ファイバーチャネル、など
  - 複数サーバーへの並列アクセス
  - サーバーアクセスの制限
    - LUNマスキング
- 付加機能
  - サーバーとは独立に各種処理ができる
    - 高速コピー
    - ミラーリング
    - 遠隔コピー



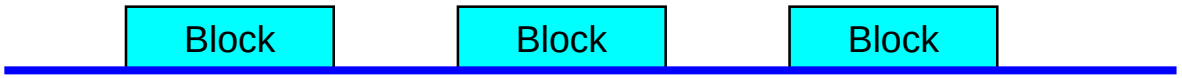
HDDの基本構造とインタフェース

HDD（Hard Disk Drive）ハードウェア構造

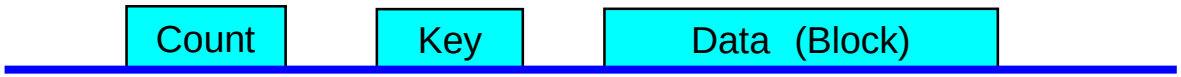


FBA方式とCKD方式

- ディスクの方式には、大きく2つのタイプがある
  - FBA方式
    - ブロック・サイズ(セクター・サイズ)を固定長にする方式
    - 現在主流となっている方式 => 通常のディスク装置

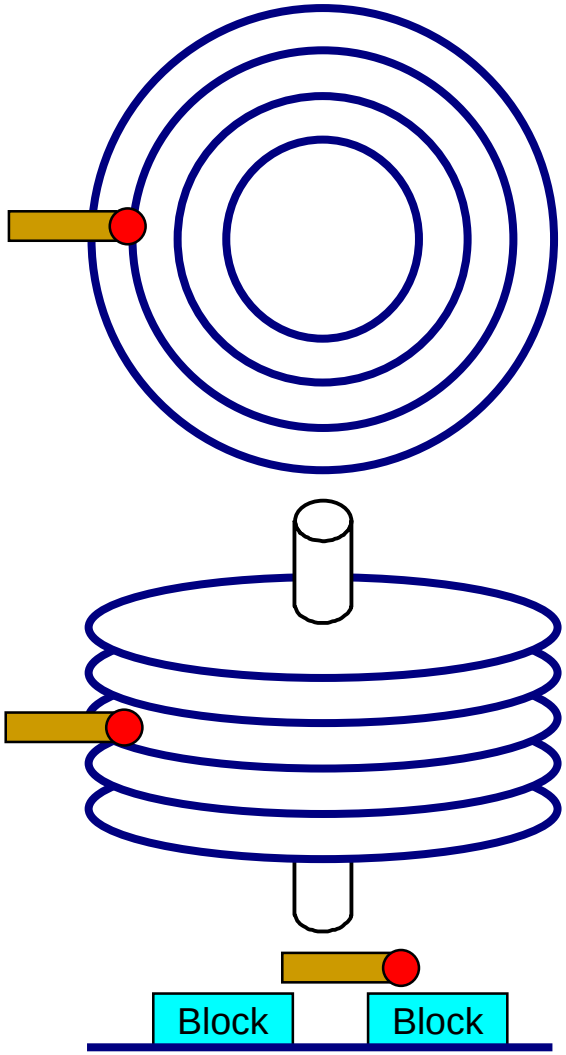


  - CKD方式
    - カウント、キー、データの順序でディスク上にデータを書き込む方式
    - ブロックのサイズは可変長な点が特徴
    - 主にホスト系のディスク装置で利用されていた
    - 現在はインタフェースとして残って入るが、実際のハードウェアとしては存在しない
      - 例：IBM 3380、IBM 3390



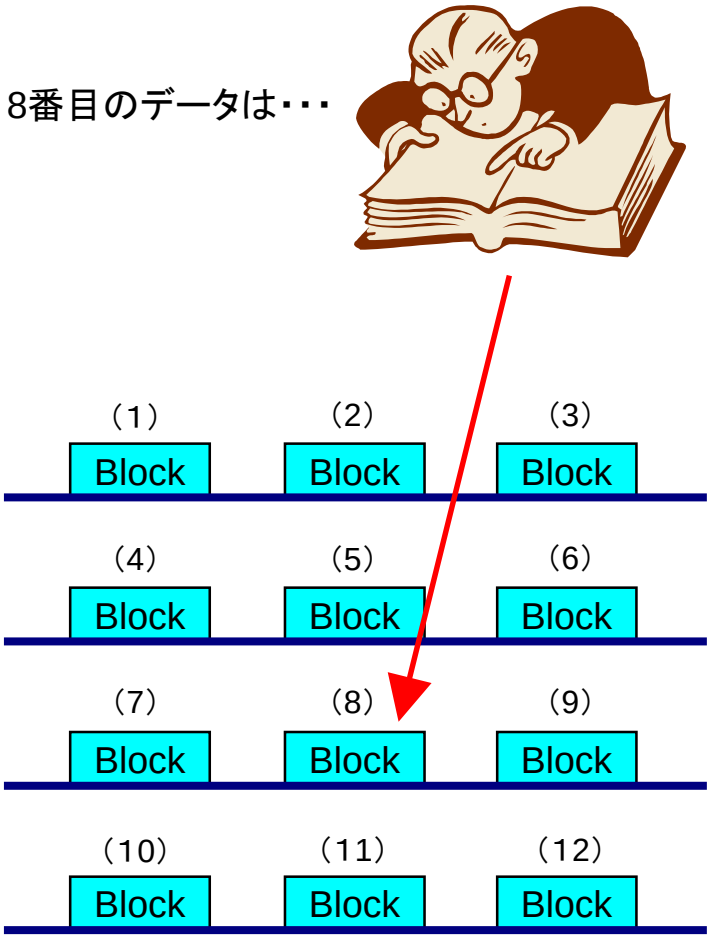
HDDのアクセス方法：CHS

- Cylinder/Head/Sector（C/H/S）
  - シリンダ番号（Cylinder）、ヘッド番号（Head）、セクター番号（Sector）の3つのパラメータを用いてハード・ディスクにアクセスする方式
  - HDDの物理構造に密接に関連するアクセス方法
  - 各用語解説
    - トラック
      - 1回転でアクセスできる1周分の円
    - シリンダ
      - プラッタの両面を使用したり、複数のプラッタから構成される場合、複数のヘッドはアームに連動して動くのでヘッド位置を決めるとアクセスできるトラックは同一円筒状に並ぶ
    - ヘッド
      - どのプラッタのどの面をアクセスするかヘッド番号により指定
    - セクター
      - 各トラックをセクター(ブロック)と呼ばれる短い単位に分割
      - セクターがHDDにおける記録単位
        - 標準的なHDDでは各セクターは512バイトの固定サイズ



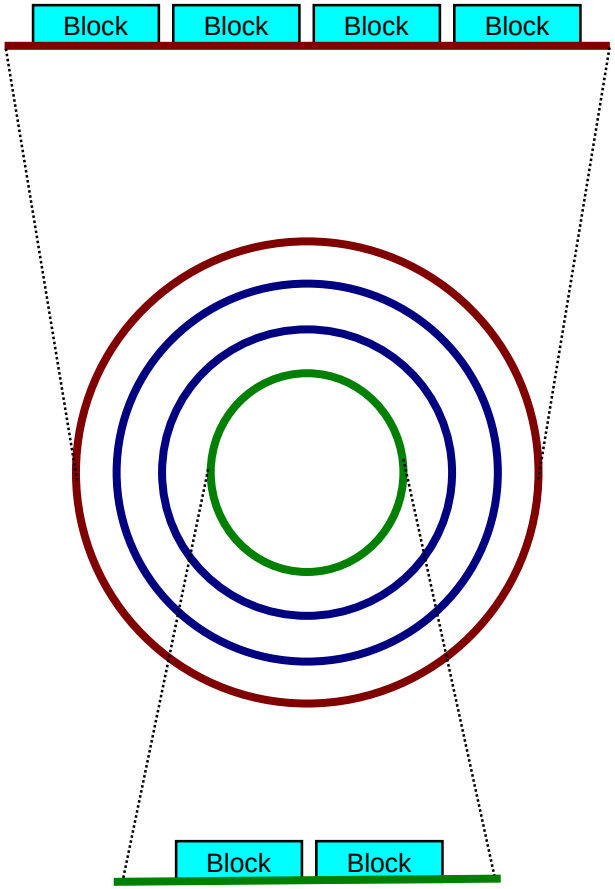
HDDのアクセス方法：LBA

- Logical Block Addressing（LBA）
  - ハード・ディスク内のすべてのセクターに通し番号を振り、その通し番号によってセクターを指定する方式
  - LBA自体にビット数の規定はないため、理論上は無限に拡張することが可能
    - IDE方式では28ビットまで、「Big Drive」方式は48ビットまで
    - SCSI方式では32ビットまで
  - IDE方式では、BIOSの規定以上の容量を持つディスクにはCHSでアクセスできないことから、現在はハード・ディスクの全セクター（ブロック）に通し番号を振るLBA方式(限界は正確には128GB)が主に使用されている



マルチ・ゾーンド・ビット・レコーディング

- ゾーン記録方式とも呼ばれる記録密度を高めるための方式
  - ディスクの最内周と最外周のシリンダ数をいくつかの領域（ゾーン）に分割
  - それぞれのゾーン毎に1トラックあたりのセクター数を定める
    - CHS方式では、最外周と最内周ではトラック長が異なる
      - 最内周トラックの記録密度が最も高く、最外周トラックの記録密度が最も低い
    - ハード・ディスクの回転数は一定なので、最内周と最外周ではデータ転送速度が異なる
      - 最外周ゾーンの転送速度が最も高く、最内周ゾーンの転送速度が最も低い
  - 近年のほとんどの大容量ドライブは8～16ゾーン程度のゾーン記録方式を採用しており、ドライブあたりの容量は従来方式と比較すると20%～50%も増加しているといわれている
- CHSアドレスは、物理的なハード・ディスクの内部構成とは異なっている

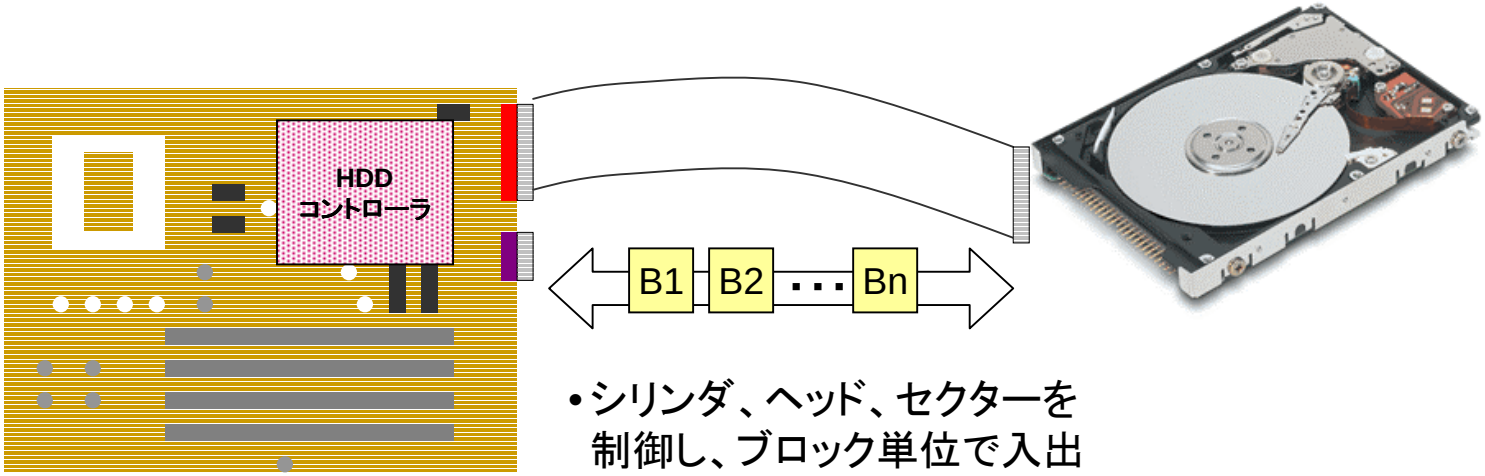


マルチ・ゾーンド・ビット・レコーディング採用ディスク例

HGST Ultrastar 146Z10

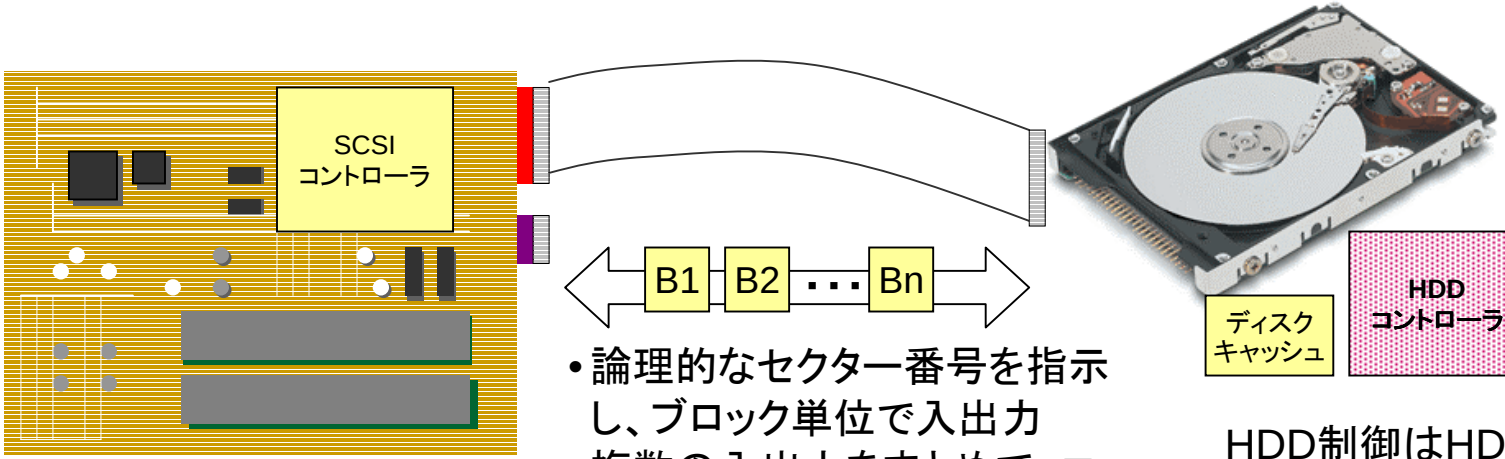
Zone	Physical Cylinders	Sectors/Track
Data Zone 0	0 - 383	864
Data Zone 1	384 - 3967	840
Data Zone 2	3968 - 5631	800
Data Zone 3	5632 - 6527	780
Data Zone 4	6528 - 8703	768
Data Zone 5	8704 - 15359	720
Data Zone 6	15360 - 18047	672
Data Zone 7	18048 - 19199	660
Data Zone 8	19200 - 21503	640
Data Zone 9	21504 - 24959	600
Data Zone 10	24960 - 27775	560
Data Zone 11	27776 - 29183	540
Data Zone 12	29184 - 30719	520
Data Zone 13	30720 - 35199	480
Data Zone 14	35200 - 36735	440

HDDの高機能化



- シリンダ、ヘッド、セクターを制御し、ブロック単位で入出力

HDD制御はコンピュータ側

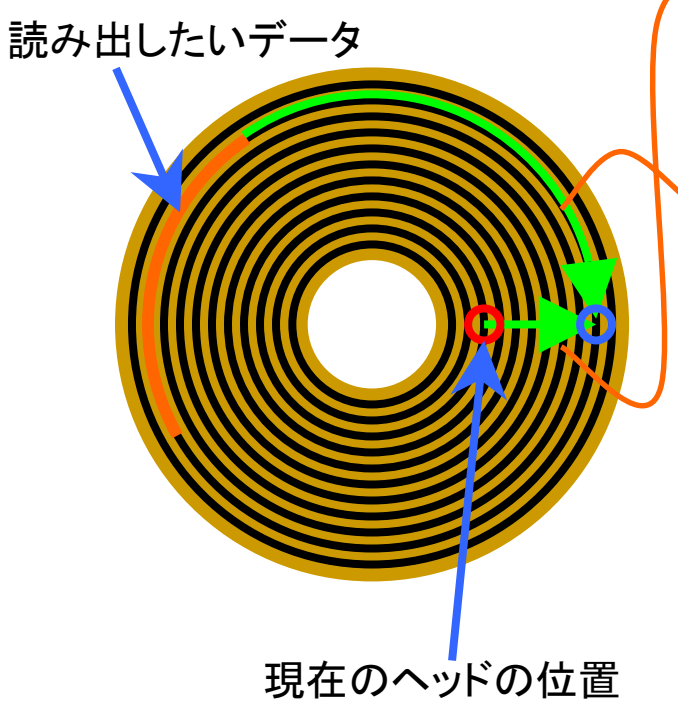


- 論理的なセクター番号を指示し、ブロック単位で入出力
- 複数の入出力をまとめて、コマンド&データで受け渡し

HDD制御はHDD側
読み書き動作をスケジューリング

HDDの物理的なアクセス時間

- シーク時間＋回転待ち時間＋データ転送時間



① 読み出すデータのあるトラックに
ヘッドを移動する
（シーク時間）

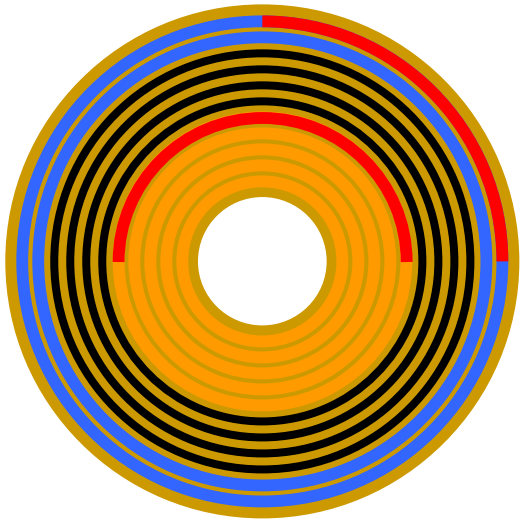
② 読み出すデータの先頭がヘッド
の下に来るのを待つ
（回転待ち時間）

③ 読み出すデータが全てヘッドの
下を通過して、ヘッドから読み出
される
（データ転送時間）

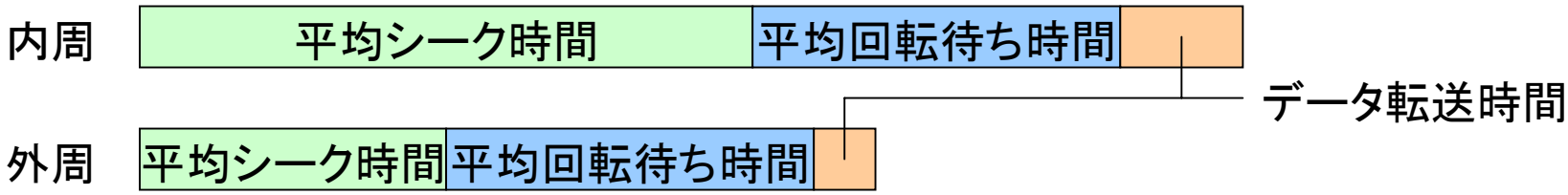
HDDの外周配置と内周配置の違い

- 外周の方が高性能
  - 平均シーク時間が小さい
  - データ転送時間が小さい
  - 使用頻度の高いデータを外周付近（トラック番号の小さいトラック）に配置するのが良い

  しかし、、、
- LBA方式IDE HDDやSCSI HDDでは、セクター番号とHDD内の記録位置を結びつけることは困難



同じデータ量を記録するトラック数は、
外周付近の方が内周付近よりも少ない。
同じデータ量を読み出すための回転量も、
外周付近の方が内周付近よりも少ない。



HDDインタフェース概要(1)

- E-IDE (ATA):Enhanced Integrated Drive Electronics
  - パソコンとハード・ディスクなどの記憶装置を接続する方式の一つ
  - Western Digital社が提唱した、IDE(ATA)方式の拡張仕様であり、IDEでは2台までだった最大接続機器数は2系統2台ずつの合計4台まで増加し、CD-ROMドライブなどハード・ディスク以外の機器も接続できるようになった(E-IDE)
  - IDEではハード・ディスクの最大容量が528MBに制限されていたが、8.4GBまでのハード・ディスクが使えるよう改善され、データ転送速度の向上も図られた
  - アメリカ規格協会(ANSI)によって、ハード・ディスク部分の仕様はATA-2、ハード・ディスク以外の機器の接続に関する仕様はATAPIとして、規格化された

- SCSI:Small Computer System Interface
  - パソコン本体と周辺機器の接続方法の取り決め
  - アメリカ規格協会(ANSI)によって規格化されている。最初の規格はShugart社(現在のSeagate Technology社)の開発したSASIをベース
  - 現在では汎用性や性能が大幅に強化された後継規格、SCSI-2やSCSI-3が普及している

HDDインタフェース概要(2)

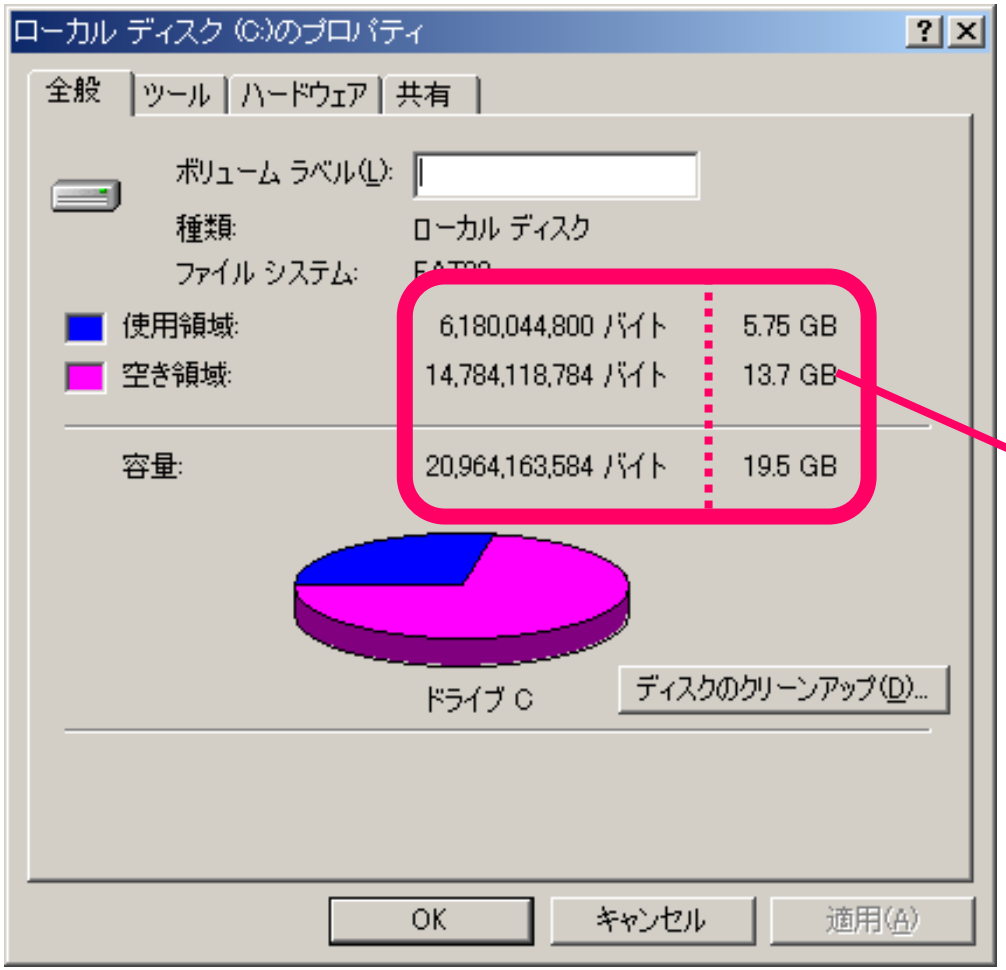
- FCP（Fibre Channel Protocol: SCSI over Fibre Channel）
  - Fibre Channel物理層上でSCSIコマンド&データを転送するための規格
  - SCSI-3規格で規定
  - 転送速度は100MB/s
  - SANの普及に伴い、現在、オープン系システムの代表的なインタフェースとなっている
- SSA：Serial Storage Architecture
  - IBM社が中心となって開発されたシリアル転送方式を採用したSCSI規格の一種
  - SCSI-3規格に含まれており、転送速度は最大160MB/s
  - 接続時の機器間の距離は最大25m、最大接続台数は96台で、ループ状の接続が可能になっている
  - ケーブルには基本的にシールド付より対線(STP)を使うが、光ファイバーケーブルを用いることで接続距離を最大2.5kmまで伸ばすこともできる
  - 外部インタフェースとしてはFibre Channelの普及が進んでいることもあって、普及率はそれほど高くない
    - 代表的な装置
      - IBM TotalStorage Enterprise Storage Server (ESS)の内部ディスク・インタフェース
      - IBM TotalStorage 7133 ディスク・サブシステム

HDDインタフェース概要(3)

- Serial ATA
  - パソコンとハード・ディスクなどの記憶装置を接続するIDE(ATA)規格の拡張仕様の一つ
  - 2000年11月に業界団体「Serial ATA Working Group」によって仕様の策定
  - Serial ATAは、Ultra ATAなどの現在のATA仕様で採用されていたパラレル転送方式を、シリアル転送方式に変更したもの
    - シンプルなケーブルで高速な転送速度を実現することができる。従来のパラレル方式のATA諸規格との互換性も持っている。
    - 従来のパラレル方式のATA仕様で転送速度が最も高速なのはUltra ATA/133の133MB/sで、パラレル方式ではこれ以上の高速化は難しいとされる
  - Serial ATAの最初の規格「Ultra SATA/1500」は1.5Gbps(約190MB/s)と、従来の約1.4倍の速度を実現する
  - Serial ATA仕様は今後も拡張を続け、2004年にはその倍の3Gbps(380MB/s)、2007年には6Gbps(750MB/s)に引き上げられる予定

ディスク容量

- ディスク容量を表す単位は2通りが混在している



147.8GB=137.6GB

HDD物理容量
1K=1,000

OSの容量
1k=1,024= 2^{10}

G=$1{,}024^3$
=$(2^{10})^3$
=2^{30}

単位:K(キロ)、M(メガ)、G(ギガ)

- コンピュータの単位:K=1,024=2^{10}
- HDDの物理容量:K=1,000

分野		K (キロ)	M (メガ)	G (ギガ)
コンピュータ メモリ LAN	2進法	1,024 2^{10}	1,048,576 $(2^{10})^2=2^{20}$	1,073,741,824 $(2^{10})^3=2^{30}$
科学技術一般 HDD 通信業者	10進法	1,000 10^3	1,000,000 $(10^3)^2=10^6$	1,000,000,000 $(10^3)^3=10^9$

※ 3.5“ 1.44MBのフロッピー･ディスクの場合は特別。
M=1,024,000=1,024×1,000=$2^{10}\times10^3$

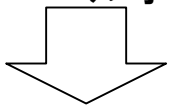
ストレージ・サーバー、
ディスク・サブシステムの制御装置機能

システム製品としてのディスク装置

- ディスク製品をHDD単体ではなく、システム製品として提供
  - ストレージ・サブシステム（ストレージ・サーバー）としてメーカーは提供する
  - 単純にHDDを提供するのではなく、可用性やパフォーマンスの面で付加価値を付けた製品としてユーザーに提供
  - 一般にHDD単体による利用に比べ、使用するユーザーのメリットは大きい
  - 多くの機能は「制御装置」、または「制御機構」と呼ばれる仕組みを保持し、それらがインテリジェンスを持ってサブシステム全体の制御を行い、各種機能を提供する
- ストレージ・サーバーの制御装置が持つ代表的な機能 (例)
  - 論理的なディスク・イメージの提供
  - RAID機能
  - キャッシュ
  - 高速コピー機能
  - 遠隔コピー機能
  - 電源/ファン/制御装置の二重化
  - マルチ・パス機能
  - LUNマスキング機能

高性能、高可用性のストレージを構成するには

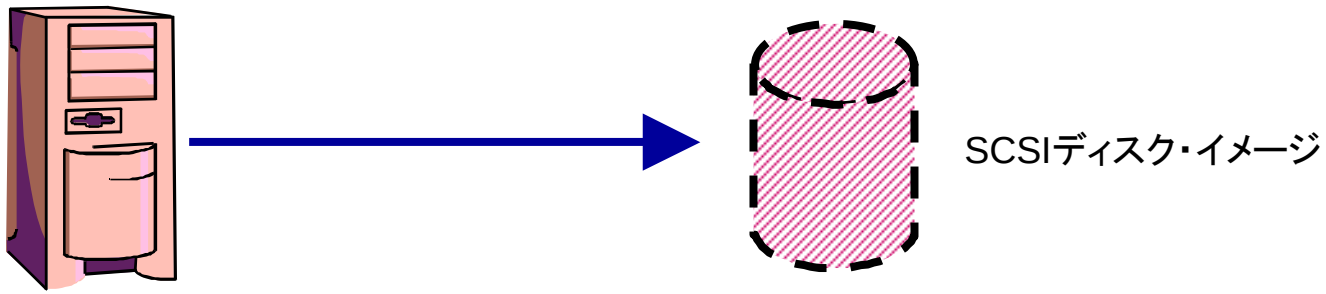
- 高性能
  - IOPS : Input Output per Second （トランザクション性能）
    - 1秒間に何回、データを書き込み/読み出しできるか
    - RAIDアレイ内の物理HDD数を多くする
  - MB/s, GB/s （スループット性能）
    - 1秒間に何MB(GB)、データを書き込み/読み出しできるか
    - 回転数が早いHDDを使用する、HDD数を多くする
- 高可用性
  - ホットスワップ
  - 自動リビルド
  - ホットスペア
    - RAID装置を集約した方がスペアドライブ配置によるオーバヘッドが少ない
  - 多重コントローラを使用した自動フェールオーバ

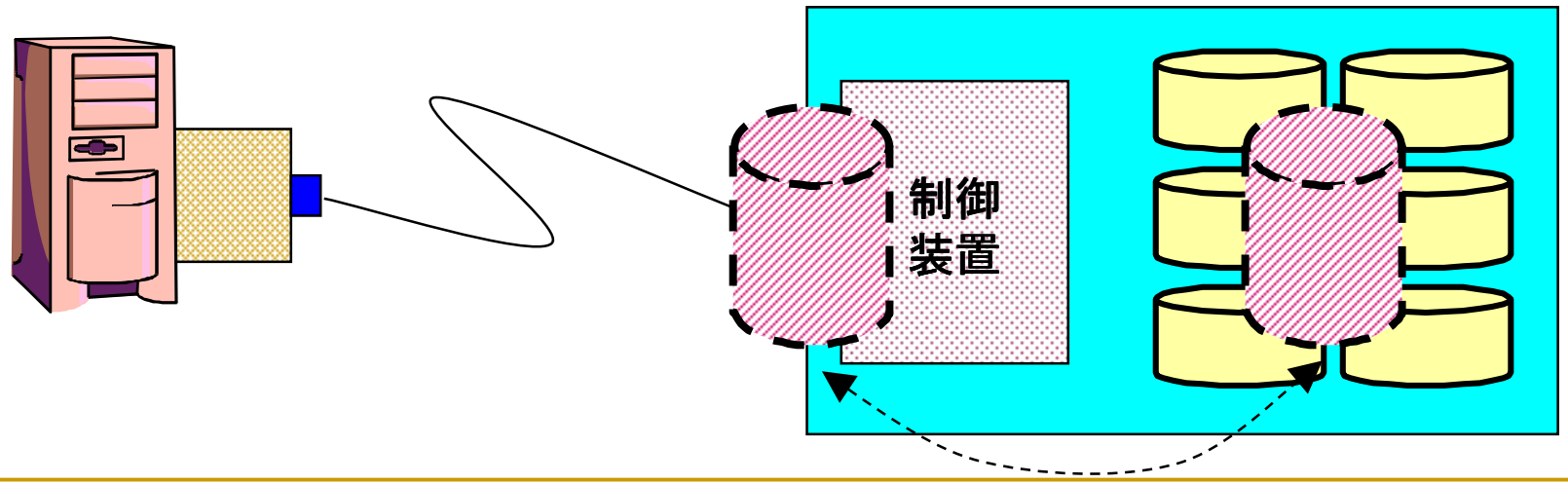


たくさんのHDDを接続でき、かつ、転送帯域が広いインタフェース
コントローラのフェールオーバに対応可能なインタフェース

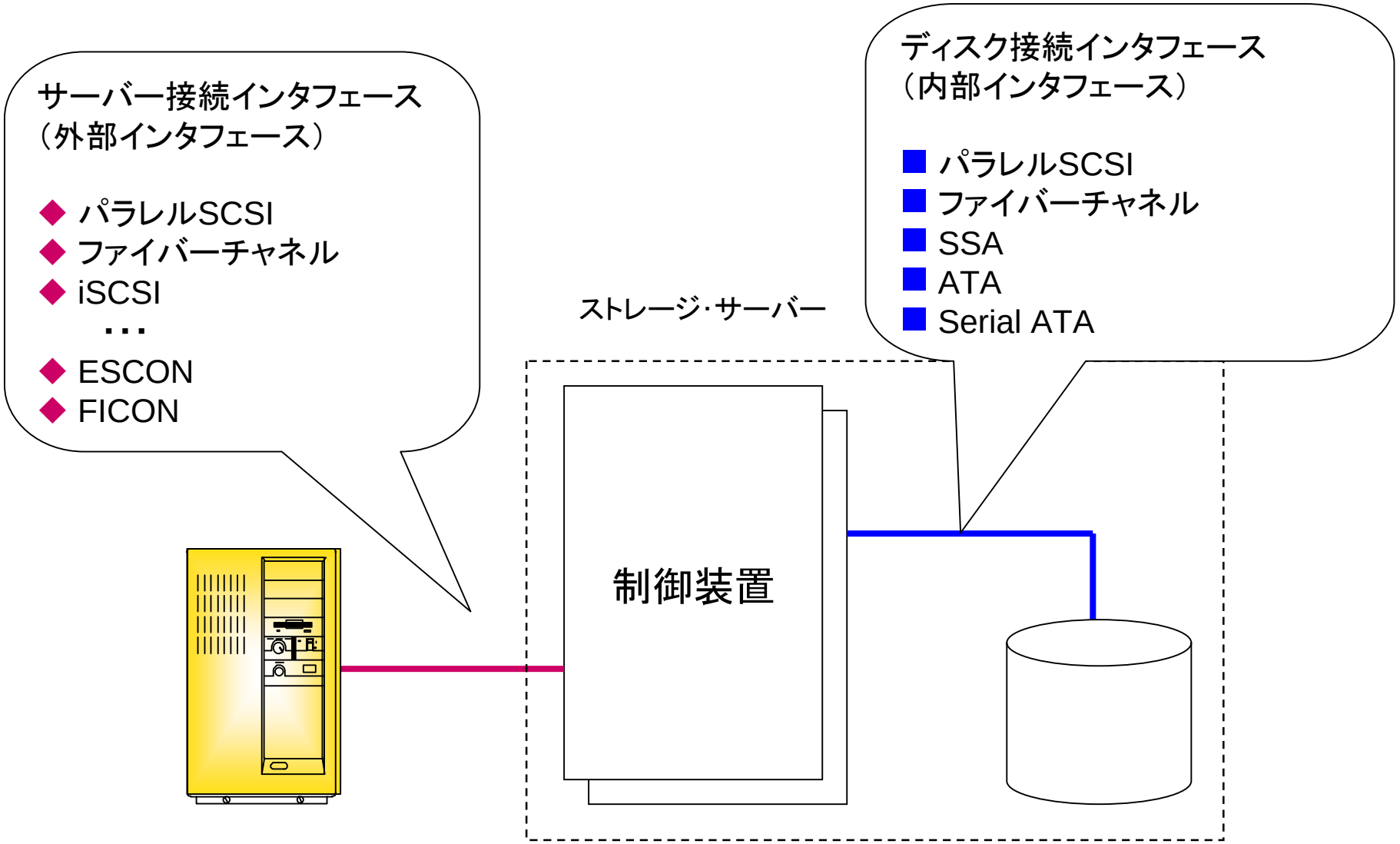
サーバー、制御装置とHDDの関係

- 制御装置はHDDとサーバーとの間で活動を行い、各種機能を提供する





インタフェースの組み合わせ

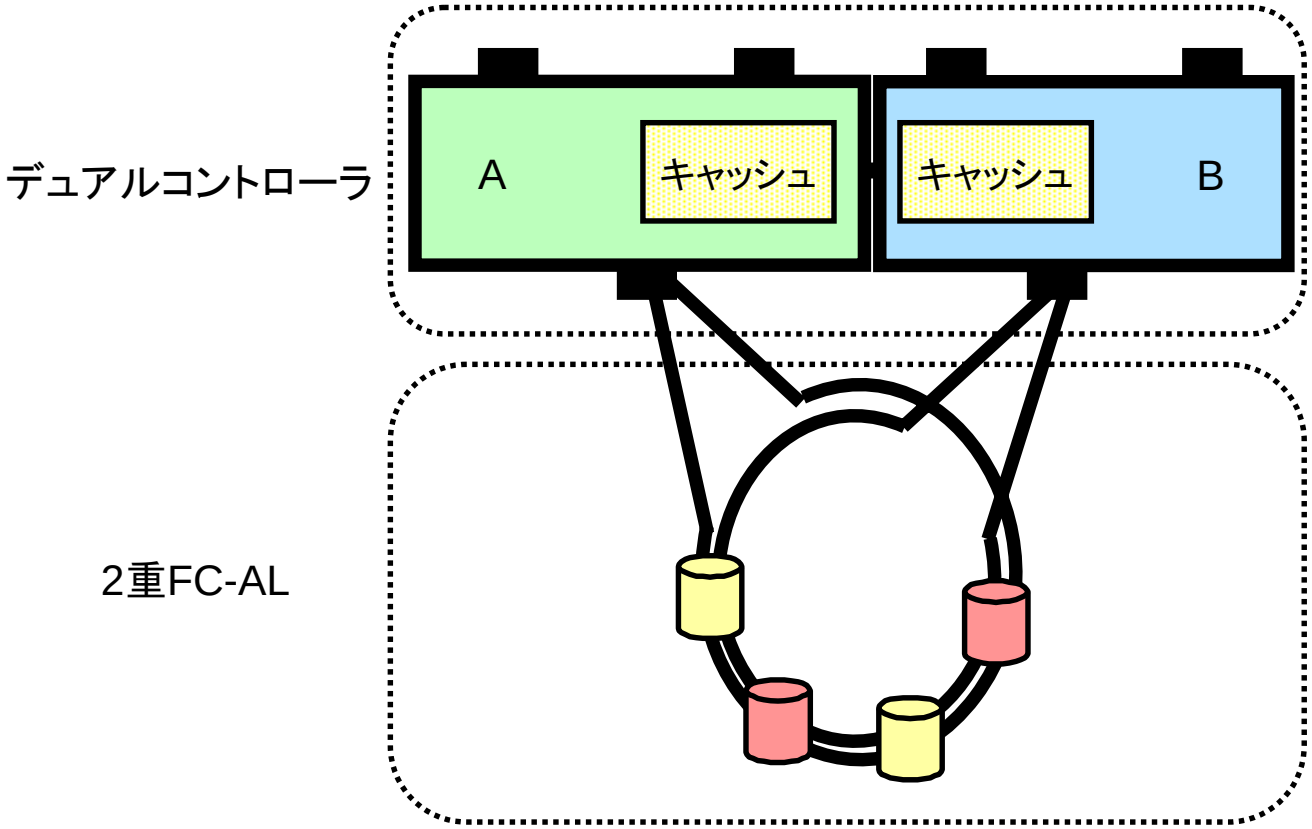


サーバーとの接続における考慮点

- 接続距離
  - SCSIは最大25m
  - ファイバーチャネルでは最大10km
    - リピータ使用により最大100km
  - IP技術を使用すれば…
- 接続可能なサーバー数
  - SCSIではチャネル当たり最大15装置
  - ファイバーチャネル
    - FC-AL構成では127装置
    - ファブリック・スイッチ構成では1600万装置
- サーバーとの接続の柔軟性
  - SCSIでは、固定されたチャネル接続
  - ファイバーチャネル
    - FC-AL構成では固定リング接続
    - FCPではスイッチを使い、ダイナミックにルートを変更させることが可能
    - ファブリックスイッチ構成では多重パス構成も可能

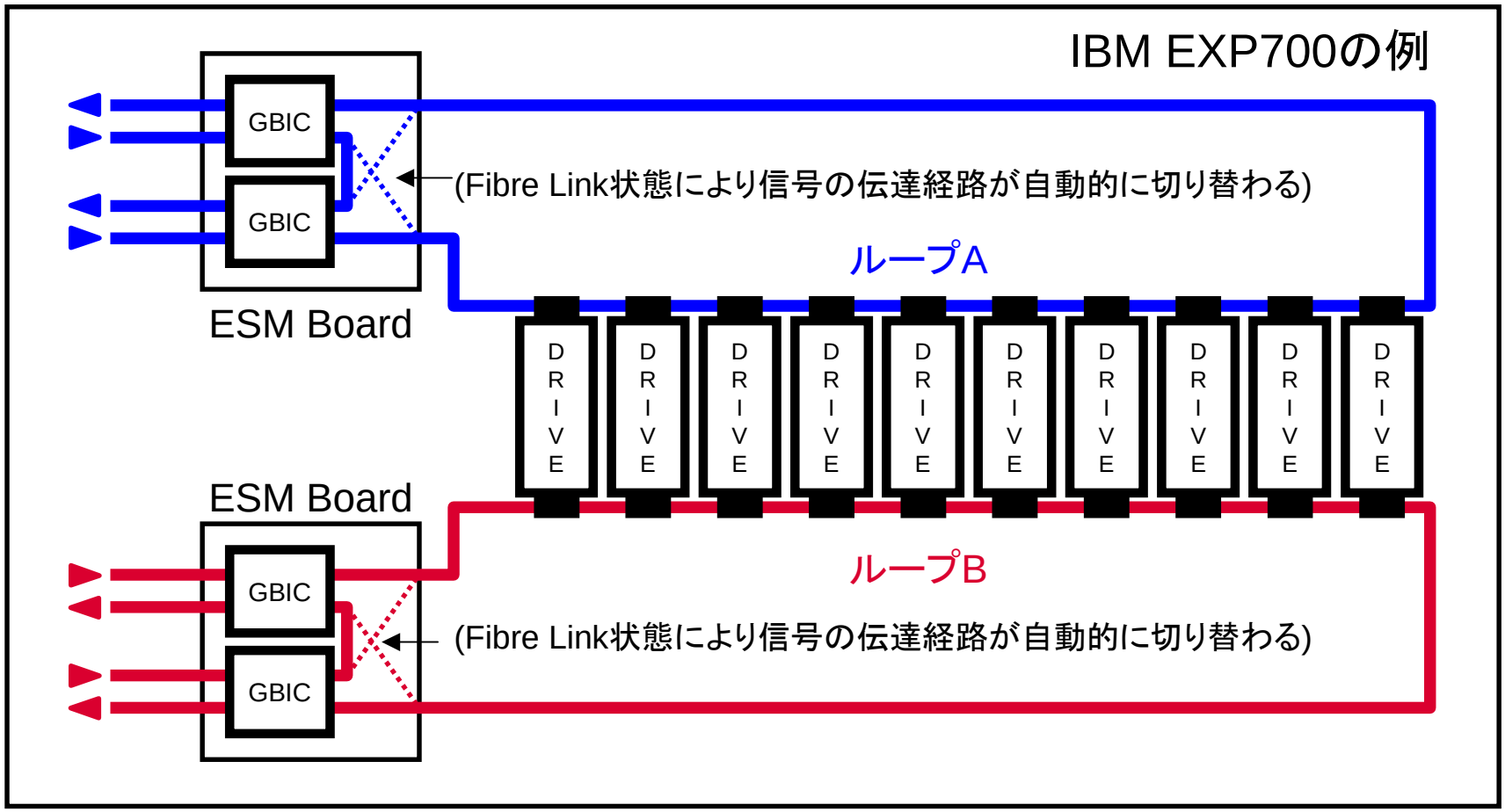
デュアル・コントローラ

- コントローラを2重化し、コントローラ故障時でも処理を継続可能
  - コントローラはホット・スワップ可能



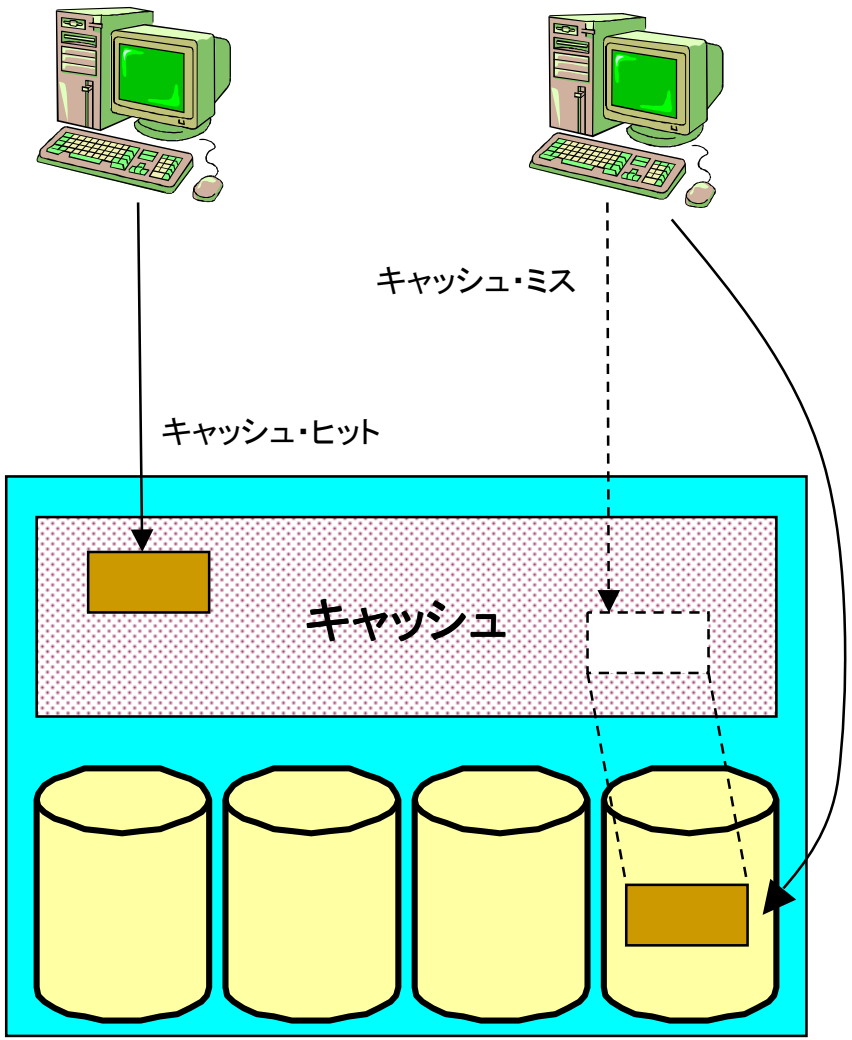
内部接続の高可用性:2重ループFC-AL構成

- ファイバーチャネルディスク拡張ユニットの内部は2重ループ構成になっている
- 各ドライブは両方のループからアクセス可能



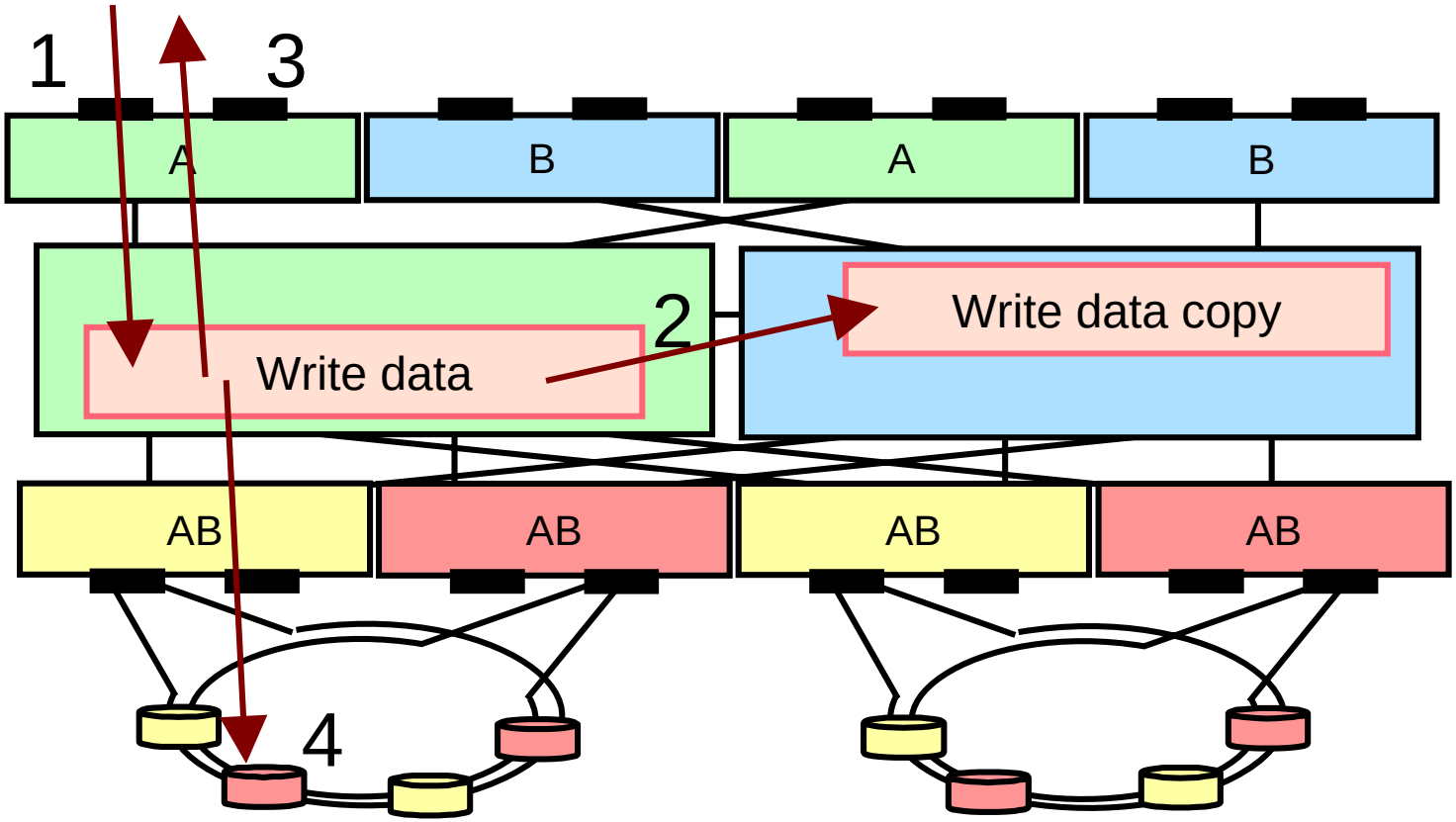
キャッシュ

- 半導体メモリーを制御装置に装備し、HDDの機械動作速度より高速に読み書きを行わせる事を可能にする機構
- キャッシュの種類
  - 読込みキャッシュ（Read only cache）
    - 一般に「キャッシュ」と言った場合はこれを指す
  - 書込みキャッシュ(Read/Write cache)
    - 通常読み込みキャッシュの機能も持つ
    - 書込みを高速化するためのキャッシュ
    - 通常二重化、不揮発性などのデータ保護機能が必要
- キャッシュ・ヒット
  - 目的とするデータがキャッシュ上にあった場合
  - 機械的動作が不要なため、高速な入出力処理が可能となる
- キャッシュ・ミス
  - 目的とするデータがキャッシュ上に無かった場合
  - HDDに直接読み書き動作をしなければならない

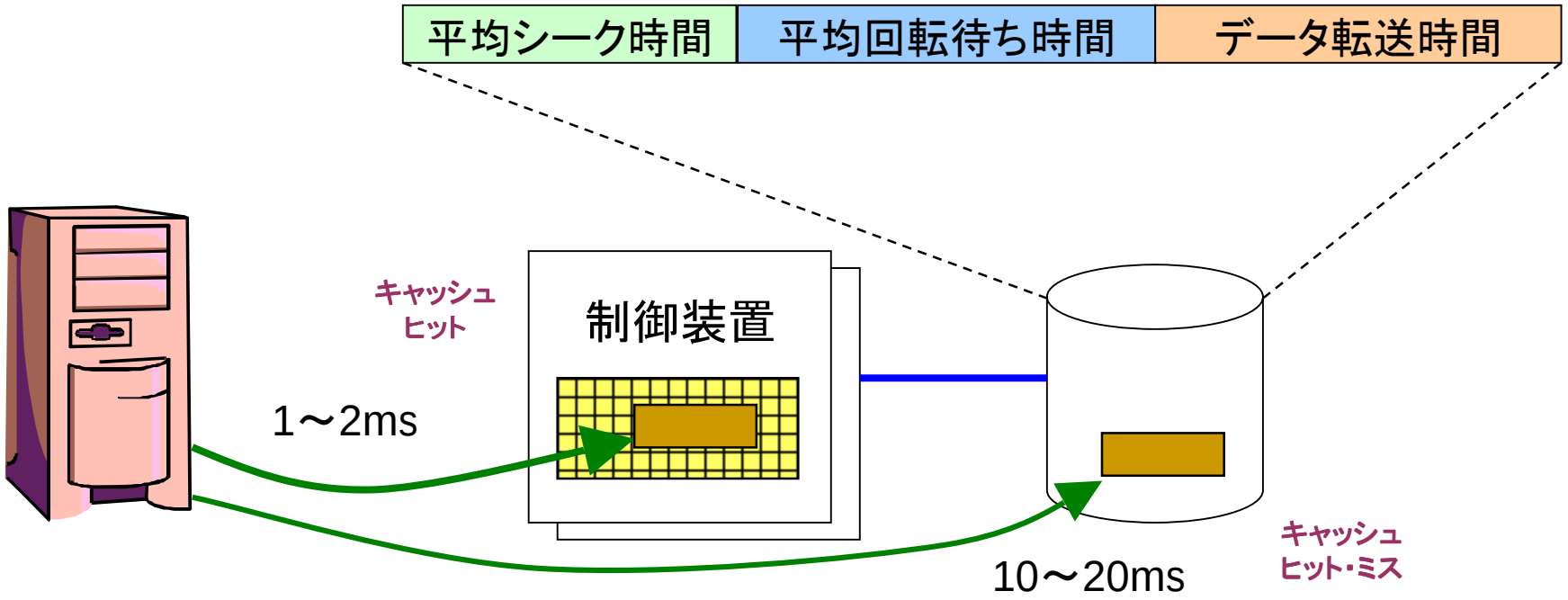


キャッシュ制御の例

書込みデータをミラーする実現例



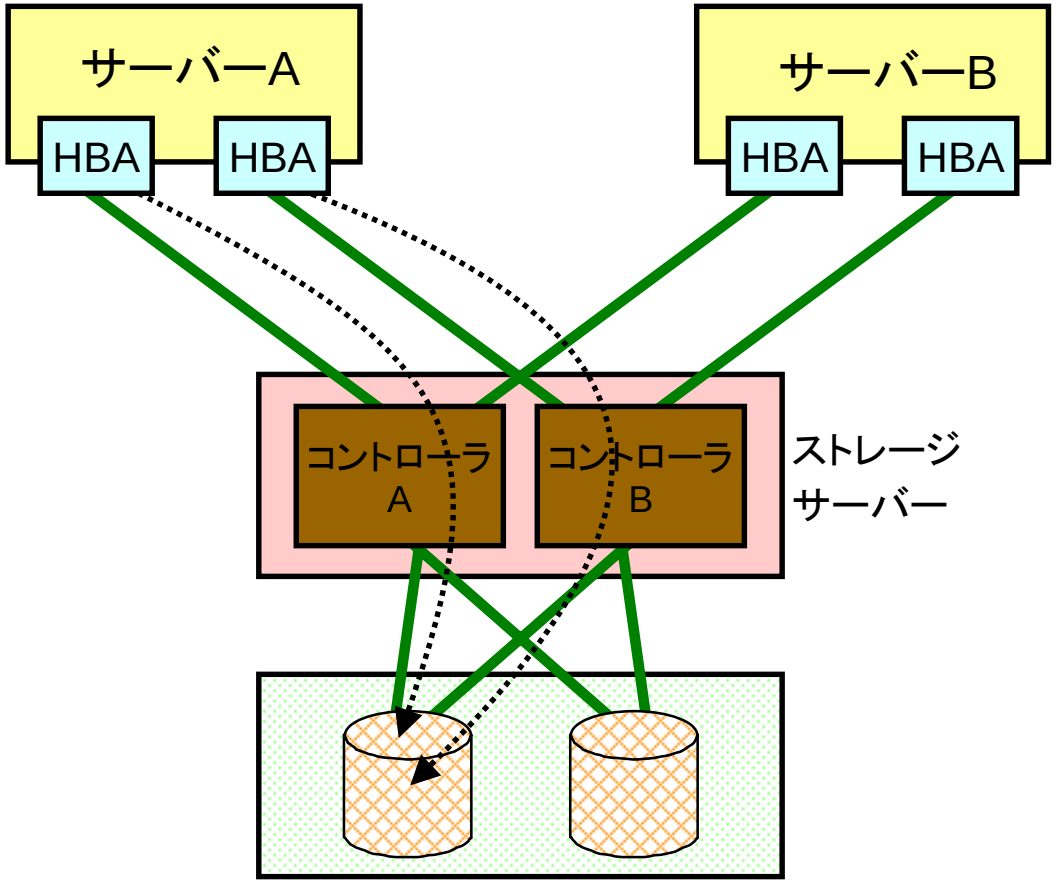
キャッシュ使用時のデータ・アクセス時間



- キャッシュ・ヒット率を考慮したレスポンス・タイム
  - 前提
    - ヒット時のレスポンスを1ms、ヒット・ミス時のレスポンスを10msと仮定
    - ヒット率を80%と仮定
  - レスポンス・タイムを計算すると・・・・
    - 1ms ＊ 80% ＋ 10ms ＊ 20% ＝ 2.8ms

マルチ・パス・アクセス

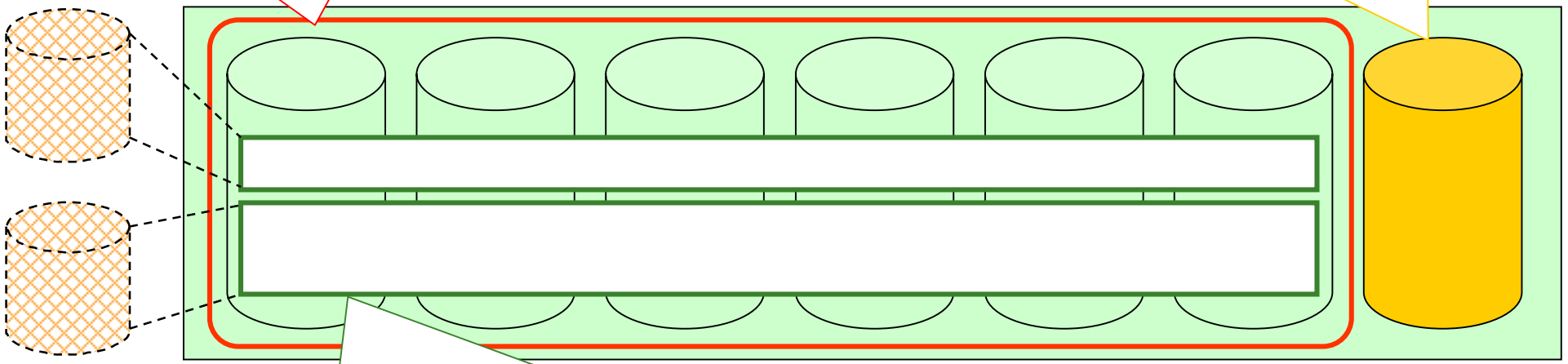
- サーバーからマルチ・パスによるアクセスをサポート
  - 可用性の向上
    - 自動フェールオーバ
  - パフォーマンスの向上
    - 負荷分散



論理ドライブ/LUN（論理装置番号）

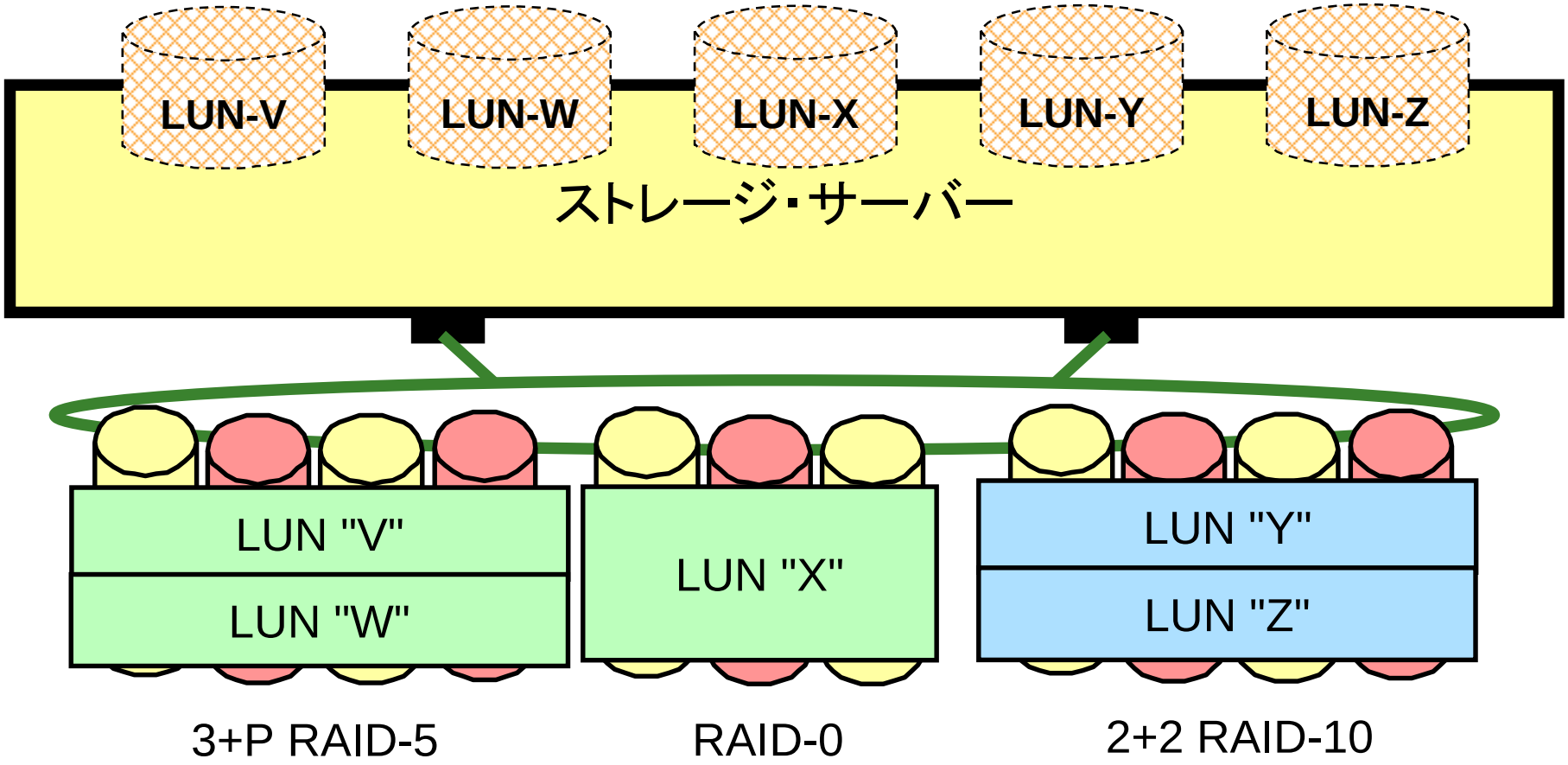
RAIDアレイ（RAID Array）
　複数の物理HDDをまとめたもの

ホットスペアドライブ（Hot Spare Drive）
障害に備えてスタンバイしているスペアディスク



論理ドライブ（Logical Drive）/論理装置番号（LUN: Logical Unit Number）
　RAIDレベルを指定して作成
　OSからは1台のドライブに見えます

LUNの構成例



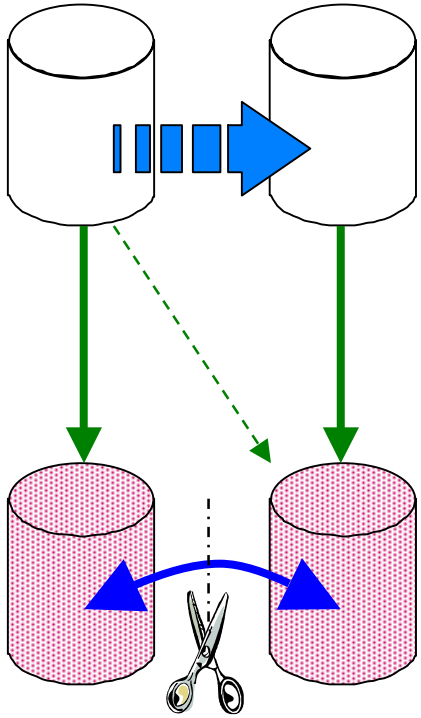
機種にもよるが、各種RAIDアレイ内にLUNを構成することも可能

高速コピー機能

- 高速コピー機能
  - サーバーを介することなく瞬時にデータの複製をストレージ・サーバーだけで作成することができる機能
    - 参考：ソフトウェアのファイル・システム・レベルで行う製品も存在する
  - スナップ・ショットとも呼ばれる
  - 用途
    - バックアップの取得
    - テスト・データの作成
    - データの並列処理
- 通常、同一ストレージ・サーバー内のボリューム(LUN)の複製を行う
  - ハードウェアは「ファイル」を認識できないためボリューム（SCSIディスクのイメージ）でコピーを作成する
- 大別すると3つの方式がある
  - どの方式も瞬時にコピーできる機能を提供するが、バックグラウンドの作業のやり方が異なる
    - スプリット・ミラー方式
    - バックグラウンド・コピー方式
    - ポインター・コピー方式

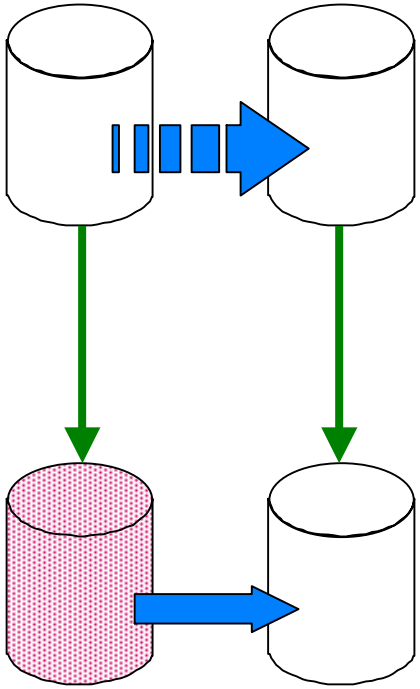
高速コピー機能の3つの方式

■ スプリット・ミラー方式



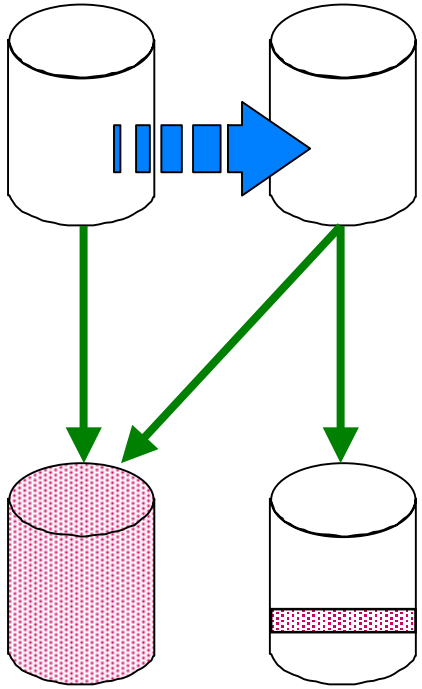
- 通常はミラーリングしている
- 高速コピー起動後にミラーを分割
- その後は個別に利用

■ バックグラウンド・コピー方式



- 高速コピー起動後、ポインターのみをコピーし、同じ内容に見せる
- 実データは後からバックグラウンドでコピーを行う

■ ポインター・コピー方式



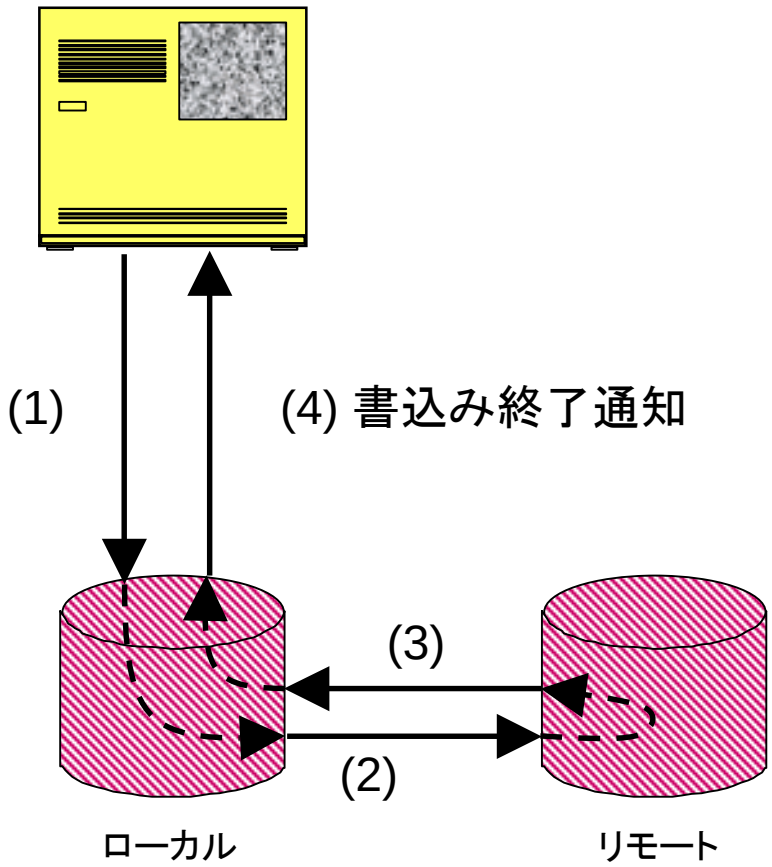
- 高速コピー起動後、ポインターのみをコピーし、同じ内容に見せる
- 実データはコピーしない
- 変更分のみ別エリアに保管

遠隔コピー機能

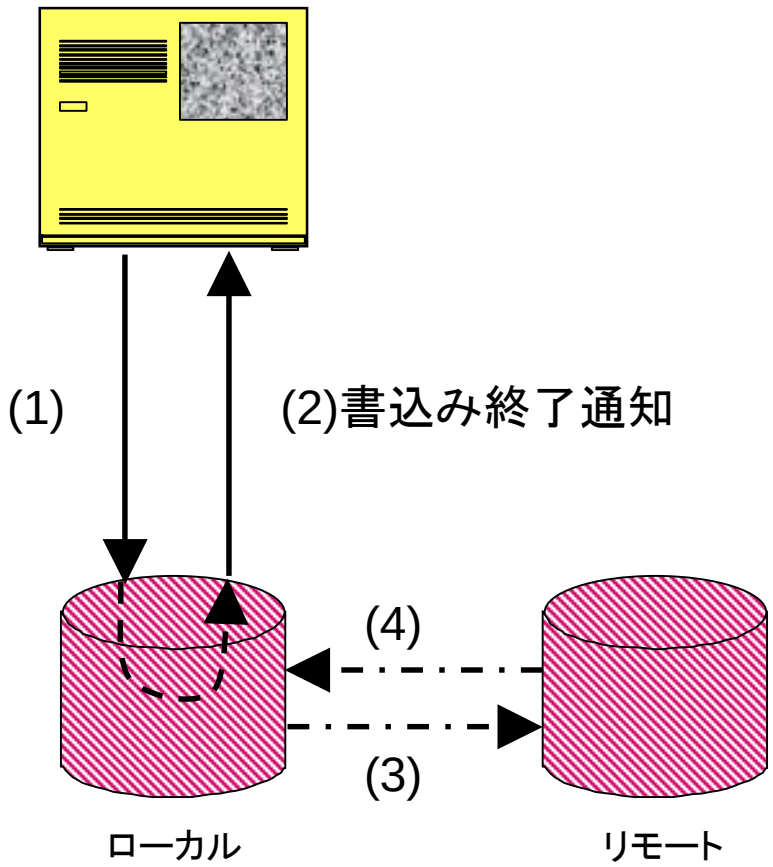
- 遠隔地間でのデータ複製機能
  - 実装方式の違い
    - ハードウェア方式
      - 一般に各ストレージ・サーバーの固有の機能のため、コピー元とコピー先は同一メーカ、同一機種である必要がある
      - サーバー資源を消費しない
    - ソフトウェア方式
      - ストレージ・サーバーの機種制限が無いというメリットがある
      - サーバー資源を消費し、システム・パフォーマンスへの影響を考慮する必要がある
  - 転送方式の違い
    - 同期方式
      - ローカルのディスクに書かれ、更に遠隔地のディスクへの書込みも確認された段階で書き込み終了とする方式
      - データの整合性が取りやすいが、パフォーマンスへの影響が大きい
    - 非同期方式
      - ローカルのディスクに書かれたことをもって書き込み終了とする方式
      - 遠隔地への転送は、非同期に転送される
      - パフォーマンスへの影響は小さいが、回復手順が複雑になる傾向がある

遠隔コピー: 同期方式と非同期方式

■ 同期方式



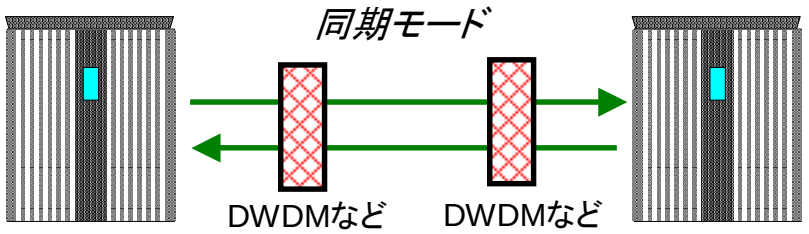
■ 非同期方式



遠隔コピー機能の実装例

IBM ESS PPRC（対等遠隔コピー）

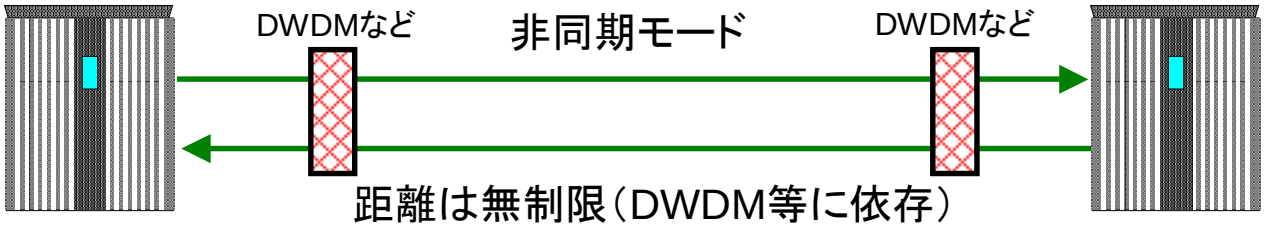
- 同期モードによる
- 最大103Kmまでの距離をサポート
  - 103Km以上が必要な場合はRPQ



距離は標準で103Kmまで

IBM ESS PPRC-XD（対等遠隔コピー拡張距離）

- 非同期モードによる
- 最大距離の制限は無し
  - 実際の制限はDWDM等の機能に依存



距離は無制限（DWDM等に依存）

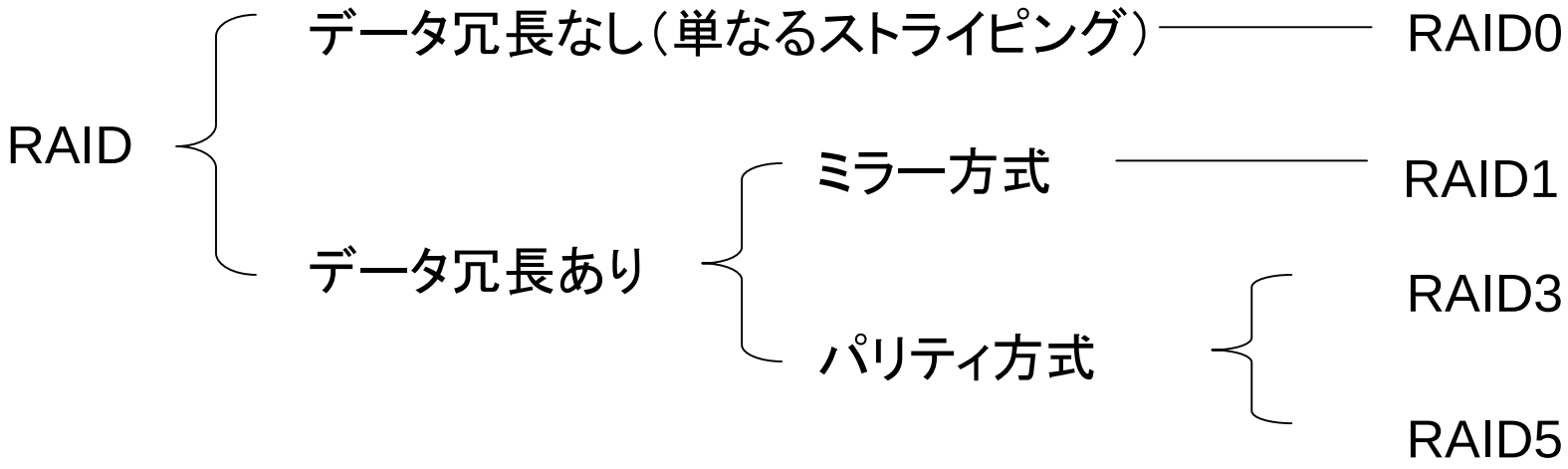
RAID

RAIDとは

- RAID=Redundant Array of Independent Disks
  - 本来は、Redundant Array of Inexpensive Disksの略であり、低価格であるが信頼性の低いHDDを組み合わせて高信頼化を実現することが目的
    - 各メーカーは一般に安いディスクは使っていないと言う意味で「Independent」を使う
  - HDDは稼働部が多いため、故障率の高いコンポーネント
  - HDDが同時に複数個故障する確率は低い
    - 単一HDD故障に対応できる仕組みができれば、可用性を向上できる
  - 筐体全体での故障やオペレーションミスによるデータ損失には対応できない
    - 外部装置（テープ装置など）へのバックアップは必要
- 一つのHDDでは不可能なことを、複数のHDDで実現する技術
  - 容量の向上
    - 複数台のHDDを1台のHDDとして取り扱う
  - 性能の向上
    - データを複数のHDDに分割、並列入出力することで性能を向上
  - 信頼性の向上
    - ドライブ間で分割した情報を重複して記憶することで、あるドライブにエラーが発生しても別のドライブから正確な情報を復元可能

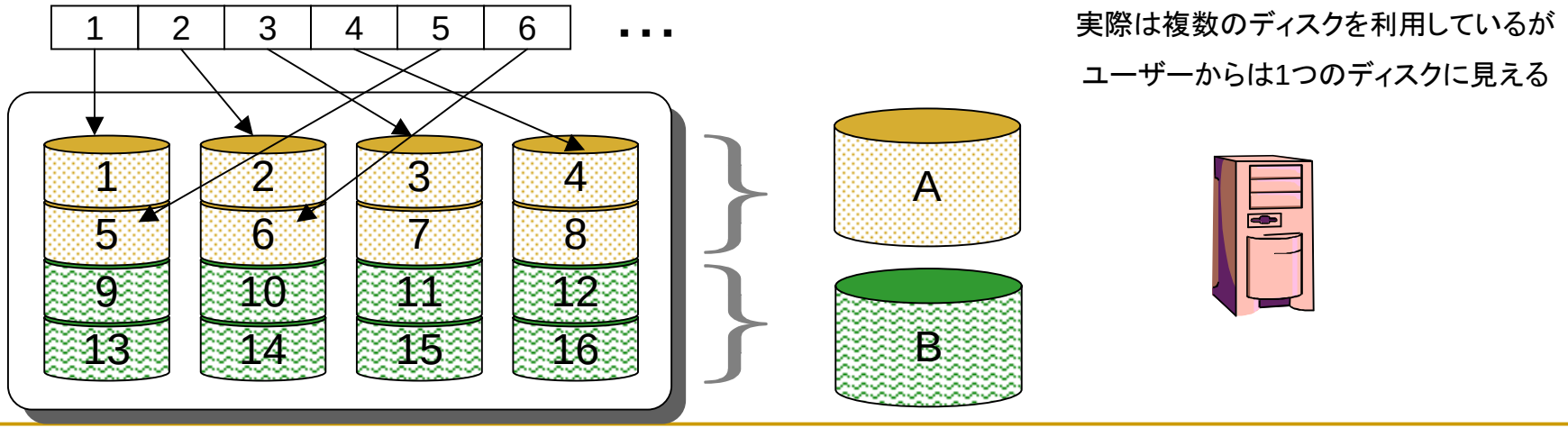
RAIDの分類

- 基本はRAID0～RAID5の6方式
  - RAID 0（ストライピングとも言う）
  - RAID 1（ミラーリングとも言う）
  - RAID 2（ECC適用方式）
  - RAID 3（パリティ保護 + ストライピング）
  - RAID 4（固定パリティ + データ単位でストライピング、）
  - RAID 5（ローテート・パリティ + データ単位でストライピング、）
  - RAID 10, RAID 1+0（RAID 1 とRAID 0 の組み合わせ）など



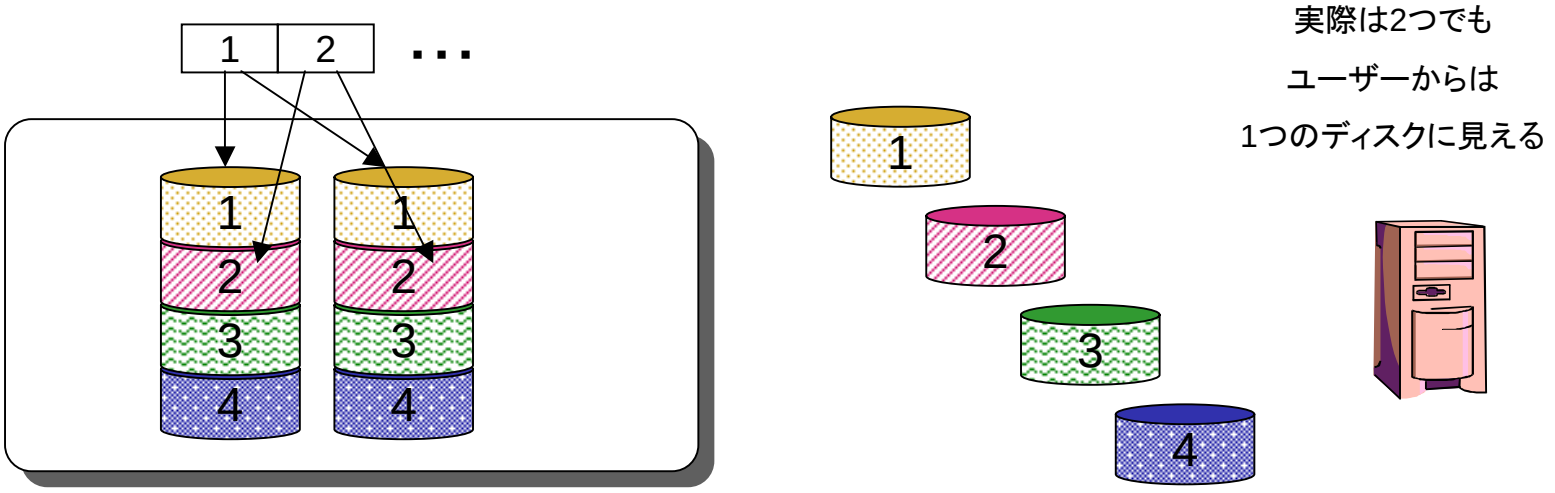
RAID 0 （ストライピング）

- 複数のHDDにデータをブロック単位で分散させて（ストライピング）記録させる方式
- 利点
  - 複数のHDDを並列に動作させるため高速に読み書きできる
    - 複数のHDDに対してコマンドとデータを送り、見かけ上シーク時間や回転待ち時間をなくす
    - HDDの数に比例して入出力のスループット性能が向上
  - 1台のHDDでは実現できない大容量のHDDを実現
- 欠点
  - 複数のHDDをデータの書き込みに使用するので、1台でもHDDが故障すると全データが読み書きできなくなる
  - 厳密には「RAID」とは呼びにくい（便宜上RAIDという用語が慣習として使われている）



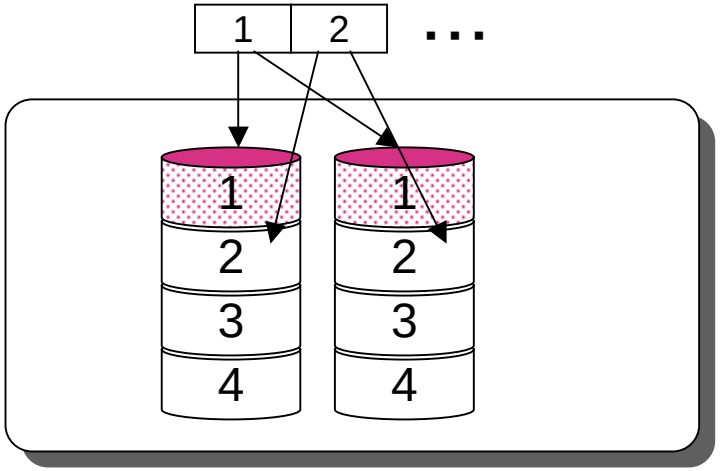
RAID 1 （ミラーリング）

- 2台のHDDに同じデータを記録し（ミラーリング）、常に同じ状態に保つ
- 利点
  - 1台が故障しても同じ内容のデータを記録したもう1台が残り、処理を継続することができる
- 欠点
  - 2台のHDDに同じデータを書き込むため多少オーバヘッドがある
  - 利用可能な容量は実装容量の半分となる

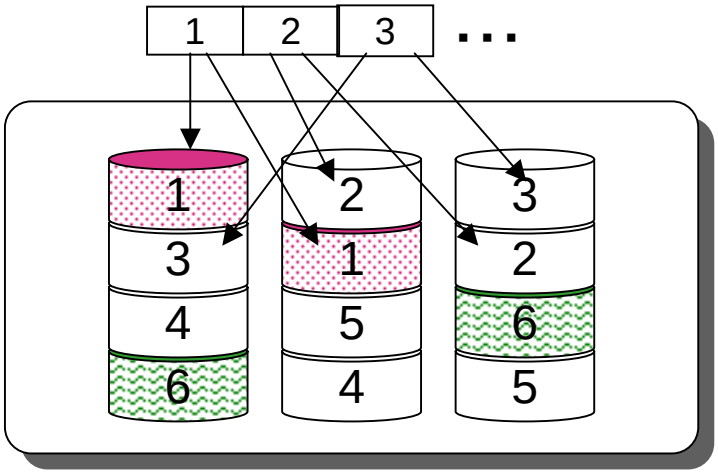


RAID 1E

- RAID 1の拡張版で、3台以上の物理ドライブを使用したミラーリングである
- 利点
  - ドライブ追加によって性能が向上する場合がある
  - 1台のHDDでは実現できない大容量のHDDを実現
- 欠点
  - 論理ドライブの容量はRAID1と同じく物理ドライブ容量の50%である



RAID1: 通常のミラーリング

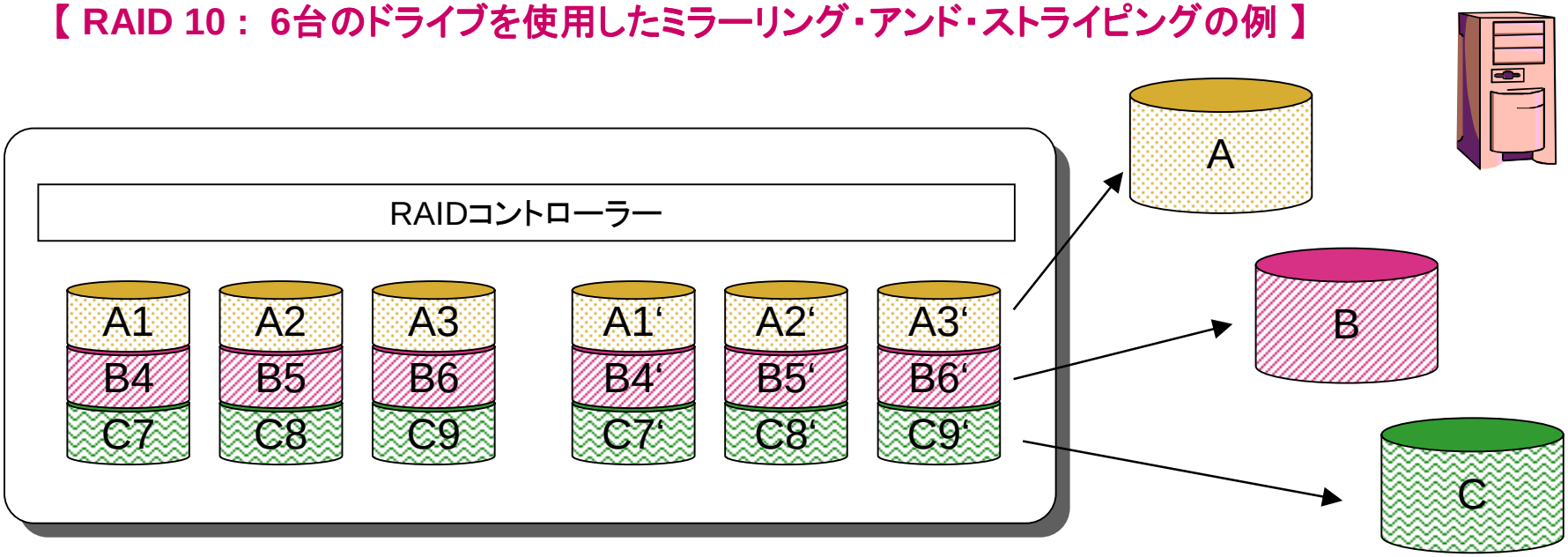


RAID1E: 3台のドライブを
使用したミラーリング

RAID 1＋0（RAID10）

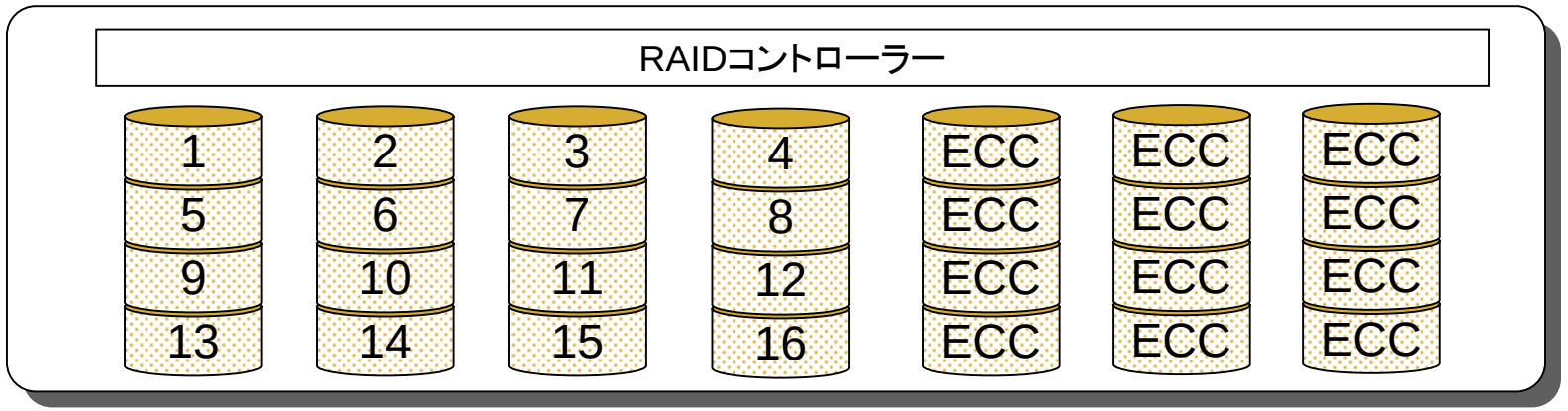
- RAID 1とRAID 0を組み合わせた技術である
  - 複数台の物理ドライブを使用したミラーリング＋ストライピングである
  - 論理ドライブの容量はRAID 1と同じく物理ドライブ容量の50%である
  - ドライブ追加によって性能が向上する場合がある

【 RAID 10： 6台のドライブを使用したミラーリング・アンド・ストライピングの例 】

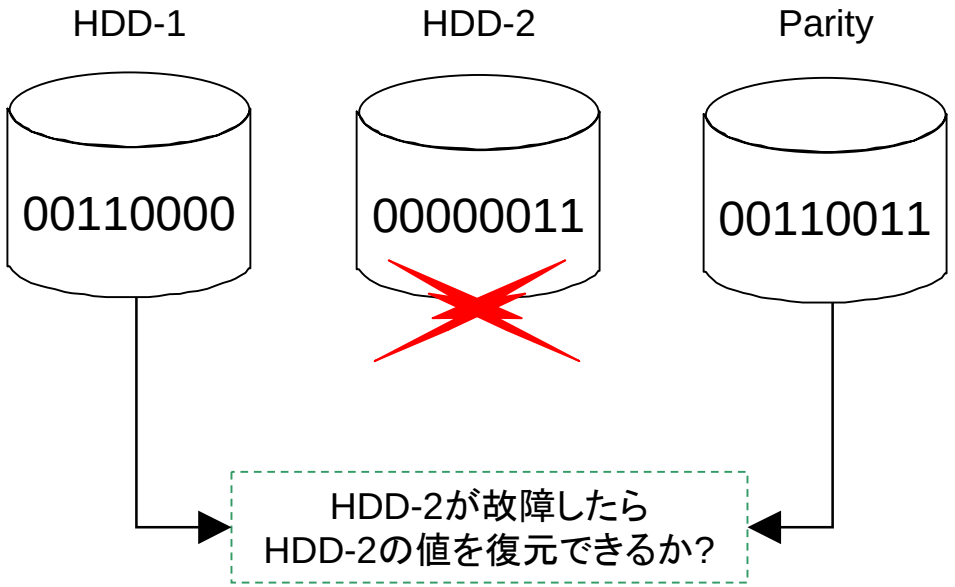


RAID 2

- メモリーなどで利用されているECCの手法をディスクに取り入れた方式
- 現在、実現された製品はない（はずである）
  - 製造回路が複雑になり、パフォーマンス及びコスト的なメリットも得にくいため



パリティを使ったデータ冗長化



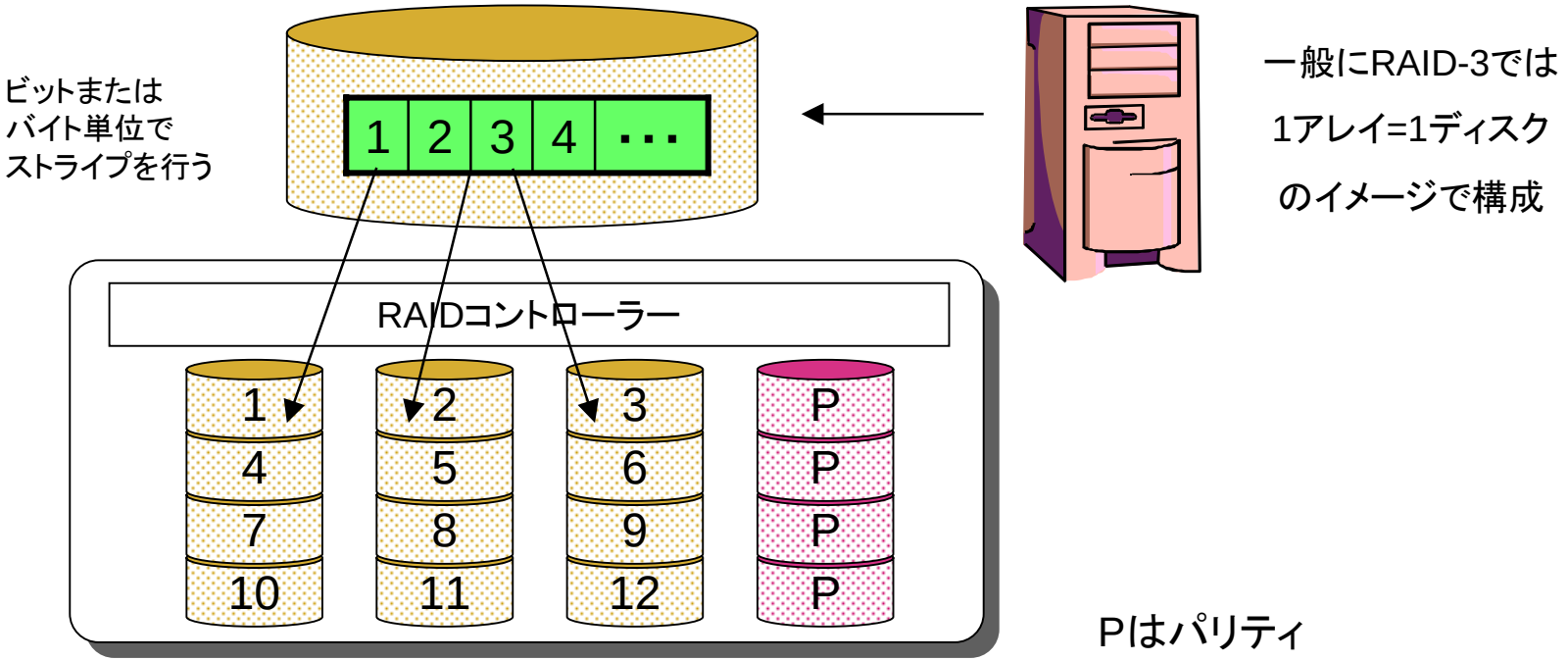
x	y	x XOR y
0	0	0
0	1	1
1	0	1
1	1	0

HDD-1	00110000	(0x30)
XOR（排他的論理和）		
HDD-2	00000011	(0x03)
パリティ	00110011	(0x33)

HDD-1	00110000	(0x30)
XOR（排他的論理和）		
パリティ	00110011	(0x33)
HDD-2	00000011	(0x03)

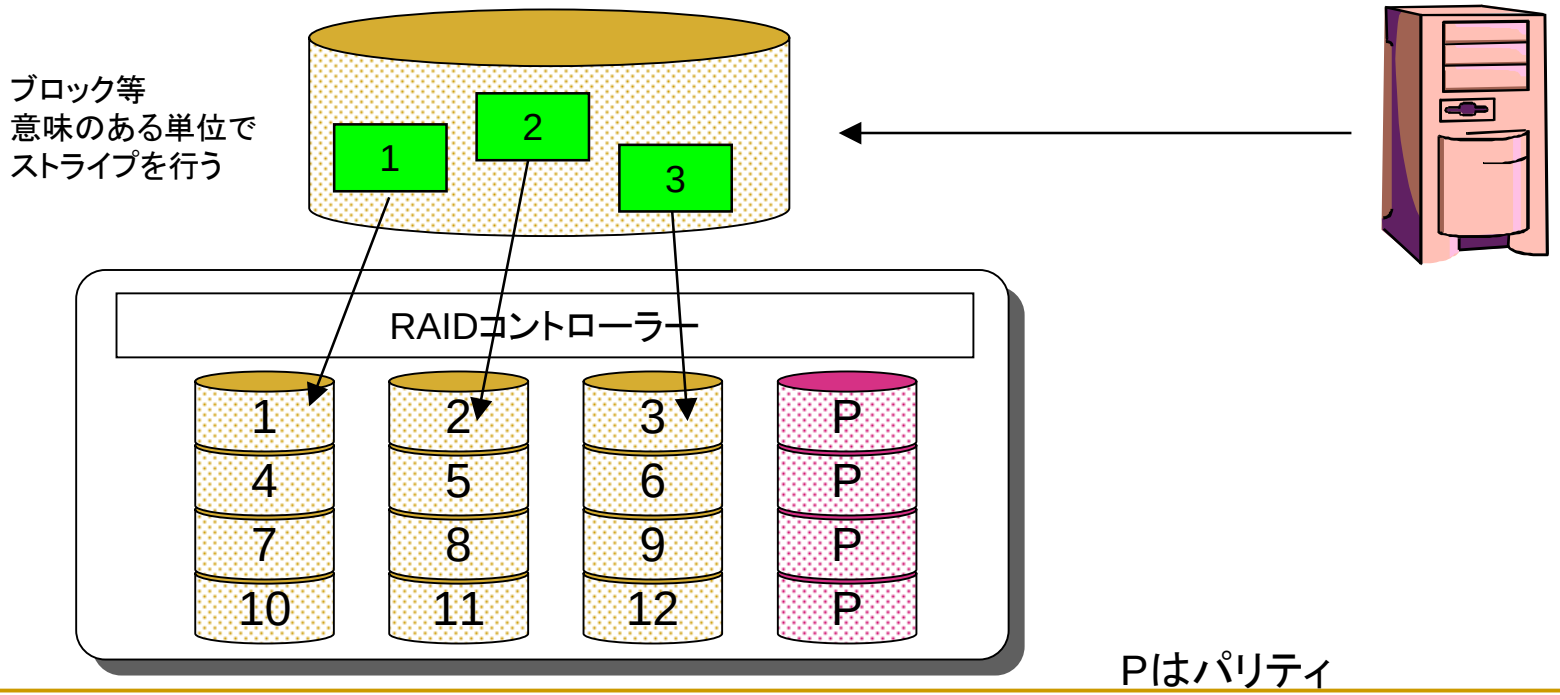
RAID 3

- 複数あるディスクのうち1台をパリティの記録に割り当て、他のディスクにデータを分散して記録する方式。
  - 意味のないデータの単位（例えば10バイト）に分割してストライプを行う
- どれか1台が故障しても交換してデータを復旧することができる
- 複数のディスクにはデータを分散して同時並行で記録するため、高速化もはかられる
  - 特に科学技術計算のアレイの読込み、書出しなどに向く



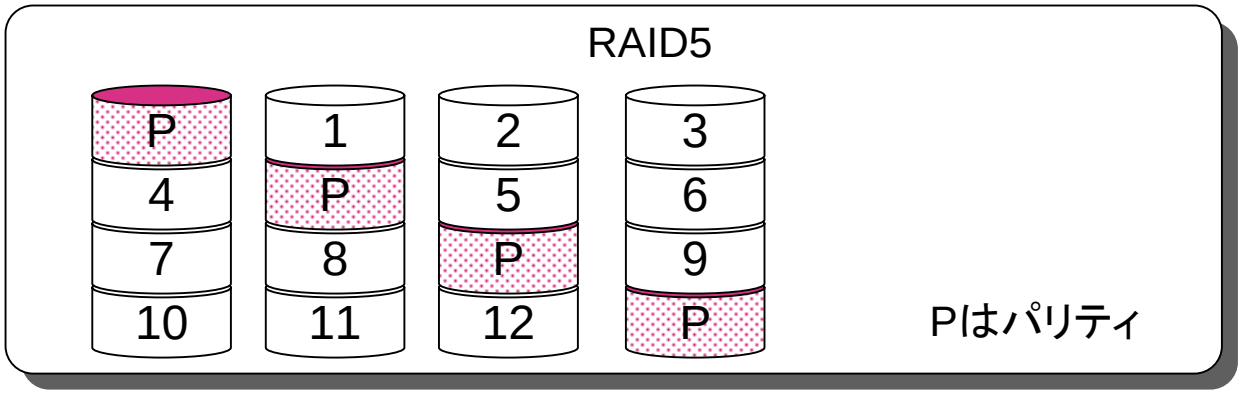
RAID 4

- 意味のある単位（ブロック、セクターなど）でストライプを実施
  - パリティーは固定
  - RAID 3 の不得手なランダム・アクセスに対応可能
- パリティ・アクセスにボトルネックが発生してしまうため、実装された製品例は少ない



RAID 5

- RAID 4の改良版、パリティーをローテートする点が特徴
- アレイの全てのドライブを越えてデータとパリティをストライプする
  - 意味のあるデータの単位（例えばブロック）に分割してストライプを行う
  - 並列アクセス（トランザクション・タイプのアクセス）に向いている
- 最も実用的なRAID方式と一般に考えられている
- 実際に使用できるディスク容量はディスク1台分（パリティデータ記憶域用）だけ少なくなりなる
- 少なくとも3台以上の物理HDDが必要
- アレイ中の1台の物理ドライブに障害が発生しても残りの物理ドライブでサービスを継続
  - ホットスペアドライブ、あるいは、交換したドライブを使ってRAID5を再構成可能



一般論としてのRAIDレベル比較

一般的なファイルサーバーの活動の統計結果： 読み取り80% 、書き込み10%、検索10%

RAIDレベル	データ冗長	ドライブ容量の使用率	ランダム読取性能	ランダム書込性能	順次読取性能	順次書込性能	コスト
RAID 0	なし	100%	△	△	◎	◎	高
RAID 1	あり	50%	○	○	○	○	高
RAID 1E	あり	50%	○	○	○	○	中
RAID 3	あり	67% - 94%	△	△	◎	◎	低
RAID 5	あり	67% - 94%	◎	△	◎	◎	低

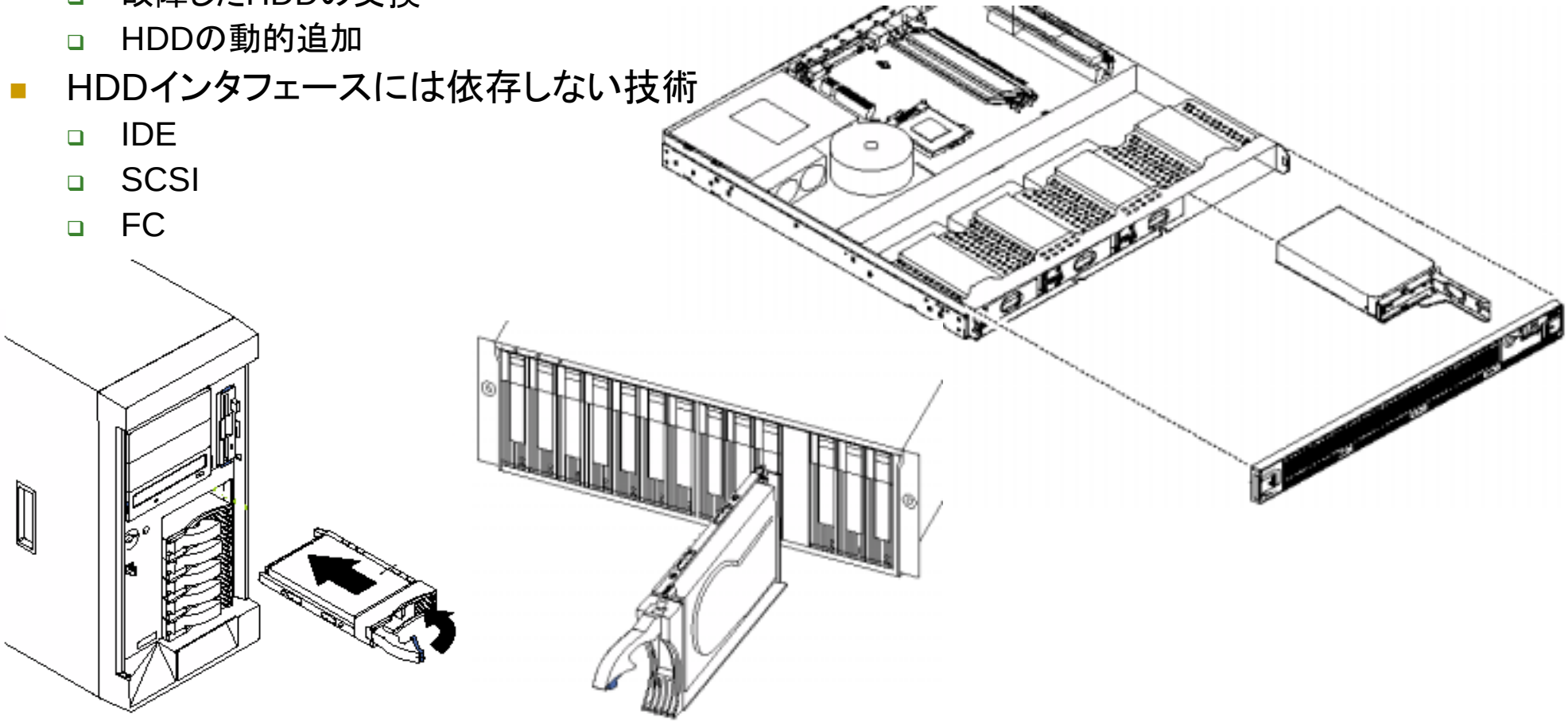
これら比較はあくまで参考と考えてください。実際の各社製品においては、
各種仕組みの実装により、この表に当てはまらないケースもある

性能良 ←→ 性能悪

◎ ＞ ○ ＞ △

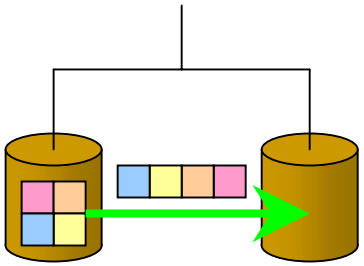
ホット・スワップ機構

- 通電中にドライブを取り外したり、取り付けたりできる仕組み
  - 故障したHDDの交換
  - HDDの動的追加
- HDDインタフェースには依存しない技術
  - IDE
  - SCSI
  - FC

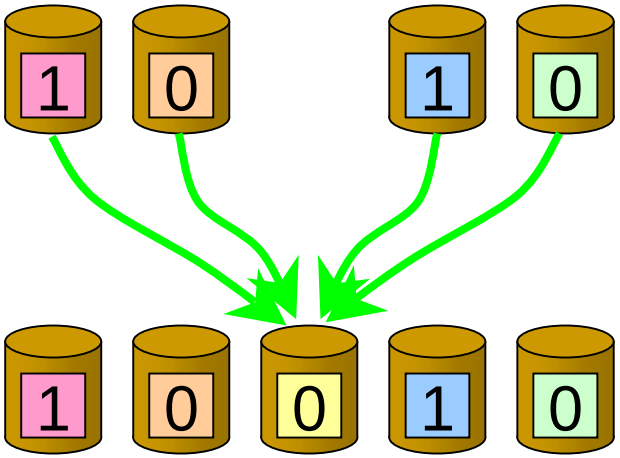


RAIDのリビルド

- RAID1、RAID5を構成しているHDDが故障した場合、その他のHDDが壊れる前に、そのHDDを交換すれば、元の信頼性を取り戻すことができる。
- 交換後、本来そのHDDにあるべきデータを再構築することをRebuild（リビルド）と言う



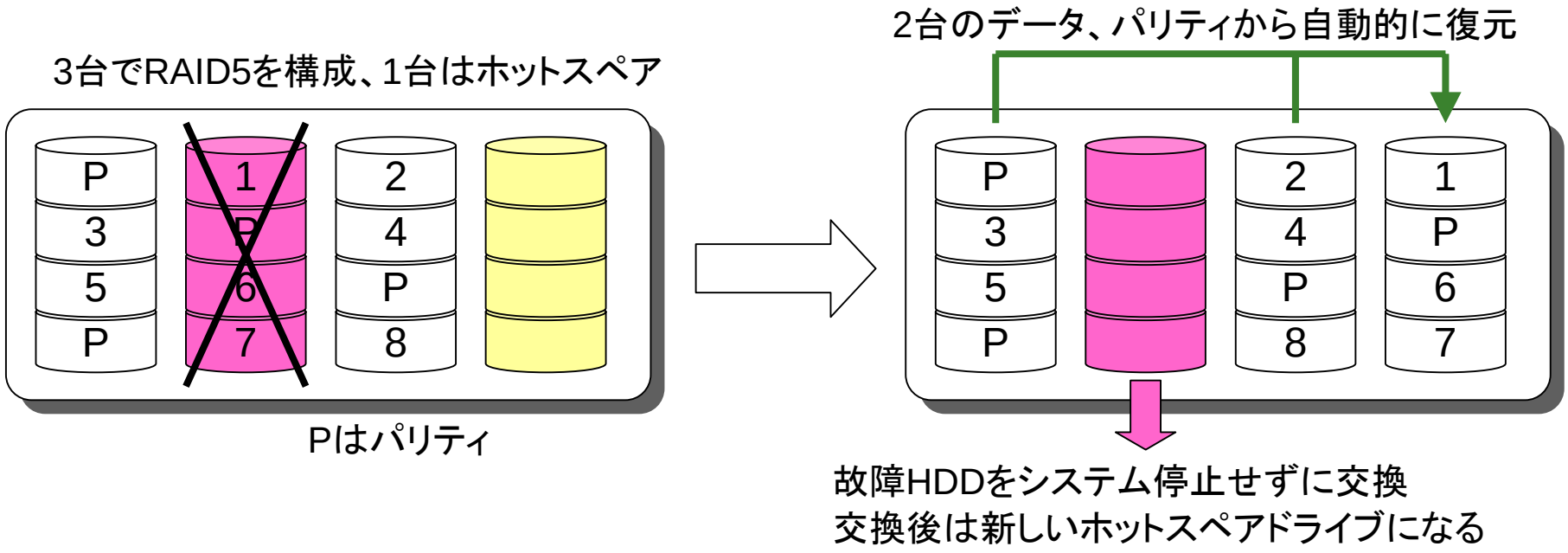
RAID1の場合、
データのコピー



RAID5の場合、他の全ての
HDDのデータを用いて、Parity
生成と同じ容量でデータを再生

ホット・スペアー・ドライブ

- ホット・スペアー・ドライブ（通常は使用されずスタンバイしている）をあらかじめ確保しておけば、HDD故障時にはオペレータが介在することなく、自動的に故障したHDDに代わりホット・スペアー・ドライブを使用してデータの復元が行われる
  - ホット・スペアー・ドライブの構成、装備数は一般に任意にカスタマイズ可能（機種にも依存）
  - RAID 0ではリビルドができないため、スペアーを準備する意味は無い
  - ホット・スペアー・ドライブの機能が無い装置も、当然世の中には存在する

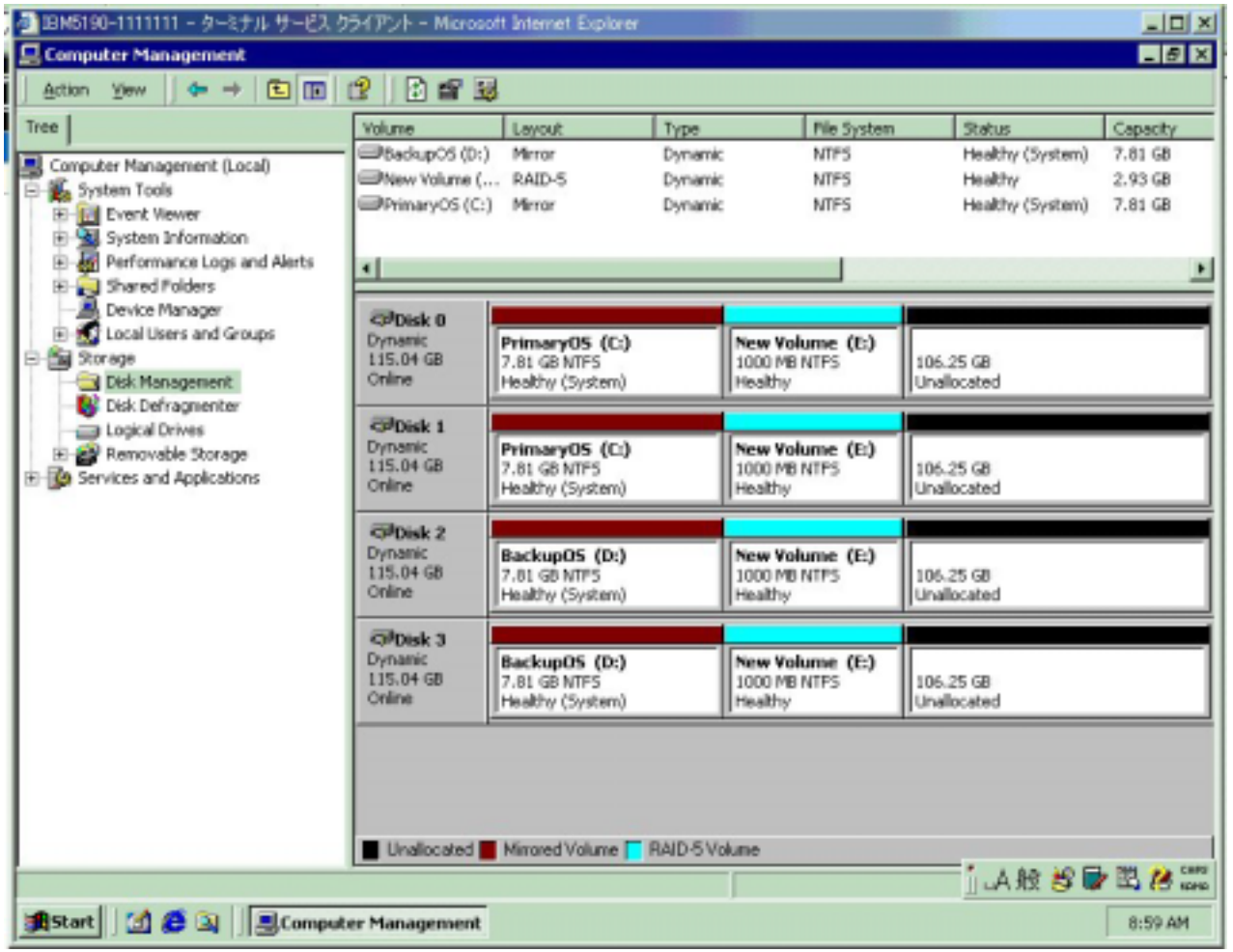


RAIDの実現方法

- サーバー直結ストレージの場合
  - ソフトウェアRAID
    - Windows 2000 Serverなどに実装されている方式で、ソフトウェア処理により、RAID1やRAID5を実現
  - ハードウェアRAID
    - RAIDアダプタ・カードなどを使用してアダプタ上のプロセッサによりRAID機能を実現
    - OSからは通常SCSIアダプタとして認識される
    - 専用のRAID制御ツールによりRAIDを構成する
- SAN接続ストレージ・サーバーの場合
  - ストレージ・サーバー内でRAID機能を実現
  - ハードウェアとソフトウェアの組み合わせ
    - 近年の高機能ストレージ・サーバーでは、ストレージ装置内での仮想化技術が取り入れられている

ソフトウェアRAIDの例

- Windows 2000サーバーなどでは、Windows NTFSのダイナミックディスク機能を使用し、ソフトウェアRAIDを実現



RAIDコントローラー・カードの例

- ハードウェアRAID PCIカード（現時点では、Ultra160 SCSI規格のものが多い）を搭載し、RAIDをハードウェアにより実現
  - RAIDアレイ構成情報の保持
  - パリティ計算
  - RAID自動復元、ホット・スペアリングなど