データ工学特論（1）

情報基盤センター
天野　浩文

この講義のテーマ

- 計算機システムにおける二次記憶装置の高速化・高機能化
  - 高速大容量二次記憶装置（システム）
    - **ディスクアレイ**
    - **並列ファイルシステム**，並列入出力機構
  - 高機能二次記憶装置（システム）
    - **インテリジェントディスク**
    - **ネットワークストレージ**，SAN (storage area network)
- データベースシステムの高速化
  - **並列問い合わせ処理**
  - **オンライントランザクション処理**

これらのテーマを通じて，現代の計算機システムにおける二次記憶装置の持つ役割を探っていく

講義の進め方

- スライドによる講義．出席者には毎回コピーを配布．
- 参考書は特に指定しない．
- 学期末（2月）にレポート提出を求める
- その他の筆記試験は行わない

背景

- **ハードウェア技術の進歩**
  ⇒1台の計算機に搭載できる資源（CPU, メモリ，ディスク）の数・容量の増大
- **大規模科学技術計算の発達**
  ⇒より高精度の計算（＝より大きなデータに対する計算）
  ⇒大容量化，高速化の必要性高まる
- **データベースの応用範囲の拡大・高度な機能の要求**
  ⇒処理量の増大，障害発生時の影響の深刻化
  ⇒高速化・高信頼化の必要性高まる

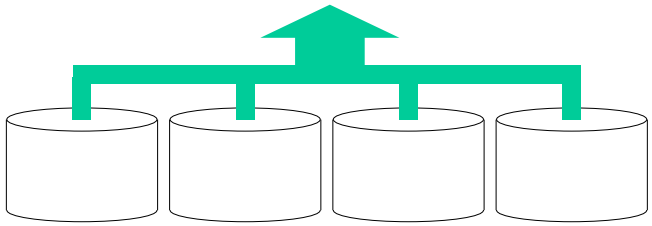
急速に大容量化する二次記憶中のデータの処理の高速化・高信頼化

課題

- データ入出力の高速化
  - **ディスクアレイ**，RAID (redundant array of inexpensive disks)
  - 並列計算機のための**並列ファイルシステム**
  - 並列アプリケーションプログラムのための**並列入出力インタフェース**
- 二次記憶装置の高機能化（処理の高速化を目指して）
  - **インテリジェントディスク**
  - **ネットワークストレージ**，SAN (storage area network)
- データベース処理の高速化
  - **並列データベースシステム**…データベース管理システム (database management system, DBMS) の並列化
    - 並列問い合わせ処理
    - オンライントランザクション処理

ディスクアレイの原理

- 複数台のディスクを並べて，
  - 「バンド幅」を上げる
  - スループットを上げる



（計算機側からは通常のディスク装置と同じように見える）

ディスクアレイの重要性（1）

- CPUとディスクの動作速度の向上
  - RISCプロセッサ…年に50%以上（最近は少し「減速」）
  - ディスクの機械動作部分…年に10%以下
  - ディスクの転送速度…年に約20%

  ↓

  CPUとディスクの速度差はますます拡大
- アプリケーションの高度化・多様化
  - より精密な科学技術計算
  - マルチメディアデータ…音声，画像 etc.
  - 複雑な構造を持つ膨大なデータ…CAD, CIM etc.

  ↓

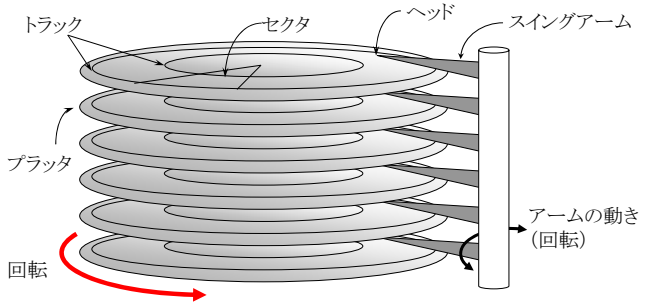
  プログラムがアクセスするデータの長大化・巨大化
- **高速な二次記憶装置の開発が極めて重要！**

ディスクアレイの重要性（2）

- ディスクの台数の増大がもたらすもの
  例：MTTF (mean time to failure) が 200,000時間（約23年）のディスクユニットが100台接続されたとき
  …全体では 2,000時間（＝約3ヶ月）のMTTF
  - 耐故障性の急激な低下

多数のディスクを接続する場合には，冗長性を付加して故障時の対策を立てておく必要がある．

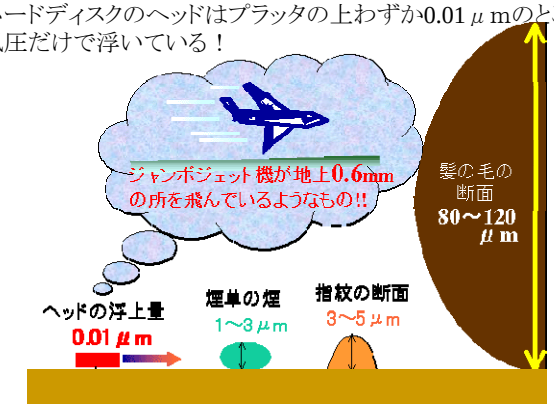
ハードディスク装置に関する基礎知識（1）



- シリンダ…同一のアーム角度に対するトラックの集合
- シーク時間…ヘッドの移動に要する時間（1～30msec）
- 回転遅延…セクタにヘッドが到達するまでの回転に要する時間（～10msec）
- データ転送時間…ヘッドでの読み出し／書き込みを行う時間

ハードディスク装置に関する基礎知識（2）

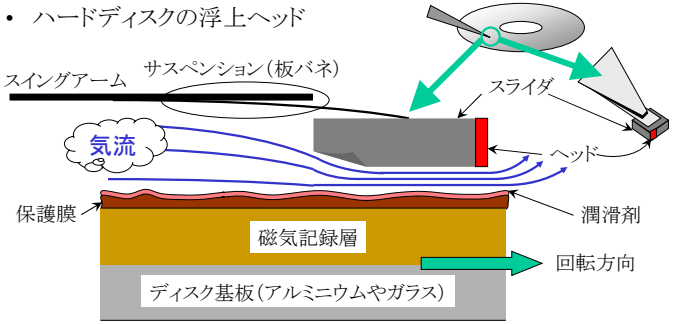
- ハードディスクのヘッドはプラッタの上わずか0.01μmのところに風圧だけで浮いている！



富士通研究所「やさしい技術講座」より

ハードディスク装置に関する基礎知識（3）

- ハードディスクの浮上ヘッド



現在のハードディスクの多くは，ケースの中に潤滑剤とともに密封（エアフィルターを通して多少の空気の出入りはあるが）されていて，停止時はヘッドがメディア表面の非記録ゾーンに接触している．

ハードディスク装置に関する基礎知識（4）

- ハードディスクの進歩
  - 記録密度の向上＝容量・転送速度の増大
  - 回転速度の向上＝転送速度の増大
- **ハードディスクの宿命：故障**
  - 装置の信頼性を向上させるために，普通は，機械動作部分を減らすのだが．．．
  - ハードディスクでは，**機械動作部分が必須**であり，これをなくすことはできない
    - プラッタを回すためのスピンドルモータ
    - ヘッドを動かすためのボイスコイルモータとスイングアーム

ディスクアレイの基本概念（1）

- ハードディスク単体の容量や転送速度は向上しているが，さらに向上させたい場合には，**並列化**が必要
- **データストライピング** (data striping)
  - データを複数のディスクに分散させ，全体の転送速度あるいはスループットを上げる
- 分散のさせ方
  - 細粒度…ビット単位
    転送速度は上がる
    小さなアクセス要求に対しても全ディスクに読み書きする
  - 粗粒度…ブロック単位
    小さな要求は一部のディスクに対するアクセスで処理
    大きな要求には細粒度と同様に高い転送速度を達成

ディスクアレイの基本概念（2）

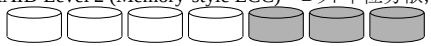
- **冗長性** (redundancy)
  - 複数のディスクユニットを搭載することによる信頼性の低下を防ぐ.
- 冗長情報のコード…パリティ，ハミングコードなど
- 冗長情報の配置のしかた
  - 集中型…冗長情報を少数のディスクにまとめる
    それらのディスクは hot spot になる
  - 分散型…冗長情報を全ディスクに一様に分散させる
- RAID **(redundant array of inexpensive disks)**

RAIDの分類 (Chen, Lee, Gibson, Kats, Patterson 1994)

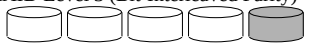
RAID Level 0 (Nonredundant) ブロック単位分散，冗長性なし

RAID Level 1 (Mirrored) ブロック単位分散，単なる二重化

RAID Level 2 (Memory-style ECC) ビット単位分散，Hamming符号



RAID Level 3 (Bit-Interleaved Parity) ビット単位分散，パリティ



RAID Level 4 (Block-Interleaved Parity) ブロック単位，集中パリティ

RAID Level 5 (Block-Interleaved Distributed Parity) ブロック単位，分散パリティ

プラッタの表示があるものはブロック単位，ないものはビット単位で分散

網掛け部分は冗長情報を表す

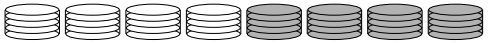
RAIDの分類についての補足

- “RAID Level *n*” を省略して，“RAID *n*” ということも多い.
- 商用RAIDのカタログなどでは，Chen, Lee, Gibson, Kats, Pattersonの分類とはやや異なる表記が用いられることも多い.

カタログ等でよく用いられる表記

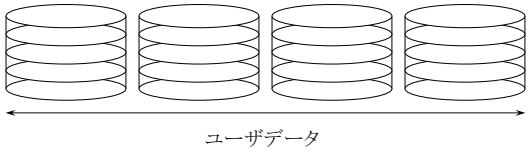
RAID 1 ストライピング無し，単一ディスクの二重化

RAID 0+1 ブロック単位分散＋二重化 ＝ RAID Level 0の二重化



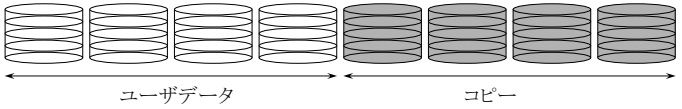
RAID Level 0 (Nonredundant)

- データを**ブロック単位でストライピング**
  冗長情報は無し(その意味では真の"RAID"ではない)
  - 冗長情報がないので記憶効率はよいが，信頼性は低い．
  - しかも，冗長情報がある RAID Level 1 に読み出し速度で負けることがある．



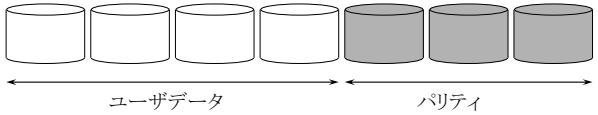
RAID Level 1 (Mirrored)

- ディスクの故障に備え，**ユーザデータの完全なコピーを持つ**
  - **ブロック単位でストライピング**
  - シーク時間と回転遅延の短いほうのセットを選択するようにすると，RAID Level 0 よりも高い読み出し性能を達成できる
  - 同じ実効容量を達成するためには，RAID Level 0 の倍の台数のディスクユニットが必要



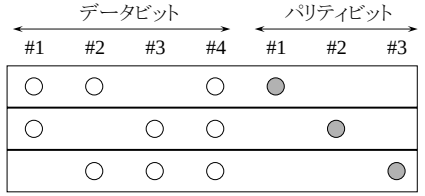
RAID Level 2 (Memory-Style ECC)

- **ビット単位でストライピング**
- メモリでエラー訂正符号 (Error-Correcting Code, ECC) として用いられる**ハミング符号**を採用
  - 単一のディスクユニットの障害を検出・回復できる



ハミング符号 (Hamming Code)

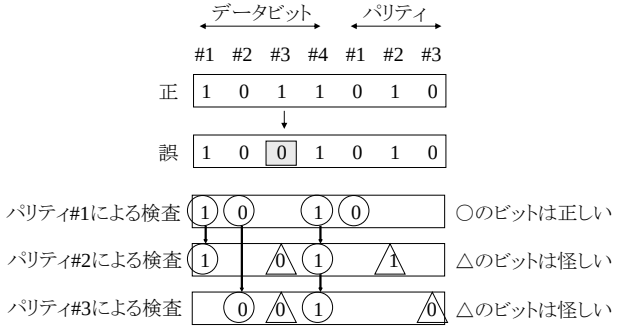
- ビット列をいくつかの部分集合にわけ，それぞれにパリティビットを付ける



- どこかのビットに異常があるときは，**いくつかのパリティビットで異常が観測される**
  - その**パターンで異常箇所を特定**できる

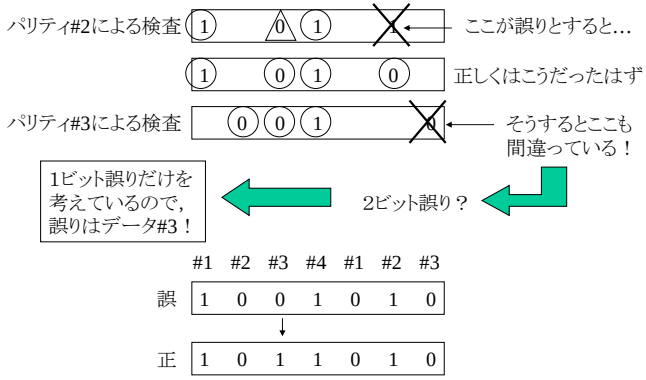
ハミング符号による誤りの検出と訂正（1）

- 正しいデータが1011であるのに，1001になってしまった場合



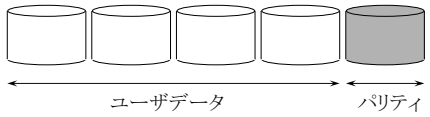
誤っている可能性があるのは△印のついた3個所！

ハミング符号による誤りの検出と訂正（2）



RAID Level 3 (Bit-Interleaved Parity)

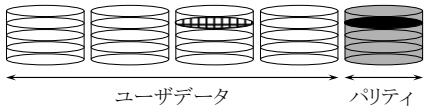
- 通常，ディスクコントローラは，どのディスクに障害が発生したかを認識できる
  - メモリの誤り訂正とは異なり，**パリティビットは1つあればよい**



- ビット単位で分散，各ストライプにパリティビットを1つ
- 高い転送速度が必要であるがスループットはあまり重要でない場合に用いられる
- RAID Level 4, 5 に比べて単純

RAID Level 4 (Block-Interleaved Parity)

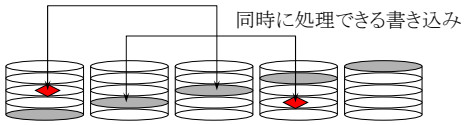
- ビット単位のストライピングでは，サイズの小さいアクセス要求が多数ある場合のスループットに難点がある⇒**ブロック単位分散**



- ストライピングユニットよりも小さい読み出し
  ⇒単一ディスクユニットへのアクセスで処理できる
- 書き込み… **Read-Modify-Write**
  (1) データディスクから変更前のデータを読む
  (2) パリティディスクから変更前のパリティを読む
  新パリティ＝旧パリティ ⊕ 旧データ ⊕ 新データ
  (3) 新しいデータを書き込む
  (4) 新しいパリティを書き込む

RAID Level 5

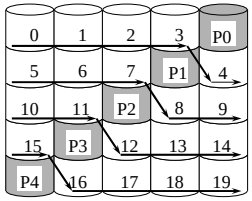
- Level 4 ではパリティディスクがボトルネックになる恐れがある.
  ⇒ **パリティをブロック単位で分散**
  (Block-Interleaved Distributed Parity)
- Level 4 と同様に, Read-Modify-Write が必要



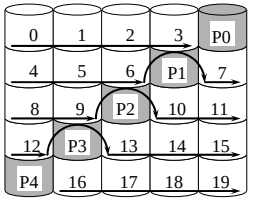
小さな読み出し …◎(冗長情報を含む中では最高)
大きな読み出し …◎(〃)
大きな書き込み …◎(〃)
小さな書き込み …△(Mirrored にやや劣る)

RAID Level 5 の最適パリティブロック配置

- 何でも順番に置けばいいというものではない！
  例えば, 先頭から順次読み出す場合のディスクのアクセス順序が崩れないようにしたほうがよい.



各ディスクを1回ずつ順にアクセスできる



アクセスがスキップされるディスクが出る一方で早々と2度目のアクセスが起こるディスクも出る

まとめ

- RAID の役割
  - **性能の向上** … 転送速度の向上に向けて
    ⇒ビット単位のストライピング
    スループットの向上に向けて
    ⇒ブロック単位のストライピング
  - **信頼性の向上**… 冗長性の付加
    (書き込み性能低下の防止が鍵)
- 市販の RAID
  - いくつかの Level のうち適切なものを選択して使えるようになっているものが多い.
  - 低価格化により, スーパーコンピュータだけでなく, ワークステーション, PC等でも利用可能になっている.