第6章 入出力アーキテクチャ

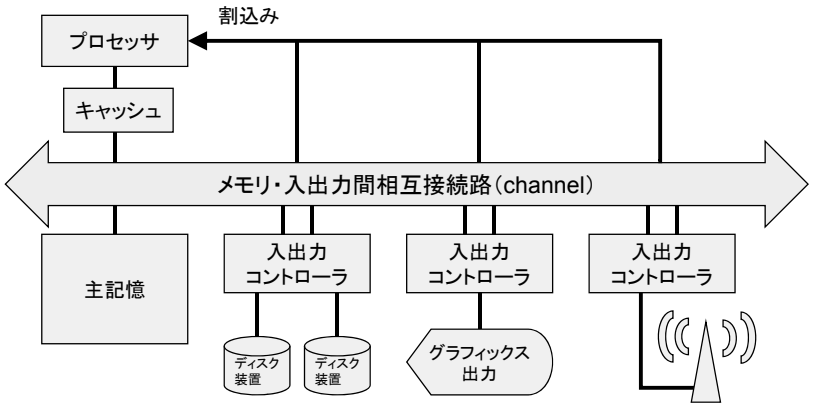
大阪大学 大学院 情報科学研究科
今井 正治

E-mail: arch-2014@vlsilab.ics.es.osaka-u.ac.jp

講義内容

- 入出力装置の概要
- ディペンダビリティ
- ディスク・ストレージ
- フラッシュ・ストレージ（Solid State Disk：SSD）
- プロセッサ，主記憶，入出力装置間の接続
- プロセッサ，主記憶，OSと入出力装置のインタフェース
- 並列処理と入出力： RAID

典型的な入出力装置



様々な入出力装置（1）

装置名	入出力動作	入出力の相手	データ転送速度（MB/秒）
キーボード	入力	人間	0.0001
マウス	入力	人間	0.0038
音声入力	入力	人間	0.2640
サウンド入力	入力	機械	3.0000
スキャナ	入力	人間	3.2000
音声出力	出力	人間	0.2640
サウンド出力	出力	人間	8.0000
レーザ・プリンタ	出力	人間	3.2000

様々な入出力装置(2)

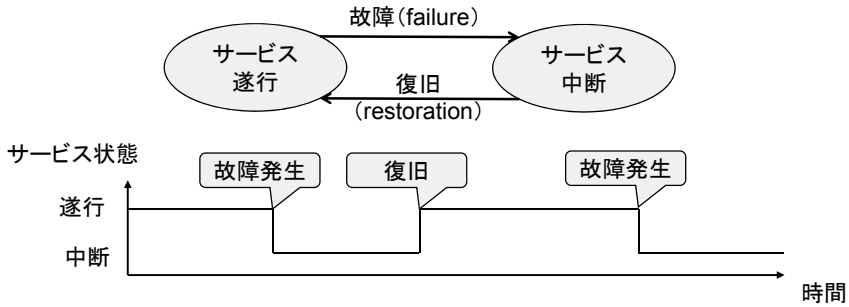
装置名	入出力動作	入出力の相手	データ転送速度(MB/秒)
グラフィック・ディスプレイ	出力	人間	800.0000～8000.0000
ケーブル・モデム	入力/出力	機械	0.1280～6.0000
ネットワーク/LAN	入力/出力	機械	100.0000～10000.0000
ネットワーク/無線LAN	入力/出力	機械	11.0000～54.0000
光ディスク	記憶	機械	80.0000～220.0000
磁気テープ	記憶	機械	5.0000～120.0000
フラッシュ・メモリ	記憶	機械	32.0000～200.0000
磁気ディスク	記憶	機械	800.0000～3000.0000

講義内容

- 入出力装置の概要
- ディペンダビリティ
- ディスク・ストレージ
- フラッシュ・ストレージ(Solid State Disk:SSD)
- プロセッサ，主記憶，入出力装置間の接続
- プロセッサ，主記憶，OSと入出力装置のインタフェース
- 並列処理と入出力: RAID

ディペンダビリティ

- サービス仕様と状態
  - サービス遂行(service accomplishment)
  - サービス中断(service interruption)



信頼度(reliability)の尺度(metrics)

- 平均故障寿命 MTTF(mean time to failure)
  - 基準点からのサービスの連続遂行可能時間の平均(故障に至るまでの時間)
- 年間故障率 AFR(annual failure rate)
  - 1年間に予想される故障率
- 平均修復時間 MTTR(mean time to repair)
  - 故障が起きてから修復されるまでの平均時間
- 平均故障間隔 MTBF(mean time before failure)
  - MTTF と MTTR の和

アベイラビリティ（availability）

- サービス遂行とサービス中断の2つの状態の入れ替わりを考慮に入れたサービス遂行の尺度

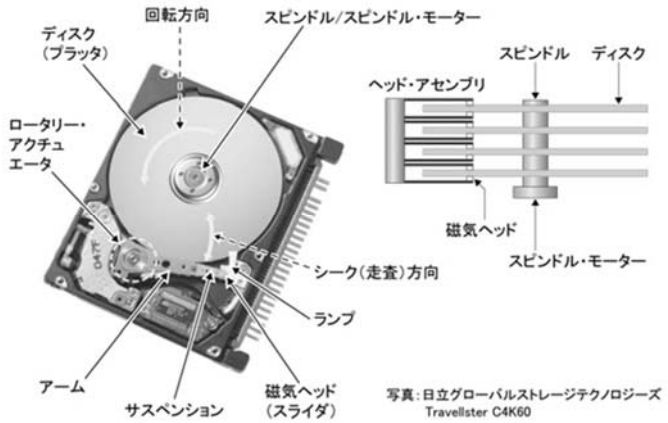
$$\text{アベイラビリティ} = \frac{\text{MTTF}}{\text{MTTF} + \text{MTTR}}$$

- MTTFの改善方法
  - 故障回避（fault avoidance）
    - 故障の発生原因を元から絶つ
  - 故障許容（fault tolerance）
    - 冗長性を持たせる
  - 故障予測（fault forcasting）
    - 故障の発生を予測して発生する前に部品を交換する

講義内容

- 入出力装置の概要
- ディペンダビリティ
- ディスク・ストレージ
- フラッシュ・ストレージ（Solid State Disk：SSD）
- プロセッサ，主記憶，入出力装置間の接続
- プロセッサ，主記憶，OSと入出力装置のインタフェース
- 並列処理と入出力： RAID

ハードディスク装置の内部構造



磁気ディスク装置（1）

- 磁気ディスク装置
  - 1～4枚のディスクから構成される
  - 回転数: 5,400～15,000 rpm（revolution per minute）
- ディスク（disk）
  - 直径 1～3.5インチ（inch）
  - 10,000～50,000本のトラック
- トラック（track）
  - 100 ～500のセクタ

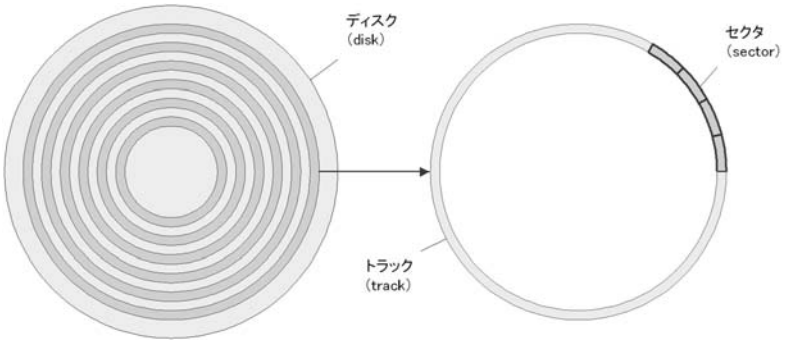
磁気ディスク装置(2)

- セクタ(sector)
  - 代表的なセクタサイズは512バイト
  - 4,096バイトに拡大しようとする動き
- セクタの構造
  - セクタ番号
  - ギャップ
  - データ
  - 誤り訂正コード(ECC: Error Correction Code)
  - ギャップ

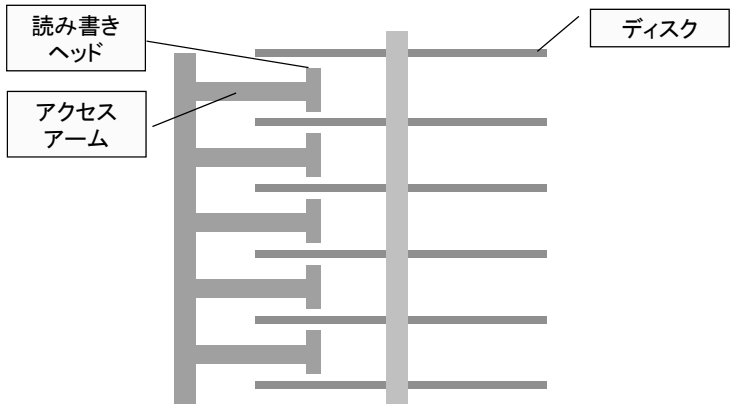
磁気ディスク装置(3)

- 読み書きヘッド(read/write head)
  - データを読み書きするためのコイル(coil)
- アクセスアーム(access arm)
  - 読み書きヘッドを移動させるための金属の板
- シリンダ(cylinder)
  - ディスク上で同じ位置を占めるトラックの集合

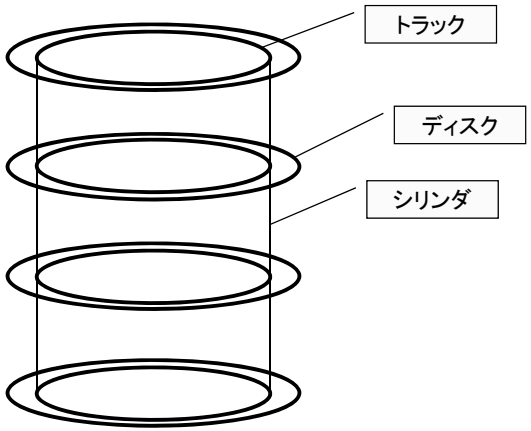
ディスク, トラック, セクタ



磁気ディスク装置の構造



シリンダ



アクセス時間(1)

- アクセス時間＝
  シーク時間＋回転待ち時間＋データ転送時間

- シーク時間(seek time)
  - アクセスアームを動かして、読み書きヘッドを適切な位置に移動させる(seek)のに必要な時間
  - 平均シーク時間は、3～13 ms

アクセス時間(2)

- 回転待ち時間(rotation latency)
  回転遅延時間(rotation delay)
  - 回転数が 10,000 rpm の場合

$$\text{平均回転待ち時間} = \frac{0.5\text{回転}}{10000\ \text{rpm} / 60} = 0.0030\ \text{秒} = 3.0\ \text{ms}$$

- データ転送時間(data transfer time)
  - データ転送速度(data transfer rate)
    - セクタ ⇔ キャッシュ: 70～125 MB/秒
    - キャッシュ ⇔ CPU: ～375 MB/秒
- ディスクコントローラを内蔵

講義内容

- 入出力装置の概要
- ディペンダビリティ
- ディスク・ストレージ
- フラッシュ・ストレージ(Solid State Disk：SSD)
- プロセッサ，主記憶，入出力装置間の接続
- プロセッサ，主記憶，OSと入出力装置のインタフェース
- 並列処理と入出力：RAID

フラッシュ・ストレージの例

特性	Kingston SecureDigital SD4/8GB	Transend Type 1 CompactFlash TS16GCF133	RiDATA Solid State Disk 2.5インチSATA
フォーマットされたデータ容量（GB）	8	16	32
セクタ当たりのバイト数	512	512	512
データ転送速度（読み/書き，Mbit/s）	4	20 / 18	68 / 50
消費電力：稼働中/待機中（W）	0.66 / 0.15	0.66 / 0.15	2.1 / -
寸法：HxWxD（inch）	0.94 x 1.26 x 0.08	1.43 x 1.68 x 0.13	0.35 x 2.75 x 4.00
重量（グラム）	2.5	11.4	52
平均故障間隔（時間）	> 1,000,000	> 1,000,000	> 4,000,000
GB/立法inch, GB/W	84, 12	51, 24	8, 16

NOR型フラッシュメモリとNAND型フラッシュメモリの比較

特性	NOR型	NAND型
代表的な用途	BIOSメモリ	USBメモリ
最小アクセスサイズ（バイト）	512	2048
読み出し時間（μs）	0.08	25
書き込み時間（μs）	10.00	消去に 1500+250
読み出しバンド幅（MB/s）	10	40
書き込みバンド幅（MB/s）	0.4	8
耐用限度（セル当たりの書き込み回数）	100,000	10,000～100,000
GB当たりの最安価格（2008年）	US$ 65	US$ 4

ウェア・レベリング（wear leveling）

- NAND型フラッシュメモリでの耐久性強化法
- コントローラを内蔵し，頻繁に書き込まれたブロックを稀にしか書き込まれていないブロックと配置し直す
- これにより，メモリの寿命が延びる
- 歩留り（yield）も向上

ハイブリッド（hybrid）型ハードディスク

- 1GB程度のフラッシュ・メモリとハードディスクの組合わせ
- OSをフラッシュ・メモリ上に配置することにより起動時間を短縮
- ハードディスクをより頻繁にアイドル状態にすることにより，消費エネルギーを節約

講義内容

- 入出力装置の概要
- ディペンダビリティ
- ディスク・ストレージ
- フラッシュ・ストレージ（Solid State Disk：SSD）
- プロセッサ，主記憶，入出力装置間の接続
- プロセッサ，主記憶，OSと入出力装置のインタフェース
- 並列処理と入出力：RAID

バス（bus）

- プロセッサ・主記憶間バス（processor-memory bus）
  - 短距離，高速，バンド幅を最大化
- 入出力バス（I/O bus）
  - 長距離，多種類の装置が接続される可能性がある
  - 接続される装置のバンド幅が広範囲にわたる
- バックプレーン・バス（backplane bus）
  - 単一のバスに，プロセッサ，主記憶，入出力装置を全て接続できるようしたバス

バスの規格

- 入出力トランザクション（I/O Transaction）
  - アドレスの送出 ⇒ データの送信/受信
  - リード・トランザクション（read transaction）
  - ライト・トランザクション（write transaction）
- 標準のバス規格
  - Firewire（IEEE Std 1394）
  - USB（Universal Serial Bus）
  - PCI Express（PCIe）
  - SATA（Serial ATA）
  - SAS（Serial Attached SCSI）

同期式（synchronous）バス

- 特徴
  - 制御線の中にクロック用の線が存在
  - クロックを基準とした固定の通信プロトコルを使用
- 利点
  - 小さいステートマシンを用いて簡単に実現できる
  - 高速，インタフェースの論理も小規模
- 欠点
  - 同じバスに接続される装置は同じクロック周波数で動作させる必要がある
  - 高速な同期式バスでは，バス長を長くできない（クロック・スキューの発生）

非同期式(asynchronous)バス

- 多種多様な装置に対応可能
  - クロック・スキュー，同期を考慮する必要がない
  - バス長を長くできる
- ハンドシェーク型プロトコル(handshaking protocol)を使用

主要な入出力規格の主要特性

特性	Firewire	USB 3.0	PCI Express	SATA	SAS
用途	外部	外部	内部	内部	外部
チャネル当たりの接続数	63	127	1	1	4
基本データ幅	4	2	レーン当たり2	4	4
理論上のピークバンド幅	50MB/s 100MB/s	1.5, 12, 480, 5000 Mbps	レーン当たり 250MB/s	300 MB/s	300MB/s
活線挿抜	可	可	形状仕様しだい	可	可
バスの最大長	4.5メートル	3メートル	0.5メートル	1メートル	8メートル
規格の名称	IEEE 1394	USB Implementors Forum	PCI-SIG	SATA-IO	T10 Committee

講義内容

- 入出力装置の概要
- ディペンダビリティ
- ディスク・ストレージ
- フラッシュ・ストレージ(Solid State Disk：SSD)
- プロセッサ，主記憶，入出力装置間の接続
- プロセッサ，主記憶，OSと入出力装置のインタフェース
- 並列処理と入出力：RAID

OS(Operating System)の機能

- ユーザプログラムによる入出力装置へのアクセスの管理(データの保護)
- 入出力装置の低水準の操作を処理するルーチンを提供し，入出力装置に対するアクセス操作を抽象化する
- プログラムおよび入出力装置が生成した例外を処理する
- 入出力装置へのアクセスをスケジューリングすることにより，システムのスループットを向上するとともに共有の入出力資源に公平にアクセスできるようにする

OS，主記憶，入出力装置との間の通信

- OSから入出力装置へのコマンドの送出
  - 読み出し，書き込み，ディスクのシークなど
- 入出力装置からOSへの状態の通知
  - 操作の完了，エラーの発生
- 入出力装置と主記憶との間のデータの転送
  - 入出力装置へのデータの書き込み
  - 入出力装置からのデータの読み出し

入出力装置に対するコマンドの送出

- 入出力装置のアドレス指定の方法
  - メモリ・マップ入出力（memory map I/O）
  - 入出力用特殊命令
- メモリ・マップ入出力
  - ユーザプログラムは，入出力装置にコマンドを直接発行出来ない
  - アドレス空間の一部を入出力装置に割当てる
  - 入出力コントローラ（I/O controller）は，これらのアドレス空間への読み出しおよび書き込みを入出力コマンドとして処理する

コマンドの実行

- コマンドを実行する際には，入出力装置の状態を調べてコマンドが正常に完了したかどうかを判定する必要がある
- 入出力装置レジスタ（I/O device register）
  - 状態情報
  - 入出力データ
- 状態情報
  - 完了ビット（done bit）
  - エラー・ビット（error bit）

入出力命令（I/O instruction）

- プロセッサは，入出力バス（I/O bus）を用いて入出力デバイスと通信を行う
- 入出力命令中の情報
  - 入出力装置の番号
  - コマンド語または，主記憶中のコマンド語のロケーション（アドレス）
- ユーザプログラムによる入出力装置へのアクセスの管理
  - 入出力命令は，スーパバイザ・モード（supervisor mode）でのみ実行可能
  - ユーザプログラムが入出力命令を実行しようとすると不正操作として例外が発生する

プロセッサとの通信

- 入出力装置の状態の検査方法
  - ポーリング（polling）
  - 割込み（interrupt）
- ポーリング
  - 入出力装置は，状態レジスタに情報をセット
  - プロセッサは，制御と作業を行う
  - 利点：実装が容易（入出力装置の動作速度が分かっている場合などには好都合）
  - 欠点：オーバヘッドが大きい

割込み駆動型入出力（interrupt driven I/O）

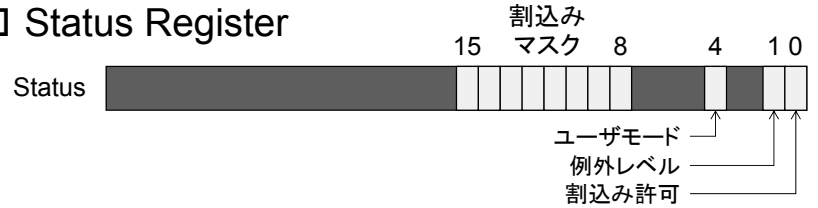
- 入出力割込みの発生
  - 入出力装置が操作を完了した場合およびプロセッサの関与が必要な場合（エラーの発生など）に発生
- 入出力割込みの特徴
  - 入出力割込みは，命令の実行とは非同期（実行中の命令とは無関係に発生する）
  - 入出力命令が発生した事実に加えて，それがどの装置で発生したかを示す情報が必要（ベクタ・アドレスまたはCause registerを使用）
  - 複数の入出力装置から割込みが発生した場合には，優先順位に従って処理を行う

割込みの優先レベル

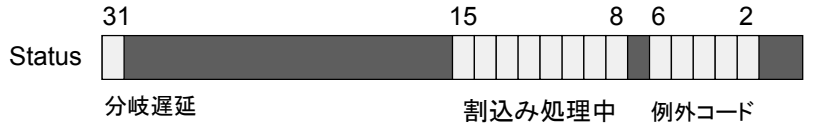
- 入出力装置の優先順位への対応
  - UNIXの場合には，4～6の優先レベル
  - 内部で発生した例外の優先度は入出力割込みの優先度より低いのが普通
  - 入出力割込みでは，高速な装置ほど優先度が高い

MIPSのStatusレジスタとCauseレジスタ

- Status Register



- Cause Register



割込み処理の手順(1)

1. 割込み処理中フィールドと割込みマスク・フィールドの論理積を取って，どの割込みが処理の許可対象になりうるかを判定する．StatusレジスタとCauseレジスタのコピーを作る．
2. 上記の割り込みの中から，優先順位が最も高いものを選択する．（左端に近いほど優先度が高い．）
3. Statusレジスタの割込みマスク・フィールドを退避する．
4. 優先度が同等または低いすべての割込みを不許可とするよう，割込みマスク・フィールドを変更する．

割込み処理の手順(2)

5. 割込み処理に必要なプロセッサの状態を退避する．
6. 選択した割込みを処理可能にするため，Statusレジスタの割込み許可ビットを1にセットする．
7. 適切な割込みルーチンを呼出す．
8. 状態を復元する前に，Statusレジスタの割込み許可ビットを0にセットする．それにより，割込みマスク・フィールドの復元が可能になる．

- IPL: Interrupt Priority Level（プロセスの優先度）

入出力装置と主記憶間のデータ転送

- ハードディスク装置などのバンド幅の大きな入出力装置では，かなり大きなブロック（数100～数1000バイト）単位でデータの転送が行われる．
- プロセッサの負荷を軽減する方法を採用する必要がある．
- ダイレクト・メモリアクセス（DMA: Direct Memory Access）
- DMAコントローラは，入出力装置と主記憶との間で，マスター（master）となって，プロセッサとは独立に直接データ転送を行う．

DMAによるデータ転送の手順

1. プロセッサがDMAコントローラの設定を行う
   - 使用する入出力装置
   - 転送データのソースまたはデスティネーションの主記憶アドレス
   - 転送バイト数
2. DMAコントローラは，当該入出力装置に対する操作を開始し，相互接続路の調停を行う．データ転送は複数回行われる可能性がある．
3. データ転送が完了すると，DMAコントローラはプロセッサに割込みをかける．
4. プロセッサはDMAコントローラはたは主記憶の状況を調べ，全操作が正常に完了したかどうかを調べる．

講義内容

- 入出力装置の概要
- ディペンダビリティ
- ディスク・ストレージ
- フラッシュ・ストレージ（Solid State Disk：SSD）
- プロセッサ，主記憶，入出力装置間の接続
- プロセッサ，主記憶，OSと入出力装置のインタフェース
- 並列処理と入出力：RAID

システムの性能に対する入出力の影響（1）

- 例題
  - 経過時間100秒のベンチマークがあり，うち90秒はCPU時間，残りの10秒は入出力時間であるとする
  - 2年ごとにプロセッサ数が倍増するが，プロセッサの処理能力は同じままであり，入出力時間は改善されないとする
  - 6年後には，このプログラムの実行はどの程度速まるか？

システムの性能に対する入出力の影響（2）

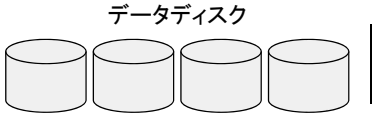
経過年数	CPU時間	入出力時間	経過時間	入出力時間の割合	経過時間の改善比
0	90秒	10秒	100秒	10%	1.0
2	90/2=45秒	10秒	55秒	18%	1.8
4	45秒/2 =22.5秒	10秒	32.5秒	31%	3.1
6	22.5秒/2 =11.25秒	10秒	21.25秒	47%	4.7

ハードディスク装置の高性能化

- RAID: redundant arrays of inexpensive disks（冗長性を持たせた安価なディスクのアレイ）
- レベル
  - RAID 0: 冗長性なし
  - RAID 1: ミラーリング
  - RAID 2: 誤り訂正コード
  - RAID 3: ビット・インタリーブ方式のパリティ
  - RAID 4: ブロック・インタリーブ方式のパリティ
  - RAID 5: 分散ブロック・インタリーブ方式のパリティ
  - RAID 6: P+Q冗長性（パリティ＋CRC）

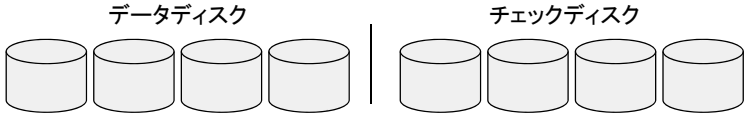
RAID 0（冗長性なし）

- 単純に複数のディスク上にデータを分散（ストライピング：striping）
- ソフトウェアからは，小さなディスクの集まりが単一の大きなディスクに見えるので，記憶装置の管理が単純化される
- 複数のディスクに同時に読み書きが可能になるので大容量データのアクセス性能が向上する



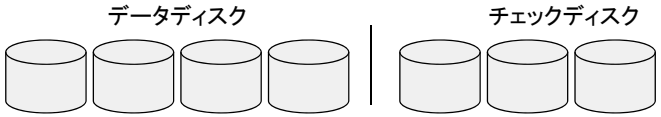
RAID 1（ミラーリング）

- ミラーリング（mirroring）または シャドーイング（shadowing）
- 故障許容性（tolerance）を持たせるため，RAID 0 の2倍の数のディスクを使用する
- データを書き込み時には，同じデータがチェックディスクにも書きこまれる
- 一方のディスクが故障したら，システムは直ちにミラーの方に切り替えて情報を取り出す

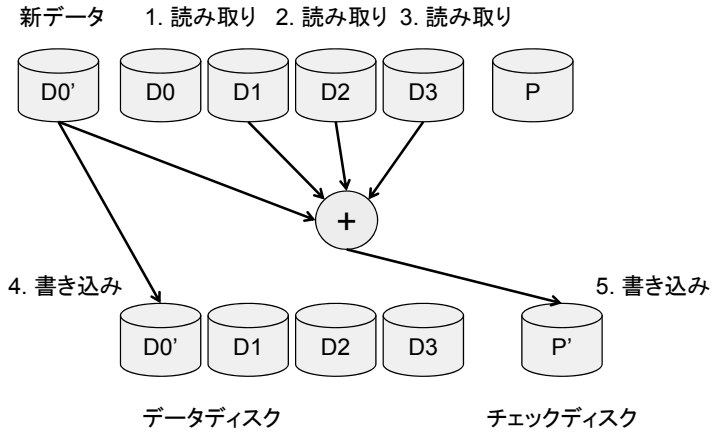


RAID 2（誤り検出コードと訂正コード）

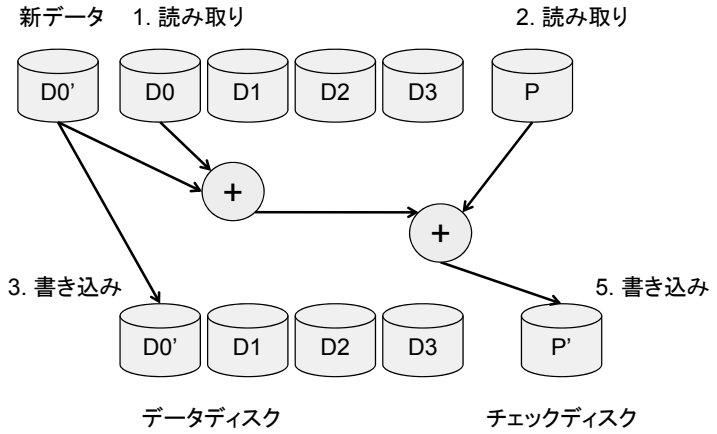
- メモリで用いられるビット誤りの検出と訂正の方法を用いる
- ディスクの使用効率が悪いので実際には使用されていない



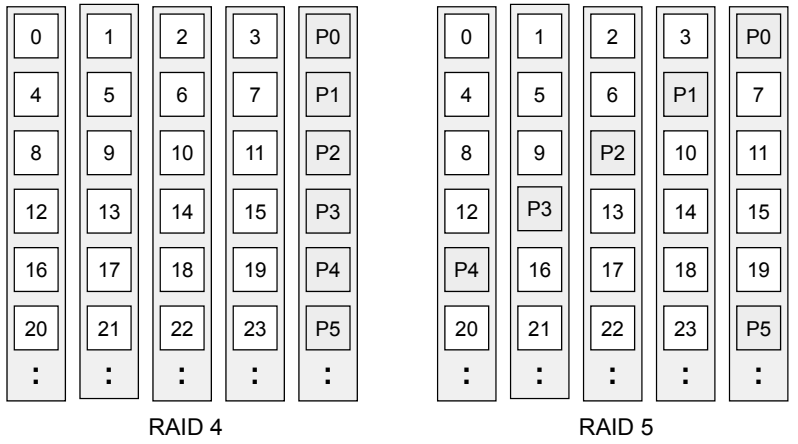
RAID 3（ビット・インタリーブ方式のパリティ）



RAID 4
（ブロック・インタリーブ方式のパリティ）



RAID 5（分散ブロック・インタリーブ方式のパリティ）



RAID 6（P+Q冗長性）

- 単一の障害からの回復だけでは不十分な場合に備えて2種類の冗長コードを用いる
  - P: パリティ（parity）
  - Q: CRC（cyclic redundancy code）
- 2種類の冗長コードを格納するために2台のディスクにアクセスを行う