データ工学特論（2）

情報基盤センター
天野　浩文

この講義に関するwebサイト:
http://isabelle.cc.kyushu-u.ac.jp/~amano/data-engineering/

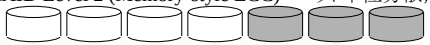
RAIDの分類（おさらい）

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

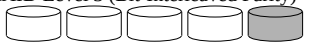
網掛け部分は冗長情報の部分を表す

RAID Level 1 (Mirrored) ブロック単位分散，二重化

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

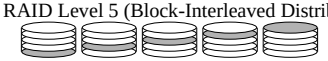


RAID Level 3 (Bit-Interleaved Parity) ビット単位分散，パリティ



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



RAIDの分類についての補足（おさらい）

- “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．

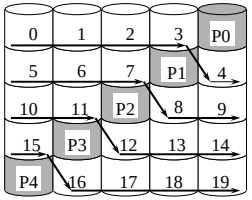
カタログ等でよく用いられる表記

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-5のLarge Read（おさらい）

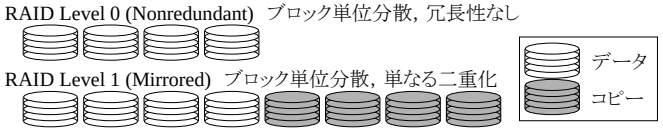
- RAID-5の最適ブロック配置
  - 先頭から順次読み出す場合のディスクのアクセス順序が崩れないようにすると，すべてのディスクが読み出しに貢献できる
  - これはRAID-0と同じ



各ディスクを1回ずつ順にアクセスできる

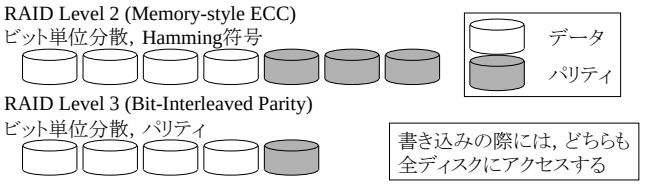
RAIDの比較（1）

- RAID-0 と RAID-1
  - コスト： RAID-0＜RAID-1（冗長ディスクの数が多い）
  - データディスク数が同一の場合のスループット性能
    書き込み： RAID-0＞RAID-1
    読み出し： RAID-0＜RAID-1



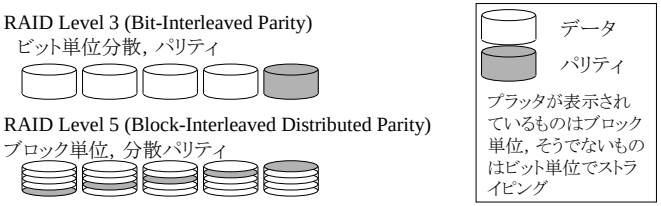
RAIDの比較（2）

- RAID-2 と RAID-3
  - コスト： RAID-2＞RAID-3（冗長ディスク数の差）
  - データディスク数が同一の場合のスループット性能
    読み出し： RAID-2＝RAID-3
    書き込み： RAID-2＜RAID-3



RAIDの比較（3）

- RAID-3 と RAID-5
  - コスト： RAID-3＝RAID-5
  - スループット性能
    小さな読み出し： RAID-3＜RAID-5
    大きな読み出し： RAID-3＜RAID-5
    小さな書き込み： RAID-3≦RAID-5
    大きな書き込み： RAID-3＝RAID-5



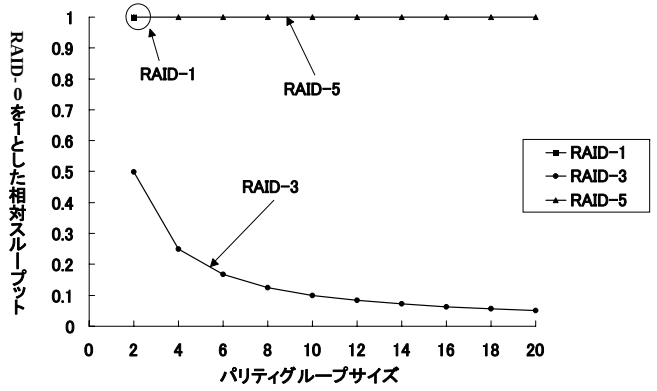
RAID-0, 1, 3, 5 のスループット性能の総合比較

- 同一コスト（＝全ディスク数）で達成できるスループット性能の RAID-0 に対する相対値

	Small Read	Small Write	Large Read	Large Write	Storage Efficiency
RAID-0	1	1	1	1	1
RAID-1	1	1/2	1/2	1/2	1/2
RAID-3	1/(G－1)	1/G	(G-1)/G	(G-1)/G	(G-1)/G
RAID-5	1	max(1/G,1/4)	1	(G-1)/G	(G-1)/G

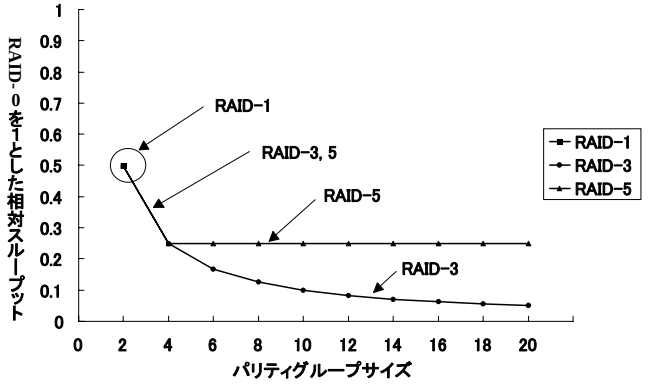
Small: 1ブロック内に収まるアクセス
Large: 1ストライプ全体に渡るアクセス
G: パリティグループのサイズ

小さな読み出しのスループット



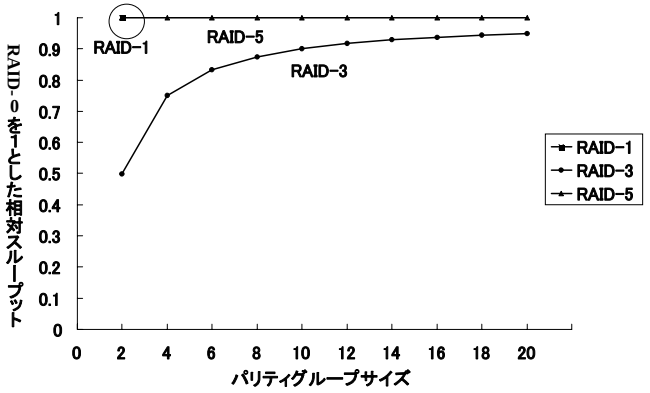
データ工学特論(2) 9

小さな書き込みのスループット



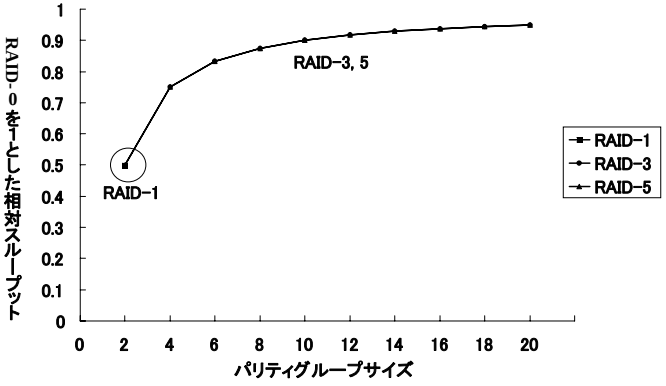
データ工学特論(2) 10

大きな読み出しのスループット



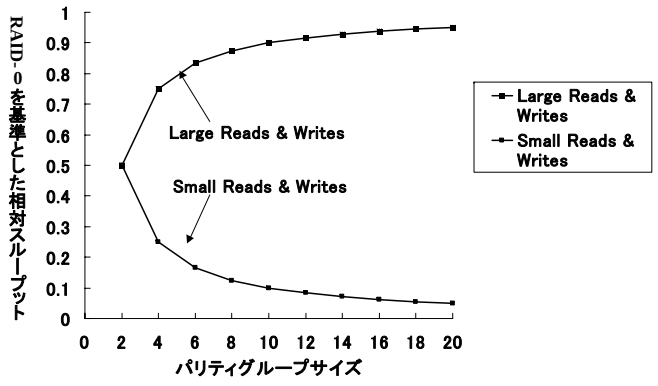
データ工学特論(2) 11

大きな書き込みのスループット

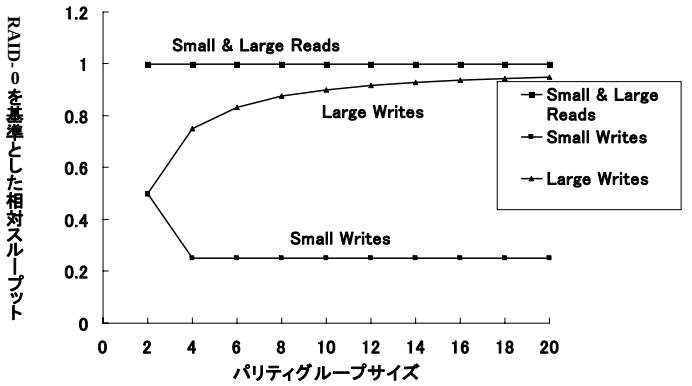


データ工学特論(2) 12

RAID-3 の相対スループット



RAID-5 の相対スループット



RAIDの信頼性を左右する要因(1)

- MTBF(mean time between failure)とMTTF(mean time to failure)
  - どちらも，故障・障害が発生するまでの平均時間のこと.
  - MTBFは，障害発生後，修理して使用を再開するような装置・部品に対して用いることが多い.
  - MTTFは，障害発生後は修理せず交換してしまうような装置・部品に対して用いることが多い. その意味で，「平均寿命」と考えることもできよう.
- MTBF(MTTF)と障害発生確率λの関係

$$MTBF\,(MTTF) = \frac{1}{\lambda}$$

RAIDの信頼性を左右する要因(2)

- 1台のディスクにランダムに発生する障害だけを考える場合
  - RAID-5 の MTTF (mean time to failure)

$$\left(\frac{N}{MTTF\,(disk)} \times \frac{MTTR\,(disk) \times (G-1)}{MTTF\,(disk)}\right)^{-1}$$

$$= \frac{MTTF^{\,2}(disk)}{N \times (G-1) \times MTTR\,(disk)}$$

(MTTR: mean time to repair)

  - N=100, G=16, MTTF(disk)=200,000時間，MTTR(disk)＝1時間とすると，
    RAID-5 の MTTF ＝約3,000年
- だが，実際には...

RAIDの信頼性を左右する要因（3）

- **システムクラッシュ**
  停電，操作ミス，ハードウェア障害，ソフトウェア障害
  - I/O操作が途中で止まる
    パリティ等の不整合が生じる
    ⇒RAID の異常ではないが，データには異常が残る
- システムクラッシュ時のパリティ不整合を防ぐ方法
  - 書き込みのたびに，パリティ回復に必要な情報をログに残す．
  - ただし，不揮発性 RAM などのハードウェアサポートがなければ大幅な性能低下は必至

RAIDの信頼性を左右する要因（4）

- **訂正不能なビットエラー** (uncorrectable bit error)
  …10^{14}に 1 ビット程度の割合
  - ディスクの磁性面の摩滅などによるデータ損失
  - read の際にパリティ異常で露見する
- 障害ディスクのデータの回復
  ⇒それ以外の全ディスクの読み出しが必要！
  再度，訂正不能なビットエラーが起こると，データは回復不能になる
  ⇒100GBの RAID では，0.8%程度のデータは回復不能
- 訂正不能なビットエラーを防ぐには...
  読み出しが危なくなってきたら，ハードウェア的に警告を出させる．読めなくなる前にオペレータが対処．

RAIDの信頼性を左右する要因（5）

- RAID-1 ～ RAID-5 …単一ディスクの障害しか考えていない
- 複数のディスクが相次いで故障することもある
  - 地震，電源異常，落雷，etc.
  - 同一の製造ラインから出荷されたディスクが同時期に故障する...
- 単一ディスクの障害が発見されてから，再構築が完了する前に次のトラブルが起こったら...
  ⇒そのデータは失われる

MTTDL (Mean Time to Data Loss)（1）

- RAIDにとり最も深刻な事態
  - データが失われること
- データが失われるケース
  - 2台のディスクに同時に障害が発生する
  - システムクラッシュ後にディスク障害が起こる
  - ディスク障害発生後，再構築中に訂正不能なビットエラーが出る
- 典型的なパラメータの例

ユーザディスク数	100台（500GB）
ディスクサイズ	5 GB
セクタサイズ	512 バイト
ビットエラー率（ビット）	$1/10^{14}$
ビットエラー率（セクタ）	$1/(2.4\times10^{10})$
全セクタが読める確率（p(disk)）	99.96%
パリティグループサイズ	16 ディスク

MTTF(disk)	200,000時間
MTTF(disk2)	20,000時間
MTTR(disk)	1時間
MTTF(sys)	1ヶ月
MTTR(sys)	1時間

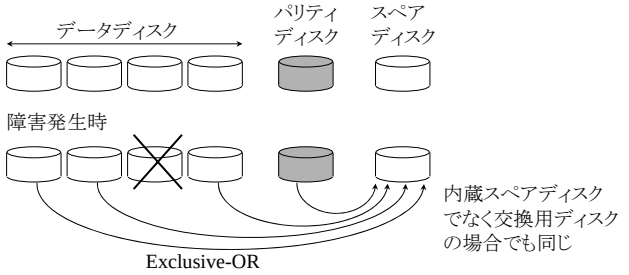
MTTDL（2）

- RAID-5 （N＝ディスク総数）

	MTTDL	10年間のデータ損失確率
ディスクの二重障害	$\frac{MTTF(disk)\times MTTF(disk2)}{N\times(G-1)\times MTTR(disk)}=285$年	3.4%
システムクラッシュ＋ディスク障害	$\frac{MTTF(sys)\times MTTF(disk)}{N\times MTTR(sys)}=154$年	6.3%
ディスク障害＋ビットエラー	$\frac{MTTF(disk)}{N\times(1-(p(disk))^{G-1})}=36$年	24.4%
ソフトウェアRAID	（上記の調和平均＝26年）	31.6%
ハードウェアRAID (NVRAM付き)	（「システムクラッシュ＋ディスク障害」を除いた調和平均＝32年）	26.8%

スペアディスク

- ディスクアレイを構成するディスクが故障した場合に，それの代替となるディスクが内蔵されていることも多い
  - 例：RAID Level 3 での構成の一例



ディスクアレイの状態に関する情報

- ディスクアレイが保持する情報
  - データ
  - パリティ(あるいはその他の冗長情報)
  - **状態記述情報**(metastate information)
    例：どのディスクが故障しているのか
    故障したディスクのうちどのセクタまで修復済みであるか
    どのセクタが修復中であるか
- 状態記述情報の保持のしかた
  - システムクラッシュがあっても正しく保持されている必要がある
  - セクタあるいはストライピングユニット単位にvalid/invalidを記録する

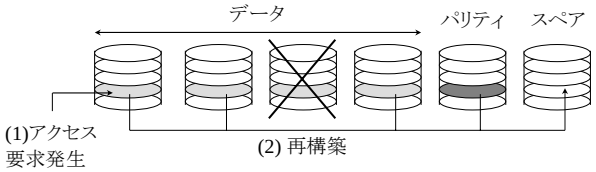
システムクラッシュ後のパリティ再構築

- 書き込み中にシステムクラッシュ
  ↓
  そのとき書き込もうとしたパリティは不正になる可能性がある．
  しかし，どのパリティセクタが不正なのか，ディスクアレイ全体をスキャンする以外に確実に判定する方法がない
- パリティセクタごとに valid/invalid を記録する．
  - 書き込み処理前にそのセクタをinvalid としておく．
  - ときどき，全セクタをvalid とするか，あるいは，パリティ書き込み処理が完了するたびにそのセクタをvalid と記録する．
  - システムクラッシュから回復したら，invalid と記録されているパリティセクタをすべて再構築する．

故障ディスクが出た場合の対処（1）

- スペアディスクがある場合
  ⇒**デマンド再構築**（demand reconstruction）
  - 故障ディスクを含むストライプへのアクセスが発生すると，故障ディスクのデータを再構築してスペアディスクに書き込む．（残りの部分はバックグラウンドプロセスで再構築）

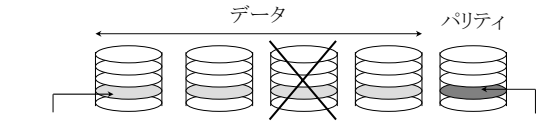
RAID Level 4 の場合



故障ディスクが出た場合の対処（2）

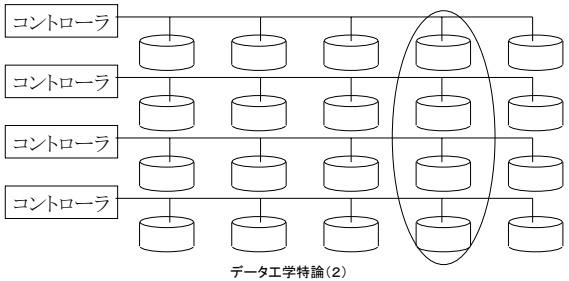
- スペアディスクがない場合
  ⇒**パリティスペアリング**（parity sparing）
  - 故障ディスクを含むストライプへのアクセスが発生すると，対応するパリティをつぶして書き込み，relocated と記録する．
  - 故障ディスクが交換されたら relocated マークのついたセクタを書き戻し，パリティを再構築し，relocated マークを消す．

RAID Level 4 の場合



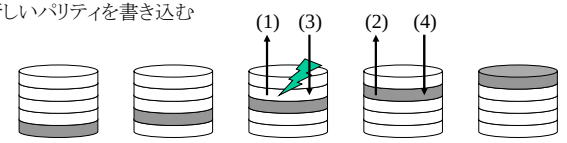
Orthogonal RAID

- 数台のディスクが1本のバスにつながる場合，1つのコントローラが故障すると，それにつながるディスクは全滅する．
- Orthogonal RAID
  - 共有されるハードウェアと「直交」する向きにパリティグループを構成する．



RAID Level 5 の書き込み性能の改良（1）

- RAID Level 5 における書き込み…4回のディスクアクセス
  - 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



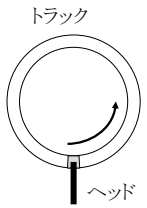
- 改良策
  - バッファリングとキャッシング
  - フローティングパリティ
  - パリティロギング

RAID Level 5 の書き込み性能の改良（2）

- バッファリング(buffering)…非同期書き込み
  - 書き込み処理が完了する前にホスト側にacknowledgeする
  - バッファが不揮発性メモリでできていなければ，システムクラッシュ時にデータが失われる
  - 低負荷時のレスポンスタイムは向上する
  - 同一ストライプへの書き込みをまとめて，フルストライプアクセスにすることができれば，スループットも多少は向上
- キャッシング(caching)
  - 通常のデータ処理では，対象データを読み込んでから新しい値を計算し，書き込むのが普通
    ⇒最初の読み込みのときにRead-Modify-Write用のデータをキャッシュしておく

RAID Level 5 の書き込み性能の改良（3）

- フローティングパリティ (floating parity)
  Menon et al. 1993
  - Read-modify-write の読み出しが終了してから書き込みが可能になるまでディスクが1回転するのを待たなければならない
  - 元のブロックではなく，もっとも近い空きブロックに書き込むようにする
  - 空きブロックが必要⇒領域の効率の問題
    空きブロックの位置の情報の管理も必要
    ディスクコントローラの中にハードウェア的に実装する



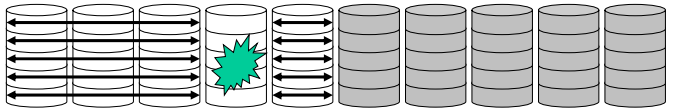
- フローティングデータ(floating data)
  - パリティだけでなくデータもフロートさせる
  - データが不連続になりやすく，領域はさらに非効率的

RAID Level 5 の書き込み性能の改良（4）

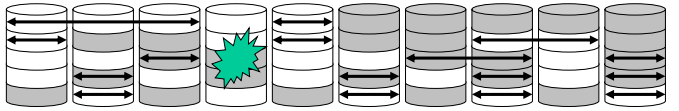
- パリティロギング (parity logging)
  Stodolsky and Gibson 1993
- 旧パリティの読み出しと新パリティの書き込みを遅らせる
  - 新旧パリティの差分（パリティ更新イメージ）のログを不揮発性メモリに記録
  - 不揮発性メモリが一杯になったら，ディスク上のファイルに転記
  - ログが一杯になったら，旧パリティとパリティ更新イメージをメモリに読み込み，まとめてパリティを更新する
  - パリティブロックに対する小さな書き込みをまとめてやや大きな書き込みにできるので，効率が向上する

パリティデクラスタリング(Parity Declustering)

- ディスクアレイの再構築中でも通常の処理を続けたい
  あるいは，サービス再開までの時間を短縮したい
- RAID Level 5 の再構築時の負荷の集中の問題
  複数台のRAID を用いる場合，再構築の行われる範囲が集中

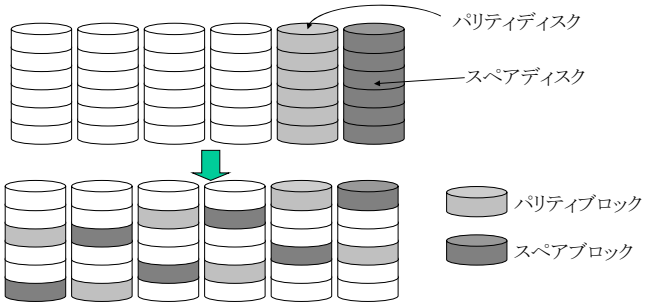


- パリティデクラスタリング Gibson et al. 1990



分散スペアリング (Distributed Sparing)

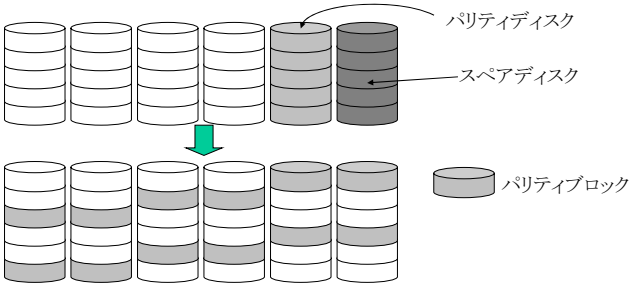
- 正常運転時には遊んでいるスペアディスクを活用する
  Menon et al. 1991



- 正常運転時の性能が上がる
- ディスクに故障がおきたときには，スペアブロックに再構築

パリティスペアリング (Parity Sparing)

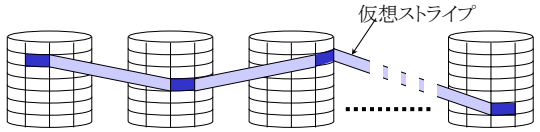
- スペアディスクもパリティに使用する
  Chandy and Reddy 1993



- 半分のサイズのディスクアレイ2つとして扱ったり，2つのパリティブロックのうちシーク時間の短いほうを動的に選んで使う

仮想ストライピング (Kitsuregwa et al. 1993)(1)

- RAID-5 の小さな書き込みが遅くなる原因
  Read-Modify-Writeのために4回のディスクアクセスが必要
- 論理データブロックを物理データブロックに固定しない
  どのブロックがどこにあるかは別にテーブルで管理



仮想ストライプ番号	Disk 0	Disk 1	…	Disk n-1	汚染ブロック数
0	1, 1	3, 6	…	1, 24	0
1	25, 3	1, 48	…	50, 19	2
2	1, 37	2, 11	…	35, 4	1
:	:	:	:	:	:
M	*, *	*, *	*, *	*, *	0

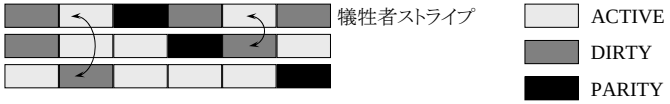
シリンダ　ブロック　空きストライプ

仮想ストライピング(2)

- *n* ドライブに対する *n* 個の小さな書き込みをまとめて，新しいストライプを作る
  ⇒フルストライプ書き込みになる(速い)
  - *i* 番目のディスクのブロックは一番近い空きブロックに書く.
  - 場所が変わったブロックは「汚染ブロック」とする.
  - 後でガーベジコレクションが必要.
- *n* データブロックの書き込み
  - RAID-5: *n* ×4回アクセス
  - 仮想ストライピング: *n* ＋1回アクセス

仮想ストライピング（3）

- 物理状態テーブル…各ブロックの状態を登録
  - **FREE**：使われていない
  - **ACTIVE**：正しいデータが入っている
  - **DIRTY**：更新されたので古い値になっている
- ガーベジコレクション
  - ACTIVEブロック数最小のストライプ（犠牲者ストライプ）のACTIVEブロックをよそのストライプのDIRTYブロックと交換，パリティを更新
  - 犠牲者ストライプのブロックをすべて FREE に



まとめ

- RAIDの比較
  - RAID-5は，小さな書き込みにやや難点があるものの，他のレベルのRAIDに比べて，バランスの取れた理論性能を持つ．
- RAIDの信頼性を左右する要因
  - MTBFとMTTF
    - 単純なMTTFの計算では，RAIDは非常に長い寿命を持つはずだが...
  - MTTDL
- RAID-5の性能を向上させるためのさまざまなアイディア