Engineer

RAID

☰ 複数のディスクを一つとして使い、通常は一台の故障に耐える仕組み

どのようなものか

RAIDは、複数のディスクを一つのブロックデバイスとして動かす仕組みで、構成するどの一台よりも、速度、容量、または故障への強さを高められます。名前の由来は、カリフォルニア大学バークレー校のDavid Patterson、Garth Gibson、Randy Katzが1987年に発表した技術報告、*A Case for Redundant Arrays of Inexpensive Disks (RAID)*（安価なディスクによる冗長アレイの提案）です。Linuxは、raid1やraid5のように、レベルを小文字で、空白を入れずに表し、mdドライバーも、その形で報告します。

報告は、不釣り合いな進歩から始まります。プロセッサーとメモリーは速くなっていましたが、要旨は、入出力が追い付かなければ、その向上が無駄になると警告します。容量は急増しても性能の向上は緩やかだった、Single Large Expensive Disk、略してSLEDという一台の大型で高価なディスクを、パソコン向けの安価なディスクで組むアレイと比較します。後者には、速度、信頼性、消費電力、拡張の余地で、大型ディスクに対しておよそ十倍の改善を期待しています。比較対象は具体的で、メインフレーム用のIBM 3380、ミニコンピューター用のFujitsu Super Eagle、パソコン用の100 MBのConner CP 3100です。

続いて、著者らが悪い知らせと名付けた節で、落とし穴を示します。ディスクが互いに独立し、一定の率で故障すると仮定すると、アレイの平均故障時間（MTTF）は、一台のMTTFを台数で割った値になります。報告によると、これは、メーカーがMTTFを示すときに置く仮定です。CP 3100は一台30,000時間とされますが、百台集めると300時間で、二週間未満です。千台なら30時間です。結論は明確です。「耐障害性がなければ、安価なディスクの大規模アレイは、信頼性が低すぎて役に立たない」というものです。

答えは、冗長性です。ディスクを信頼性グループに分け、それぞれに、冗長情報を保持する追加のチェック用ディスクを置きます。一台が故障したら交換し、その情報から新しいディスクへ内容を再構築します。これにかかる時間が平均修復

時間で、故障時に電子的に切り替えて使う、ホットスタンバイの予備ディスクが、時間を短くします。報告は、五つの構成を示し、それぞれをレベルと呼びます。レベル1は、すべてのディスクをミラー化します。レベル2は、メモリーシステムと同様に、ハミング符号で誤りを訂正します。レベル3は、グループ当たり一台のチェック用ディスクだけを必要とします。レベル4は、小さい読み書きを、別々のディスクで同時に行えます。レベル5は、データとチェック情報を、すべてのディスクへ分散し、一台のチェック用ディスクが書き込みを待たせることを避けます。

実際の運用では、その後に種類が増えました。LinuxのソフトウェアRAIDを管理するmdadmは、linear、raid0（またはstripe）、raid1、raid4、raid5、raid6、raid10などを受け付けます。現在レベル0と呼ぶ、冗長性のないストライピングは、ミラー化したディスクから始まる1987年の分類には含まれません。報告は、ハードウェアとソフトウェアのどちらでアレイを構築するかを未決とし、費用と便益だけの問題としました。Linuxのmdはソフトウェア方式で、OSへ一台の大きいディスクとして見せるコントローラーカードは、ハードウェア方式です。

概念を判断に生かす

「RAIDとは？」を実際の判断に生かしましょう。その概念を、開発または購入するシステムに結び付けて考えるお手伝いをします。 あると信じるだけでなく、復元を試したバックアップと、サーバーが起動しない朝に使える手順書を用意します。

サービスを見る お問い合わせ →

用語の背後にある前提を検証する

解説とリンク先の実績を参考に、プロジェクトについての質問を整理してください。ツールを選ぶ前に、前提を一緒に検討できます。 Barmanで重要データのバックアップ100件を自動化 — 100 データベース全体で自動化した重要バックアップ

Barmanで重要データのバックアップ100件を自動化 お問い合わせ →

なぜ重要なのか

ディスクは劣化し、一台だけのサーバーでは、その故障によってマシンが止まり、データを取り戻せるかどうかは、存在するバックアップにかかっています。報告の信頼性の節も、まさにこの点から始まります。ディスクの信頼性が低く、システム管理者は、毎晩すべてをバックアップしている、という説明です。冗長性があれば、一台の故障は、サービス停止ではなく、修理作業になります。残りのディスクでアレイは動き続けます。Linuxがdegradedと呼ぶ縮退状態で、交換用のディスクを追加して再構築すると、完全な状態に戻ります。

この保護が続くのは、残りのディスクが持ちこたえる間だけです。報告の計算は、最初の一台を修復する前に、同じグループの二台目が故障すると、データを失うものとします。そのため、修復にかかる時間は、チェック用ディスクの数と

同じくらい重要です。ホットスペアを用意し、修復を短くすると、数値は目を引くものになります。報告は、レベル5のアレイを1,000台で組んだ場合でも、十台ずつのグループに分け、5パーセントの待機用の予備を持たせれば、MTTFは90年になると計算しています。一方、別の箇所では、地震や電圧サージによって、同じアレイ内のディスクが同時に故障し得ると警告しています。計算式は、各ディスクの故障が互いに独立して起こると仮定しています。

レベルの選択は、容量、速度、費用の間のトレードオフです。報告自身の結論では、記憶容量の半分までしか使わないトランザクション処理には、ミラー化したディスクが適します。より多くの容量が必要な場合、スーパーコンピューターの処理、または、それとトランザクション処理の組み合わせには、レベル5が最もよさそうだとしています。

冗長性を使い切った後に残るのは、復旧です。SDカード、ハードディスク、SSD、RAIDからのデータ復旧のポートフォリオ項目は、故障したRAIDアレイを、停止したSSDや故障した磁気ディスクとは別の問題として扱い、媒体に応じて適切な方法が変わるとしています。

実際の活用例

二台のディスクに、それぞれ `/dev/sda1` と `/dev/sdb1` というパーティションを一つずつ持つサーバーで、データ用のボリュームをミラー化する例です。Linuxのmdでは、三つのコマンドで、アレイを作り、ファイルシステムを置き、設定に記録します。

```
mdadm --create /dev/md0 --level=1 --raid-devices=2 /dev/sda1 /dev/sdb1
mke2fs -t ext4 -L data /dev/md0
mdadm --detail --scan >> /etc/mdadm/mdadm.conf
```

最初のコマンドは、 `/dev/md0` というRAID1アレイを二つのパーティションから作り、カーネルは、作成後に毎回行うように、冗長情報を計算し直す再同期を始めます。二つ目は、一台のディスクと同じようにアレイをフォーマットし、ファイルシステムからは、一つのデバイスに見えます。三つ目は、アレイを記述する行を、mdadmの設定ファイルへ追記します。場所は `/etc/mdadm.conf` か `/etc/mdadm/mdadm.conf` です。マニュアルは、その出力に不要な詳細が含まれ得るため、使う前に確認するよう求めています。

アレイの状態は、三つのファイルで分かります。

```
cat /proc/mdstat
cat /sys/block/md0/md/degraded
cat /sys/block/md0/md/sync_action
```

`/proc/mdstat` は、稼働中の全mdデバイスと状態を一覧にし、mdadmは、アレイの発見と再構築の追跡に読みます。各アレイは、 `/sys/block/` の下にもディレクトリーを持ち、その中の `degraded` が、不在の構成ディスク数を保持します。すべてそろっていれば0、一台なければ1です。 `sync_action` は、何もしていなければ `idle` 、作成後や正常に終了しなかった後は `resync` 、失われたディスクの内容をホットスペアが引き継ぐ間は `recover` 、全体の検証中は `check` か `repair` になります。ここへ `check` を書き込むと、アレイの全ブロックを読む検証が始まります。

```
echo check > /sys/block/md0/md/sync_action
cat /sys/block/md0/md/mismatch_cnt
```

その後、 `mismatch_cnt` には、書き直したセクター数、または、書き直すことになったはずのセクター数が表示されます。カーネルのドキュメントは、多くのレベルが、セクターではなくページ単位で動くため、実際の誤りの数より大きくなり得ると説明しています。

故障を誰かに知らせることも必要です。 `mdadm --monitor` は、アレイを定期的に調べ、検出した内容をイベントにして、指定の宛先へメールで送るか、プログラムへ渡せます。systemdで動くシステムについて、マニュアルは、 `mdmonitor.service` を主な方法として挙げ、メールには `MAILADDR` を `mdadm.conf` で設定するよう示しています。メールを送るイベントは、 `Fail` 、 `FailSpare` 、 `DegradedArray` 、 `SparesMissing` 、 `TestMessage` の五つだけです。最後のものは、通知経路が機能することを確かめるためにあります。

```
MAILADDR ops@example.com
```

```
mdadm --monitor --scan --oneshot --test
mdadm --detail --test /dev/md0
```

最初のコマンドは、見つけた各アレイについてテストメッセージを送り、終了します。二つ目は、スクリプト向けです。`--test`を付けると、終了ステータスは、正常なら0、少なくとも一台が故障していれば1、故障が多すぎてアレイが使えなければ2、アレイの情報収集中にmdadmでエラーが起きれば4です。

アレイが備えるのは、ディスクが一台故障するという出来事だけで、それ以外には備えません。バックアップと復元テストの自動化の項目は、別の復旧設計を説明します。暗号化して重複を除いたresticのスナップショットを、SFTP経由で別の事業者のストレージへ保存し、systemdタイマーと、既知のファイルを一時領域へ復元してバイトを比較する月次ワークフローを使うものです。復元のテストが設計に含まれるという説明であり、実行済みという証拠ではありません。2026-10-27付の実装記録は、リポジトリーが未設定のままタイマーが無効だったとし、完了した実環境の復元を記録していません。

問題が起きるところ

アレイが想定する出来事は、構成するディスクの一台が故障することです。データを失う原因の多くは、その出来事のすぐ隣にあります。

アレイをバックアップだと考えます。 ミラーは、すべての変更を両方のディスクへ同時に書きます。誤って消したファイル、不適切なマイグレーションが削除したテーブル、ランサムウェアが暗号化したボリュームも含みます。その結果、すべてのディスクが、同じ損傷を保持します。対策は、ホストの外へ別のバックアップを置き、定期的に復元を試すことです。アレイはハードウェアの故障に、バックアップは、それ以外に備えます。

縮退したアレイに気付きません。 縮退してもアレイはサービスを続け、次のディスクが故障するまで、見た目は変わりません。mdadmは、宛先かプログラムがなければ、メールを送りません。マニュアルによると、`--scan`を使い、そのどちらも設定していない場合、何も監視しません。また、`DegradedArray`は、mdadmが初めて見たときに既に縮退しているアレイだけが対象です。後でディスクが故障すると、`Fail`イベントになります。`MAILADDR`を設定し、`mdmonitor.service`が動いていることを確認し、`--test`メッセージを送って、メールが届くことを確かめます。

再構築中に、二台目が故障します。 再構築は、ほかのすべてのディスクから、失われたディスクの内容を組み直します。報告のモデルでデータを失うのは、まさに、その修復中に、同じグループ内の別のディスクが故障した場合です。故障したディスクを一週間放置すれば、その危険が続く期間も一週間になります。対策は、期間を短くすることです。アレイ内にホットスペアを置きます。mdadmで

は、作成時に `--spare-devices` で用意できます。故障したディスクを速やかに交換し、定期的な `check` を行い、アレイがまだ訂正できる間に、問題のあるディスクを見つけます。

ディスクの故障が独立していません。 MTTFの計算は、ディスクがそれぞれ単独で故障すると仮定します。報告自身が、そうならない例に、地震と電圧サージを挙げています。同じ筐体の二台は、電源、コントローラー、部屋、建物を共有します。対策は、計算値を上限と捉え、別の場所にあるコピーなど、筐体の外側で共通の原因に備えることです。

RAID 5か6で、dirty状態と縮退状態が重なります。 一台が欠けた状態でマシンが異常終了すると、パリティーを信頼できず、一部のブロックを再構築できません。そのため、カーネルのドキュメントは、検出できない破損を警告し、mdは、アレイの起動を拒否します。 `mdadm --assemble --force` や、ルートファイルシステムの場合の `start_dirty_degraded` モジュールパラメーターで強制することは、その危険を受け入れることです。対策は、整合性の方針を先に決めることです。mdは、レベル4、5、6で、ジャーナル用のデバイスへ書き込みを記録できます。レベル5には、部分パリティーログもあり、ドキュメントによると、書き込みホールを防ぎます。強制するしか方法がないときは、データを疑わしいものとして扱い、バックアップと照合します。

RAID 0を、データを保護する仕組みと見なします。 mdは、raid0とlinearのアレイには冗長情報がないとし、mdadmの監視は、 `Wrong-Level` を報告します。監視が対象とする、故障、予備ディスク、再同期に関する操作のどれにも対応しないからです。失われたディスクの内容を再構築する材料はありません。レベル0は、捨ててもよい、または再作成できるデータにだけ使い、1987年の報告がいう意味でのRAIDとしては、扱わないようにします。

ハードウェアのコントローラーが、ディスクを隠します。 コントローラーカードの背後では、OSからは一つの論理デバイスに見え、そのデバイスへディスクの健全性を尋ねても、届くのはコントローラーで、ディスクではありません。たとえばsmartctlは、MegaRAIDコントローラーの背後のディスクNを指定するために、 `-d megaraid,N` を必要とします。マニュアルによると、この指定はDell PERCカードにも使えます。対策は、コントローラー自身のインターフェースを通して、物理ディスクを一台ずつ監視し、アレイがまだ完全な状態のうちに、故障しつつあるディスクを見つけることです。

参考資料

- 安価なディスクによる冗長アレイ（RAID）の提案 — 要旨とPDFを掲載する、バークレー校の技術報告のページです。
- UCB/CSD-87-391、報告全文（PDF） — MTTFの計算、信頼性グループ、五つのレベルです。
- RAIDアレイ、Linuxカーネルのドキュメント — `/sys` 以下のmdのファイル、dirty状態かつ縮退状態のアレイ、整合性の方針です。
- mdadm(8) — アレイの作成、監視イベント、メール、 `--detail --test` の終了コードです。
- smartctl(8)、smartmontools — `-d` オプションによる、RAIDコントローラーの背後のディスクへのアクセスです。

用語集

実際の制約に合わせて選ぶ

デプロイの流れ、運用上のリスク、緊急対応の担当者をお知らせください。検査、復旧手段、明確な責任の所在を伴う変更の範囲を定められます。

料金を見る お問い合わせ →

用語を実務に結び付ける

「RAIDとは？」がお客様の組織にどう当てはまるかをご相談ください。現在のシステムの具体例が一つあれば、話を始められます。

サービスを見る

welcome@engineer.company +45 50 10 61 84 デンマーク、コペンハーゲン

このページを持ち歩く

連絡先カード

心を込めて丁寧に制作： D.

CVR 45969703 · Tagensvej 105C, 3. th, 2200 København N

著作権 © 2025 – 現在 · Engineer ApS · すべての権利を留保します。

印刷元： https://engineer.company/ja/glossary/raid/