Engineer

ファイルシステム

ディスクを名前付きのファイルとディレクトリーにする構造

どのようなものか

ファイルシステムは、そのままではバイトを番号付きのブロックでしか保存しないディスクを、ディレクトリーへ並べた名前付きファイルへ変える、コンピューターの仕組みです。英語ではfilesystemと一語でも書き、Linuxカーネルの文書は、主にその表記を使います。Unix系システムが従う標準、POSIXは、2024年版で「ファイルと、その属性の一部からなる集合」と定義します。続けて、ファイルシリアル番号は、ファイルシステムの提供する名前空間に属するとします。各ファイルを、保存先のファイルシステム内で、一意に識別する番号です。

残りは、標準のいくつかの定義から分かります。ファイルは、書き込み、読み取り、または両方ができるオブジェクトで、アクセス権や種類などの属性を持ちます。種類には、通常ファイル、ディレクトリー、シンボリックリンク、FIFO、ソケット、文字型やブロック型の特殊ファイルがあります。ディレクトリーは、その中でエントリー名が重複しない、ディレクトリーエントリーの集合からなるファイルです。POSIXがハードリンクとも呼ぶエントリーは、ファイル名をファイルへ結び、複数のエントリーが同じファイルを指せます。そのため、名前とファイル自体は、別に保持します。Linuxでファイル自体の記録となるのはinodeです。GNUのdfのマニュアルは、所有者、権限、タイムスタンプ、ファイルシステム上の場所を保持すると説明します。名前は、それを指すディレクトリー側にあります。

Unix系システムは、すべてのファイルシステムを、名前の階層からなる一つの木構造として見せます。最初のものがルートディレクトリーを提供し、ほかは、木の中のディレクトリーへ接続します。それがマウントポイントで、接続後、その下のパスは、接続したファイルシステムに属します。Linuxカーネル内では、プログラムと、多数の対応ファイルシステムの間に、Virtual File Systemがあります。カーネルの文書は、Virtual Filesystem Switchとも呼び、プログラムへファイルシステムのインターフェースを提供し、異なる実装を共存させる層と説明します。プログラムが `open(2)` 、 `stat(2)` 、 `read(2)` 、 `write(2)` を呼ぶと、VFSがパ

スを解決し、ファイルを持つ実装へ仕事を渡します。ディスクのパーティション上のext4や、inodeがメモリー内だけにある疑似ファイルシステムなどです。実行中のカーネルに登録された種類は、`/proc/filesystems` へ並びます。

概念を判断に生かす

「ファイルシステムとは？」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。 サーバー、メール、アカウント、更新を日々管理し、誰もが頼る仕組みを平穏に動かし続けます。

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。 UbuntuでPostgreSQL・MS SQLサーバー10台を運用 — 10 導入し正常に保ったDBサーバー

UbuntuでPostgreSQL・MS SQLサーバー10台を運用 お問い合わせ →

なぜ重要なのか

次の再起動後もサーバーが保つものの、ほぼすべてを保存するのが、ファイルシステムです。データベースのデータファイル、ログ、アップロードした文書、設定などです。停電、容量不足、ドライブの故障に耐えて残るかは、作成時とマウント時の選択が、大きく左右します。後から変えられない選択もあります。

作成時には、二種類の容量を決めます。ブロックが内容を、inodeがファイル自体を保持します。ext2、ext3、ext4を作るmke2fsは、一定のバイト数の領域ごとに、一つのinodeを作ります。マニュアルは、このinodeあたりのバイト数の比率を、作成後は変更できないと警告します。内容を置く空き領域があっても、新しいファイルを作る余地が尽きる場合があります。使用容量の割合だけを見ていると、意外に感じます。

マウントは、クラッシュの代償を決めます。ext4はジャーナルを持ち、文書は三つのデータモードを挙げます。既定の `data=ordered` は、参照するメタデータをジャーナルへ確定する前に、ファイルデータを本体のファイルシステムへ書き出します。`data=writeback` は、その順序を保証しません。正常に終了しなかった後、その直前に書いたファイルへ古いデータが現れ、一部ではセキュリティ上の危険にもなり得ると警告します。`data=journal` は、データもジャーナルを通しますが、非推奨で、2028年1月に対応を削除する予定です。`errors=` は、ファイルシステム自体にエラーを見つけたとき、読み取り専用へ再マウントするか、続行するか、パニックを起こしてマシンを停止するかを決めます。

復旧も、この仕組みに左右されます。ファイルシステムが壊れても、内容はディスクにあり、名前を付ける構造が壊れていることがあります。データ復旧は、その状況から始まります。SDカード、ハードディスク、SSD、RAIDからのデータ復

旧の事例は、媒体ごとに壊れ方も、データを取り出す方法も異なり、Macのファイルシステムにも固有の癖があると述べます。そこで重要だったのは、単一の技法より、作業台にある媒体に合う方法を選ぶことでした。

実際の活用例

新しいパーティション、`/dev/sdb1`へ、サービスのデータを置き、`/srv/data`から使うとします。ext4を作り、手でマウントするには、三つのコマンドを使います。

```
mke2fs -t ext4 -L data /dev/sdb1
mkdir -p /srv/data
mount -t ext4 /dev/sdb1 /srv/data
```

`-t ext4`が、カーネルのext4の文書と同じように種類を指定し、`-L`がラベルを付けます。mke2fsは、ラベルを16バイトに制限します。ほかは既定値です。例えば、e2fsprogsが配布する設定では、各inodeは256バイトです。二つの`df`コマンドが、二種類の容量を示します。

```
df -hT /srv/data
df -i /srv/data
```

最初は、種類と、使用中および使用可能なブロックを、1024の累乗に基づく単位で出します。`-H`なら、1000の累乗です。二つ目は、ブロックの代わりに、inodeの総数、使用中の数、利用可能な数、使用率を出します。

手で行ったマウントは、次の再起動までです。恒久化するには、mount、umount、fsckが上から順に読む設定ファイル、`/etc/fstab`へ、一行を加えます。

```
LABEL=data  /srv/data  ext4  defaults,nofail  0  2
```

最初のフィールドは、`/dev/sdb1`ではなく、ラベルで指定します。fstabのマニュアルは、`LABEL=`か`UUID=`を勧めます。デバイス名は、ハードウェアの検出順に従うことが多く、ディスクの追加や取り外しで変わるためです。`defaults`は、通常、読み書き可能と、ほかのカーネルの標準オプションを意味します。

`nofail` は、デバイスがなくてもエラーを報告しないよう、mountに指示します。五番目のフィールドはdump(8)が読み、0なら対象から外します。六番目は、起動時にfsckが検査する順序で、ルートは1、ほかは2、検査なしは0です。systemdのシステムでは、編集後の `systemctl daemon-reload` を勧めます。 `mount -a` が、 `noauto` 以外の全エントリーをマウントするため、追加行の誤りを、まだ人がキーボードの前にいる間に見つけられます。

空き領域を見ることも、運用の一部です。通信が途絶えたら警報を出すホスト監視は、forgeとウェブサーバーを動かす464 MBのマシンが止まる、現実的な二つの原因に、メモリー不足とディスク容量不足を挙げ、どちらも自らは知らせないと指摘します。対処は、タイマーが一時間に四回、毎回少し時刻をずらして起動する、小さいスクリプトです。ディスクの空き容量と使用可能メモリーを測り、両方がしきい値を超える間だけ、外部サービスへpingを送り、止まったことで警報が出ます。同じ形の検査で、容量の列とともに、 `df -i` のinodeの列も読めます。どちらが先に限界へ達するか、決まっていないためです。

問題が起きるところ

問題の多くは、満杯のディスク、昨日マウントできたものをマウントできないマシン、クラッシュ後のデータ消失として現れます。作成時に、誰も確かめなかった設定へ、原因をたどれるものも多くあります。

空き領域があるのに、満杯になります。 各ファイルにinodeが必要で、mke2fsは、作成時に固定数を用意します。何百万もの小さいファイルを持つボリュームは、 `df -h` が空きを示しても、inodeを使い切り、以後、新しいファイルを一切作れなくなる場合があります。 `df -h` とともに `df -i` も監視します。小さいファイルを大量に置くものは、mke2fsの実行時に、 `-i` で比率、または `-N` で個数を設定します。後から比率は変えられません。サイズの変更でも比率は保たれ、それに合わせてinodeが増減します。

マウント中に、検査します。 e2fsckのマニュアルは、マウント中の実行は、一般に安全ではないとします。唯一の例外は、読み取り専用で開き、すべての質問へ、いいえと答える `-n` です。それでも、マウント中に出す結果は、有効ではありません。e2fsckが、マウント中の実行を続けるかと聞いたら、マニュアルの答えは、いいえです。先にアンマウントするか、ルートなら、起動時に実行させます。fstabの六番目のフィールドが、そのためにあります。

「読み取り専用」が、ディスクへ書き込みます。 ext4を `ro` でマウントしても、書き込みは止まりません。読み取り専用でも、ジャーナルを再生し、パーティションへ書くと、文書は述べます。救出中のディスクでは、まだ必要な何かを、書

き込みが上書きしかねず、特に重要です。`ro,noload` なら書き込みを防ぎますが、同じ文書は代償も示します。正常にアンマウントしなかった後、再生を省くと、不整合が残ります。稼働の継続より内容が大切なディスクは、`ro,noload` でマウントし、必要なものを取り出し、読んだ内容には不整合があり得るとして扱います。

ディスクの名前が変わります。 `/dev/sdb1` をマウントする行が指すのは、今日、カーネルが `/dev/sdb1` と呼ぶものです。fstabのマニュアルは、追加や取り外しで変わり得ると警告します。古いディスクの場所へ新しいものをマウントしたり、起動時に古いものが見つからなくなったりします。最初のフィールドへ `LABEL=` か `UUID=` を使います。ラベルは、`mke2fs -L` で、または後で `e2label` などで設定します。

書き込みがディスクへ届く前に、停電します。 ext4は、既定では五秒ごとにジャーナルを確定します。停電で、直近五秒のメタデータの変更を失うことがありますが、ジャーナルが、ファイルシステム自体の破損を防ぎます。ファイルデータは、さらに古い時点の分まで失われることがあります。ブロックの遅延割り当てでは、`/proc/sys/vm/dirty_expire_centisecs` で決める間隔の後で、書き出しが始まるためです。文書は、`fsync()` なしでファイルを置き換えるアプリを、壊れていると呼びます。ext4の既定の `auto_da_alloc` は、新しいファイルを書いて古いものへリネームする場合と、その場で切り詰めて書き直す場合の、二つの典型的な形を検出し、クラッシュ後に空になるのを避けます。直すべき場所は、書くプログラムです。安全網に頼らず、保存したと伝える前に、`fsync()` を呼びます。

タイムスタンプが、2038年で止まります。 mke2fsのマニュアルは、128バイトのinodeでは、2038年1月19日より後を保持できず、256バイト以上なら、拡張タイムスタンプに対応するとします。新しいものは、既定で256バイトですが、古いものや、`-I 128` で作ったものは、小さいままの場合があります。その日が来る前に、対象を見つけます。ext4の文書は、128バイトのinodeで作ったものを、`tune2fs -I 256` で変換する方法を示します。

参考資料

- 定義：POSIX.1-2024、The Open Group基本仕様第8版 — ファイルシステム、ファイル、ディレクトリー、ハードリンク、マウントポイントの定義です。
- Linux Virtual File Systemの概要 — カーネルのVFSがパスを解決し、各実装へ仕事を渡す仕組みです。
- ext4の一般情報 — ext4の作成とマウント、データモード、`commit=`、`ro` とジャーナルです。

- mke2fs(8) — inodeあたりのバイト数、inodeの大きさ、2038年の限界、ボリュームラベルです。
- e2fsck(8) — ext2、ext3、ext4の検査と、マウント中に実行してはならない理由です。
- fstab(5) — 六つのフィールド、`LABEL=`、`UUID=`、`nofail`、起動時の検査順です。

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

「ファイルシステムとは？」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/file-system/