3. ZFSの信頼性/可用性

ZFSの信頼性・可用性

■ZFSの信頼性と可用性を支える仕組み

✓コピーオンライト
ファイルへの書込み時に元データを上書きしない仕組み

✓トランザクションファイルシステム
データの更新時に整合性を確保する仕組み

✓チェックサム
End-to-Endのチェックサムによる信頼性の高いデータ保護

✓スクラブ
ストレージプール内の全データの完全性をチェックする機能

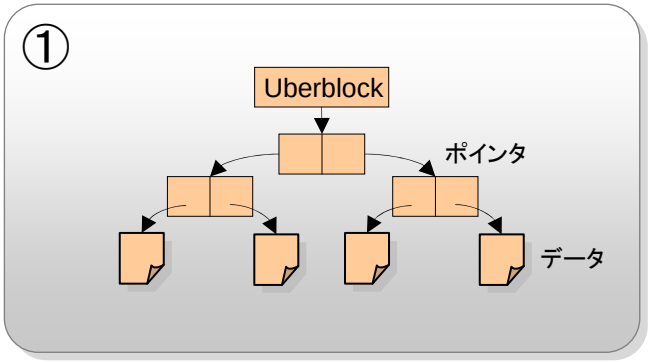
✓RAID-Z RAID-Z2
物理ディスクの損傷時にデータを守るZFS特有のRAID構成

コピーオンライト

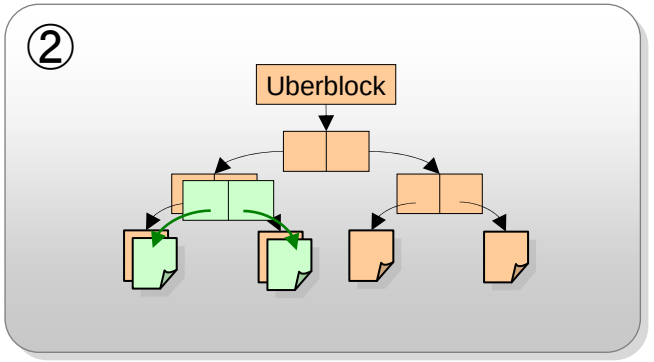
■コピーオンライトによるデータ更新

ZFSは、コピーオンライト(Copy On Write)というデータの更新方法をとっており、ファイルシステムやファイル自体に矛盾が発生しない仕組みになっています。

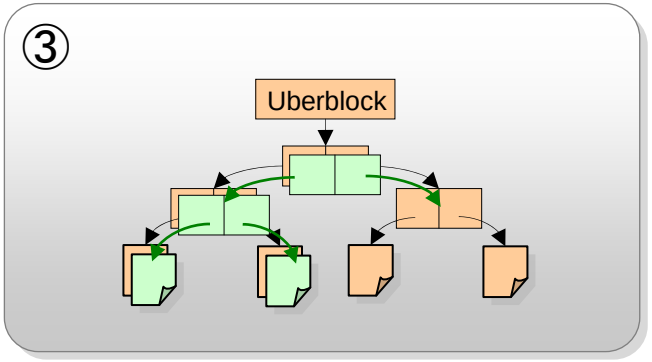
コピーオンライトでは、いったん書き込んだデータを変更するとき、まず元のデータのコピーを作成しそのコピーの方を変更します。その後、システムにとって区切りの良い時点でデータブロックのリンクを更新します。



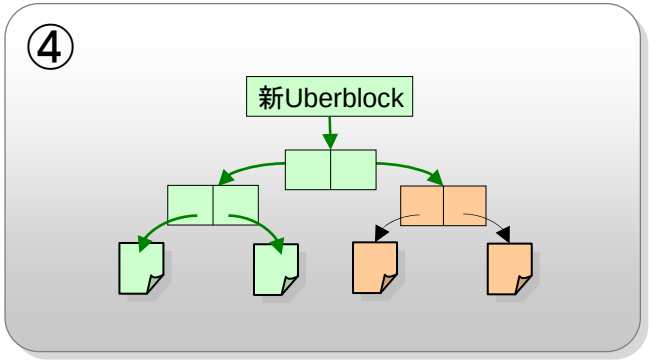
初期状態のブロック構造



データを含むブロックを複製し、更新する



ポインタを含むブロックを複製し、新しいデータにリンクする



Uberblockを更新する

トランザクションファイルシステム

■ ZFSはトランザクションファイルシステムを採用

ZFSファイルシステムは、一連の書き込み動作を一つの更新単位（トランザクション）として扱います。

トランザクションファイルシステムの利点

✓**データの不整合が発生しない**
データ更新が「全て成功」か「全て失敗」のどちらかであることが保障される

✓**ファイルシステムの整合性チェックが不要（UFSファイルシステムでは必須）**
不意にシステムがクラッシュしてもファイルシステムを保護します

トランザクションファイルシステムの留意点

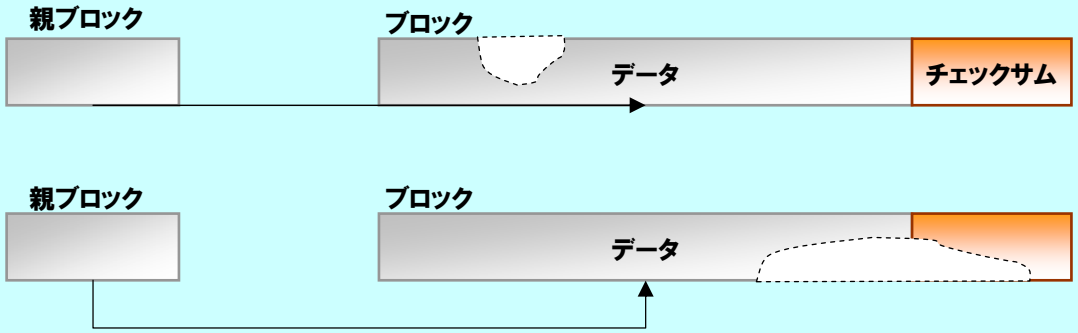
✓**ディスクへの書き込みが非同期である**
一連の処理が終了してからディスクへ書き込み（I/O発生）が行われるため、実際のディスク容量を確認したときに差異が発生することがあります（dfコマンド、duコマンドなど）

チェックサムとデータの自己修復

■End-to-Endのチェックサムによりファイルシステムの保護機能を強化

データとチェックサムを異なるブロックに保存することで、チェックサム自身の信頼性を向上しています。

従来のファイルシステム



データの損傷範囲がチェックサムにまで及ぶとデータ修復ができない。

ZFSファイルシステム

データとチェックサムが物理的に離れて別々に読み込まれる。チェックサム自身も上位ブロックによって修復可能。

- ディスク故障による読み出しエラーを検出するだけではなく、ソフトウェアのバグなどによる論理的な不整合が発生しても検出と訂正が可能です。
- データの読み書き時は、チェックサムによって不正データが検出されます。不正データが検出された場合、冗長構成(ミラー、RAID-Z、RAID-Z2)であれば複製データから自動的にデータを修復します(Self-Healing)

スクラブ

■ZFSファイルシステムのデータの完全性を明示的にチェックする機能

基本的にZFSではチェックサムによってデータの損傷は自動修復（Self-Healing）されますが、明示的にファイルシステムを検査する場合には「スクラブ」を実行します。
スクラブは全てのデータを順にたどってみて、すべてのブロックが読み取り可能であることを確認します。

スクラブの特長

✓スクラブによってハードウェア障害やディスク障害に起因する非表示のエラーも検出します
✓定期的にスクラブを実行することで、ストレージプール上のディスクの完全性を継続的にチェックさせることができます。
✓スクラブ実行中にI/Oが発生するとスクラブの優先順位は低くなります（業務へ影響を与えません）。

●スクラブ実行直後

```
# zpool scrub rpool
# zpool status rpool
 pool: rpool
 state: ONLINE
 scrub: scrub in progress for 0h0m, 0.02% done, 11h58m to go
 config:

        NAME        STATE     READ WRITE CKSUM
        rpool       ONLINE       0     0     0
          mirror    ONLINE       0     0     0
            c0d0s0  ONLINE       0     0     0
            c0d1s0  ONLINE       0     0     0

errors: No known data errors
```

●スクラブ完了後

```
# zpool status rpool
 pool: rpool
 state: ONLINE
 scrub: scrub completed after 0h13m with 0 errors on Thu Feb  4 15:52:42
2010
config:

        NAME        STATE     READ WRITE CKSUM
        rpool       ONLINE       0     0     0
          mirror    ONLINE       0     0     0
            c0d0s0  ONLINE       0     0     0
            c0d1s0  ONLINE       0     0     0

errors: No known data errors
```

商標について

- SPARC Enterpriseは、米国SPARC International, Inc.のライセンスを受けて使用している、同社の米国およびその他の国における商標または登録商標です。
- UNIXは、米国およびその他の国におけるオープン・グループの登録商標です。
- Sun、Sun Microsystems、Sunロゴ、SolarisおよびすべてのSolarisに関連する商標及びロゴは、米国およびその他の国における米国Sun Microsystems, Inc.の商標または登録商標であり、同社のライセンスを受けて使用しています。
- すべての SPARC 商標は、SPARC International, Inc. のライセンスを受けて使用している同社の米国およびその他の国における登録商標です。SPARC 商標が付いた製品は、Sun Microsystems, Inc. が開発したアーキテクチャーに基づくものです。
- SPARC64 は、米国 SPARC International, Inc. のライセンスを受けて使用している同社の登録商標です。
- ORACLE, SQL＊Plus, SQL＊Forms, SQL＊Net, Pro＊C, Pro＊FORTRAN, Pro＊COBOLは、ORACLE Corporationの登録商標もしくは商標です。
- その他各種製品名は、各社の製品名称、商標または登録商標です。

留意事項

- 本書の内容は、改善のため事前連絡なしに変更することがあります。
- 本書の内容は、細心の注意を払って制作致しましたが、本書中の誤字、情報の抜け、本書情報の使用に起因する運用結果に関しましては、責任を負いかねますので予めご了承願います。
- 本書に記載されたデータの使用に起因する第三者の特許権およびその他の権利の侵害については、当社はその責を負いません。
- 無断転載を禁じます。