TINTRÍ

Ensuring Robust Data Integrity

TECHNICAL WHITE PAPER

総括

企業の業務データを格納する特化型ストレージアプライアンスは、付加価値のある独自のOSとファイルシステムソフトウェアを搭載した汎用的なハードウェアで設計されています。アプライアンスを構成するコンポーネント（ハードウェアとソフトウェア）は時として、高確率で障害を引き起こします。例えば、デュアルコントローラシステムにおける障害はデータ処理に影響がなくても、直接ユーザーが感じてしまう障害です。その他にも、ファームウェアエラーは、わずかな変化が後にデータ破損を引き起こす場合もあります。

Tintri™ VMstore™アプライアンスは、コスト効率を考慮したマルチレベルセル（MLC）フラッシュのソリッドステートドライブ（SSD）とハイキャパシティのSATAハードディスクドライブ（HDD）を最大限活用した、VMware専用のファイルシステムです。Tintri OSはVMstoreアプライアンス上で稼働し、ハードウェアおよびソフトウェアコンポーネントの誤作動を防ぐための包括的なデータ保全と信頼性の高い機能を搭載しています。これらの機能により、最高水準のパフォーマンスとシステム可用性を実現しています。

本ホワイトペーパーは、Tintri OSとそれを支えるファイルシステムがどのように１つの筐体で数百のVMを稼働させながら高水準のデータ保全性とパフォーマンスを提供するかを解説しています。また、Tintriが保有する包括的なデータ保全性のための、独自のRAID構造についても解説しています。

始めに

ストレージ、特にプライマリストレージにおいてデータ保全性は必要不可欠です。どんなハードウェアおよびファームウェアのエラーにも対処するため、何重にも堅牢な対策を講じる必要があります。想定される障害は以下の通りです。

- コントローラ障害またはドライブ障害。
- HDDやSSDの個別のコンポーネントによる多数の不自然な動作。例えば、ドライブがリード処理において壊れたデータを返したり、ライト処理を行えなかったりする障害。
- 物理ストレージに書き込まれた後のデータの頻繁な移動。データの移動は潜在的にデータの保全性を欠損させる恐れがある。例えば以下のような場合。
  - 非対称なリード／ライト処理によって生じる、ガベージコレクションのようなSSDの複雑な内部のメカニズムから発生する障害。
  - 不良セクターによるHDDのリマップ障害。
  - 重複除外処理によるファイルシステムのガベージコレクション障害。
- 重複除外処理のような高度な機能により、小さなエラーが大きな問題に発展。重複除外処理を行った結果、多数のファイルが同じデータブロックを参照するため、１つのブロックに起こる不具合が関連する全てのファイルに影響を及ぼす。

包括的なデータ保全の概念は、コンポーネントの不具合のような簡単な問題だけでなく、上記のようなケースを全て網羅する必要があります。データ保全の機能は、エンドトゥーエンドの保全性を保証するため円滑に機能しなければなりません。図１でTintri OSのアーキテクチャーコンポーネントの概要を示します。

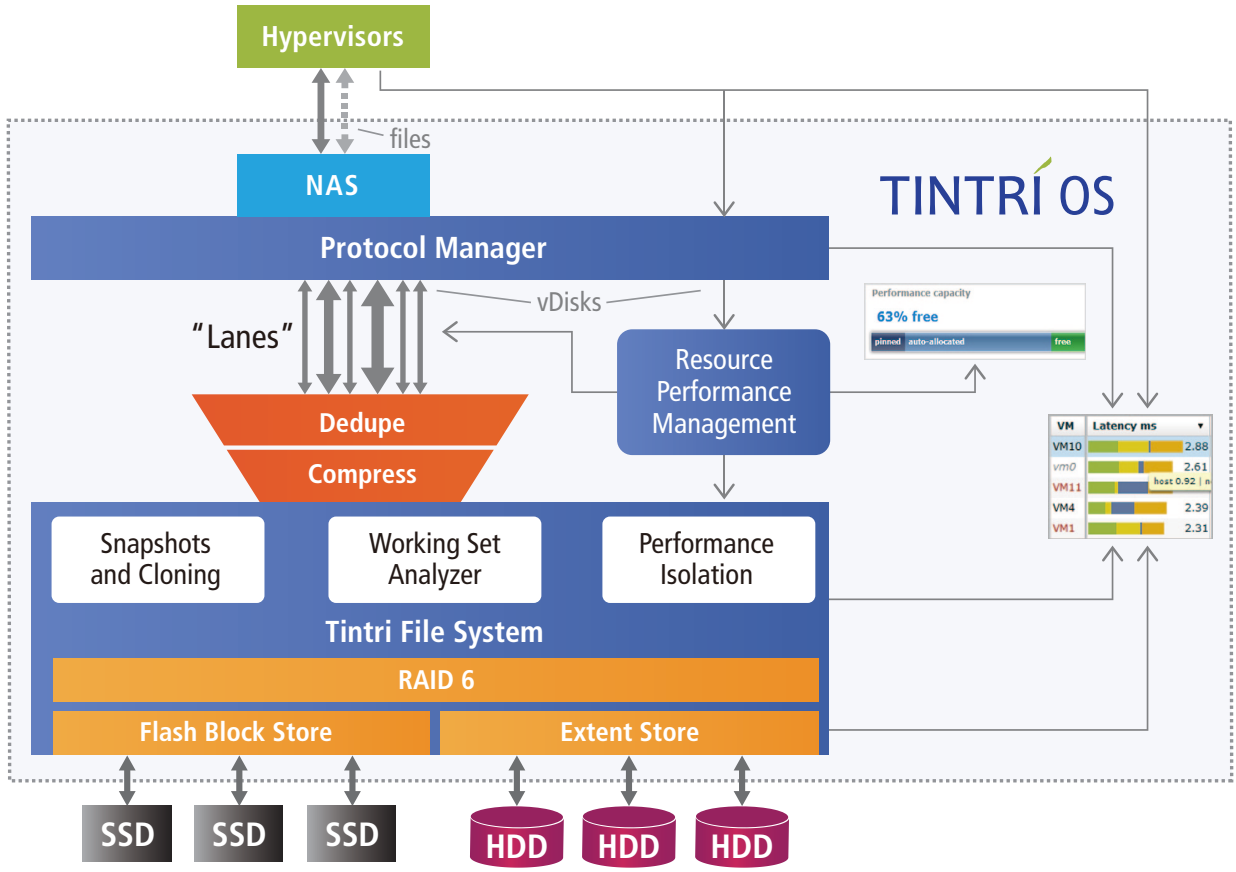

図 1：Tintri OSのアーキテクチャー概要

Tintri OSにおけるデータ保全性

Tintri OSの包括的なデータ保全の概念には以下のような特長があります。

- VMに特化して開発されたOS
- エラーを回避する能動的な設計
- RAID 6によるリアルタイムなエラー修復
- インラインの保全性とベリフィケーション
- 自動修復ファイルシステム
- 高可用性

VM専用に開発されたアプライアンス

従来の汎用的なストレージシステムとは異なり、Tintri OSを搭載したVMstoreアプライアンスは、VMを理解した上で、稼働させる目的のためだけに開発されました。その設計は、Tintri OSがハイパーバイザーにのみデータを供給するため、VMにおける環境をシンプルにし、今後の設計を予測しやすくします。少ない稼働部品と、厳密な照合によってハイパーバイザーとTintri OS間の整合性が保証され、ソフトウェアとインフラストラクチャーの整合性といった、環境による相互作用から発生するエラーを排除します。現在、Tintri OSはNFSデータストアを利用したVMware vSphereハイパーバイザーに対応し、Citrix XenServerといった他のハイパーバイザーへの対応も予定しています。

エラーを事前回避する設計

ファイルシステムの保全性における最大のリスクは、新しいデータを書き込む際のソフトウェアエラーです。ガベージコレクションのように、データを上書きしてしまったり、メタデータのアップデートによって既存の構造を台無しにしたりする可能性があります。Tintri OSとそのファイルシステムは、書き込みのバッファリングにおいてはNVRAM (nonvolatile RAM)を利用し、RAID 6グループに組み込まれたSSDに直接データを書き込むことで上記のようなエラーを回避します。主な利点としては以下のような点が挙げられます。

- **部分ストライプ書き込みが発生しない**：Tintri OSのファイルシステムは、RAIDストライプにおいては1つのデータブロックだけをアップデートすることはありません。全ての新しい書き込みは、新しいRAIDストライプに(SSDにもHDDにも)そっくりそのまま書き込まれます。複雑な部分ストライプ書き込みと比較すると、ドライブで障害が発生した場合にもディスクの再構築が保証され、一定してデータが失われることはありません。以下の図では、フラッシュブロックストア(SSD)と外部ストア(HDD)で利用されるフルストライプ書き込みのスキームを示しています。

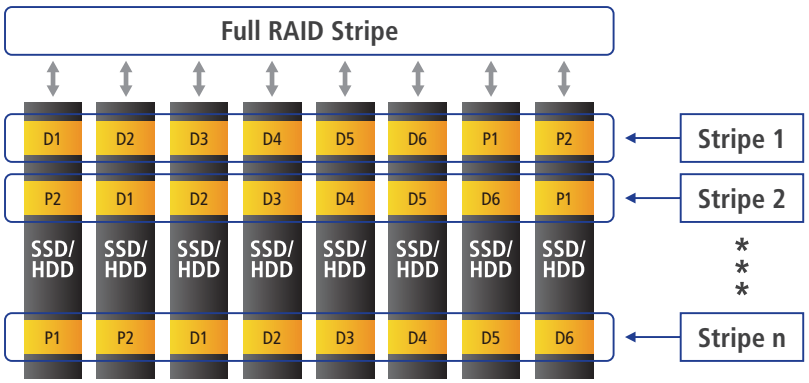

図2：SSDにもHDDにもフルストライプ書き込みが行われます

- **高速なバッファリングを行い、障害を防止するNVRAM**:
  - Tintri VMstoreアプライアンスは、全てのデータブロックがSSDに格納場所を確保される前にNVRAM書き込み用のバッファを準備します。NVRAMは、オンボードのフラッシュメディアに書き込みを行うことで電源の障害からデータを守ります。データブロックはNVRAMにバッファされ、SSDの耐用性と信頼性を高めるために大きな塊毎に順次フラッシュに書き込まれます。
  - Tintri OSのファイルシステムは、安全なファイルシステムの再起動を実現するためにNVRAMを搭載しています。再起動が始まると、Tintri OSのファイルシステムは、データをファイルシステムに適用する前にNVRAMバッファの中にあるデータの保全性をベリファイし、ファイルシステムの再起動によって失われたデータがないことを確認します。NVRAMで障害が発生した場合には、VMが安全に再起動できるかファイルシステムが保全性を確認します。

RAID 6とリアルタイムなエラー修復

二重ドライブ障害プロテクション

RAID 6はSSDとHDD上のデータ保護の基礎であり、継続的にエラーを検知し、リアルタイムに修復を行います。RAID 6における二重ドライブ障害プロテクションの構造は、RAID 1やRAID 5のような単体ドライブ障害の対処と比べて大きなメリットをもたらします。画一化されたRAID 6は、複数のRAIDレベルをシングルアレイに導入するストレージシステムとは異なり、容易に設定をすることができます。

RAID 6は、二重ドライブ障害の際にも可用性を実現します。そのため、VMstoreアプライアンスは2つのSSDと2つのHDDが同時に障害を引き起こしても稼働し続けることができます。従来のディスクベースのシステムとは異なり、アプリケーションのパフォーマンスに与える影響は最小限で済みます。SSDは早急なバックグラウンドでの再構築には驚異的なパフォーマンスを発揮し、99%以上のデータをフラッシュから提供します。

従来のストレージの多くはシングルパリティのRAIDを採用しており、2つのディスクで同時に障害が発生した際にデータを失ってしまう可能性がありました。シングルパリティのRAIDでは、1つのディスクで障害が発生し、同時に別のディスクでも障害が発生すると、データを失ったり壊したりしてしまう可能性がありました。RAIDプロテクションを採用せず、リードキャッシュのためだけにフラッシュを利用しているストレージシステムは、パフォーマンスを劇的に損なってしまう可能性があります。これらのシステムにおいては、全ドライブおよびキャッシュにある全てのデータは、障害が発生したドライブを交換した後で再構築する必要があります。

リアルタイムなエラー修復

ストレージのメディアにデータの一部が残されていたり、ファームウェアのエラーが発生したりしていると、データブロックを読み込めない場合があります。多くの場合、ドライブは潜在的な問題を抱えています。例えばライト処理を密かに失っていたり、誤った場所にブロックを書き込んだり、誤った場所からデータを読み込んだりします。ファイルシステムはこういったエラーを検知し、修復する機能を備えていなければなりません。Tintri OSのRAID 6ソフトウェアはリアルタイムにエラーを検知し、修復処理を行います。図3でRAIDを使ったリアルタイムなエラー検知と修復を図解しています。

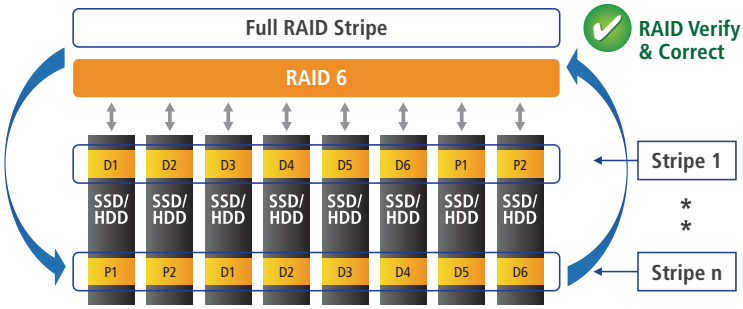

図3：RAID 6のリアルタイムなエラー検知と修復

Tintri OSの基本的なファイルシステムは、全てのオブジェクト（メタデータとデータ）を強力なチェックサムと共にブロック毎に管理しています。全てのリード処理において、Tintri OSのファイルシステムは、チェックサムを算出することでディスクから読み出されたオブジェクトをベリファイし、算出された値と合致するか確認します。問題が見つかれば、RAID 6が修復処理を行いシステムを自動的に修復します。

インラインの保全性保護とベリフィケーション

フラッシュを利用したストレージシステムは、消耗（MLC SSD）と信頼性を考慮しなければなりません。SSDはその非対称なプログラムにより、ストレージシステムをより複雑にする代わりにガベージコレクションのような内部の複雑性を排除します。フラッシュを組み込んだだけのボルトオン・プライマリストレージシステムは、最も基本的な障害防止（例えばRAID）さえも搭載していません。これはコストとパフォーマンスに大きく影響します。このようなシステムは高価なシングルレベルセル（SLC）フラッシュを使用していたり、リードキャッシュのためだけにフラッシュを利用していたりするからです。

Tintri OSがライト処理のリクエストを受信すると、データが既に保存されているかを確認し、チェックサムと一緒にデータブロックをSSD上に格納します。重複が見つかったブロックに関しては、データが同じであるか整合性を確認するためにフラッシュから既存のブロックにリード処理が施されます。チェックサムが算出され、データオブジェクト毎に（フラッシュにもディスクにも）格納され、オブジェクトが読み込まれる度にベリファイされます。一般的な自己修復プログラムとは異なり、DMAエラーや内部のメタデータ欠損などのためにディスクからリード処理を行う場合、自己完結チェックサムは必ずしも有効というわけではありません。つまり、インラインのチェックサムでは全てのデバイスエラーを見つけることはできないのです。Tintri OSは参照整合性を確認することでデータの欠損を見つけ、誤ったデータを返すことで起こる大きな問題を回避します。参照整合性を保証するために、データオブジェクトは読み込まれたオブジェクトのチェックサムに対してベリファイされた別のチェックサムを含んでいます。

これらの技術はSSDとHDD上のデータがリード可能で、データを検知するのに使われたファイルシステムのメタデータがリード可能であることを保証します。ディスクコントローラが悪いデータを返すような潜在的な問題は素早く検知され、修復されます。多くの場合、5ページの「リアルタイムなエラー修復」で解説した自動修復機能で修復可能です。

自動修復ファイルシステム

RAIDアシストのリアルタイムなエラー検知は、アクティブなデータに対しては効果的に機能しますが、長時間にわたってスナップショットに参照されるデータブロックのようなコールドデータに対してはエラーを検知しません。データ欠損を防ぐために、VMstoreアプライアンスは継続的なバックグラウンド処理の中で能動的にSSDとHDDのデータ保全性をリベリファイします。

HDDに格納されたデータにおいては、エラーを検知し修復するためのスクラブ処理が2種類あります。

- 新しいデータとそのチェックサムが書き込まれると、バックグラウンドの処理がディスクに書き込まれたRAIDストライプのデータ全体を読み込み、保全性のためにチェックサムをベリファイします。
- 毎週、自動的にスクラブ処理が行われ、ディスクに格納された全てのデータをリベリファイし、エラーを検知した際には修復処理を行います。これにより、コールドデータによるエラーを修復することができます。

SSDに格納されたデータにおいては、定期的にフルRAIDストライプのデータを読み込むためにバックグラウンドでスクラブ処理が継続的に行われ、算出されたチェックサムを比較します。エラーが検知されると、RAIDはリアルタイムにエラーを修復します。RAIDストライプ内のデータオブジェクト毎のチェックサムも個別に算出され、SSDから抽出された値と照合されます。図4でRAIDを利用したリアルタイムのエラー検知と修復のフローを図解しています。

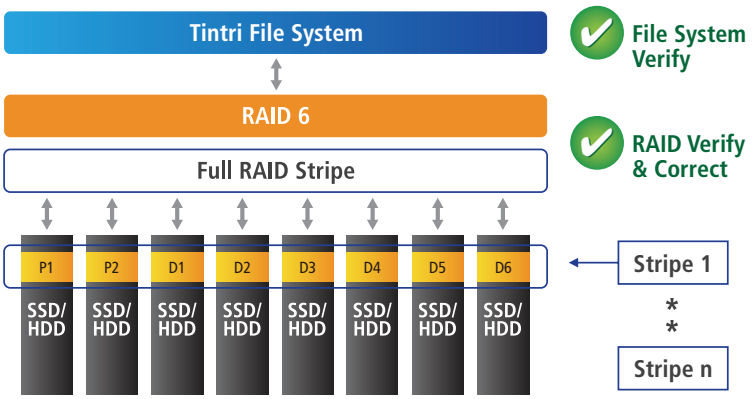

図 4：RAID 6 を使用したエラー検知と修復の継続的なスクラブ処理

RAID 6 のリアルタイムなエラー修復と継続的にスケジュールされたデータ処理により、ストレージメディアが生み出すエラーのほとんどは、ファイルシステムまたはストレージシステムの操作に影響を与えることなく検知され修復されます。

参照整合性と復元性

Tintri のファイルシステムは、SSD と HDD にデータを格納します。データの属性情報が記述されているメタデータは SSD 上（ページプールに整理されたページ毎）に格納されます。全てのオブジェクト（データブロックまたはデータとメタデータページ）は、チェックサムとディスクリプター（オブジェクトの自己記述モジュール）を保持しています。データのディスクリプターはオブジェクトが属するファイルとオフセットを示し、同様にメタデータページのディスクリプターは、メタデータオブジェクトが属するページプールを示します。Tintri のファイルシステムは、オブジェクトとそのディスクリプターを紐付けるチェックサムを格納するため、欠損したライト処理、誤った場所のリード処理、またはその他のエラーによってデータが壊れることはありません。データとメタデータの自己記述により、ディスクやファームウェアの誤作動の発生を抑えます。図 5 で Tintri のファイルシステムの自己記述構造を図解しています。

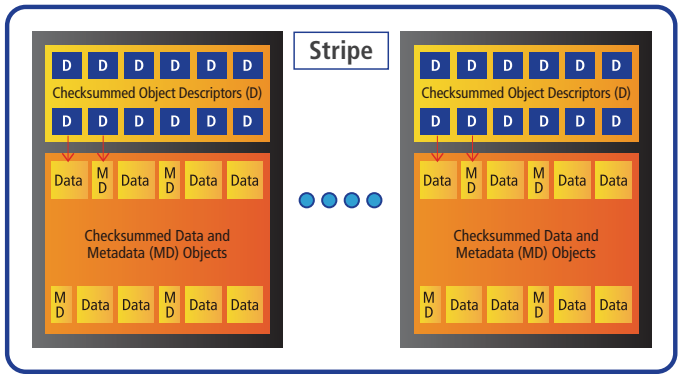

図 5：Tintri のファイルシステムの自己記述構造

Tintri のファイルシステムは、最も低い階層にあるブロックやデータを利用した参照階層で構成され、メタデータがそれらの情報を上位の階層にマッピングしています。また、エラーを検知するために強力なチェックサムを利用して各階層の参照整合性が維持されます。そのため、チェックサムは誤ったデータブロックに向けられたファイルのようなエイリアス問題を防ぎます。さらには、メタデータオブジェクトはメタデータページにバージョンナンバーを保持し、同様のエイリアス問題を防ぎます。

前述したように、データブロックやデータはフル RAID ストライプユニット毎に SSD や HDD に書き込まれます。自動修復ファイルシステムの技術により、コールドデータのエラーが検知され修復されます。それとは異なるケースが生じ、修復不可能なディスクエラーが孤立または欠損したオブジェクトを発生させた場合には、自己記述オブジェクトのスキャンが問題を検知し修復処理を行います。

高可用性

Tintri VMstoreアプライアンスは、デュアルコントローラのアーキテクチャーを有し、高可用性を実現しています。Tintri OSは、高可用性のアーキテクチャーによる複数のデータ保全の機能を提供します。まず、Tintriのファイルシステムは、アクティブコントローラ上のNVRAMとスタンバイコントローラのNVRAMを同期し、ファイルシステムのコンテンツが合致しているかを確認します。データが同期されるとプライマリコントローラ上のNVRAMのコンテンツにおいて強力なフィンガープリントが算出され、セカンダリコントローラによって個別に算出されたフィンガープリントとベリファイされます。

Tintri OSがライト処理を受信すると、データはプライマリコントローラ上のNVRAMデバイス内にバッファされ、セカンダリコントローラ上のNVRAMデバイスに転送されます。データが安全にNVRAMに格納されたという確認がセカンダリコントローラに届くと、プライマリコントローラはライト処理が行われたことをハイパーバイザーに伝達します。このように、高速なバッファリングと障害防止のためのNVRAM技術を組み合わせることで、コントローラの障害がデータ保全性に影響を与えることはありません。

まとめ

現代のストレージシステムは、データの保全性を保証するために異なる階層（ハードウェア、オペレーティングシステム、ファイルシステム）において複数の技術を搭載していなければなりません。複数の技術を組み合わせた堅牢な構成により、ストレージシステムの様々なコンポーネントで起こり得るエラーを防止します。従来のストレージシステムとは異なり、VMstoreアプライアンスはデータ保全性を設計主旨とし、VMに特化して開発されました。Tintri OSは、データの保全性を提供するために多数の技術を搭載し、こういった技術の組み合わせは汎用的なその他のプライマリストレージシステムでは実現されていません。

VMに特化して開発されたという点は、サポートできるハイパーバイザーの種類を限定してしまいますが、念入りな証明により、環境の相互作用から発生するエラーを削減します。エラーを回避する開発概念と堅牢な導入により、まずはソフトウェアのエラーが発生する機会を最小限に抑えます。専用に開発されたファイルシステムとフルストライプ書き込みにより、データは常に守られ、ソフトウェアであってもその他のコンポーネントであっても潜在的なエラーの発生を防止します。

SSDとHDDに書き込まれた全てのデータとメタデータオブジェクトはチェックサムにより守られ、全ての書き込み処理においてベリファイされます。データの欠損はRAID 6を利用して即座に検知され、修復されます。さらに、メタデータを参照することで整合性を確認し、メタデータを欠損させ得るソフトウェアエラーを検知します。

SSDとHDDにおけるTintri独自のRAID 6の搭載により、二重ドライブ障害を防止し、データのリードエラーがあっても失敗したドライブを再構築し、リード処理中にリアルタイムにエラーを修復します。継続的なスクラブ処理が潜在的なエラーを能動的に検知し、自動修復します。

発売元

ノックス株式会社
www.nox.co.jp

本　　社 〒152-0023 東京都目黒区八雲2-23-13　Tel. 03-5731-5551　Fax. 03-5731-5552
西日本支社 〒533-0033 大阪市東淀川区東中島1-17-5 Tel. 06-4809-5544　Fax. 06-4809-5547

●本製品に関するお問い合わせ：**営業本部**
●メールでのお問い合わせ：tintri@nox.co.jp

お問い合わせ先

201209-A