特集

Oracle Solarisの仮想化技術

Virtualization Technology of Oracle Solaris

あらまし

ICTインフラは，システム最適化に向けたコスト削減が急務となっており，仮想化技術を用いたシステムリソースの有効活用が大きく注目されている。Oracle Solaris Operating Systemでは，サーバを仮想化するSolarisコンテナおよびOracle VM Server for SPARCと，ストレージを仮想化するSolaris ZFSを標準提供している。Solaris ZFSは，高信頼ファイルシステムであり，ストレージプール技術やスナップショット/クローン技術により，システムのディスク使用を効率化する。

本稿では，Oracle VM Server for SPARCのドメイン管理，リソース管理，ドメインマイグレーションと，Solarisコンテナのゾーン管理，リソース管理，ゾーンマイグレーションについて触れる。さらに，Solaris ZFSのZFSストレージプールおよびZFSファイルシステムについて紹介する。

Abstract

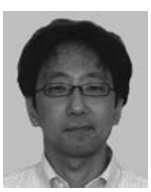
There are urgent needs to reduce costs aiming for system optimization in ICT infrastructure. To this end, there are high expectations for virtualization technology. By default, the Oracle Solaris Operating System offers Solaris Containers and Oracle VM Server for SPARC, which realize server virtualization, and Solaris ZFS which realizes storage virtualization. Solaris ZFS is a highly reliable file system that enables efficient disk use by storage pool technology and snapshot and clone technology. This paper covers domain management, resource management, domain migration for the Oracle VM Server for SPARC, and zone management, resource management, and zone migration for Solaris Containers. Moreover, it introduces the ZFS storage pool and ZFS file system for Solaris ZFS.

小澤彰利（おざわ　あきとし）

エンタプライズサーバ事業本部
UNIXソフトウェア開発統括部 所属
現在，仮想化ソフトウェアの企画・開発に従事。

鈴木克之（すずき　かつゆき）

エンタプライズサーバ事業本部
UNIXソフトウェア開発統括部 所属
現在，仮想化ソフトウェアの企画・開発に従事。



垰田雅己（たおだ　まさみ）

エンタプライズサーバ事業本部
UNIXソフトウェア開発統括部 所属
現在，仮想化ソフトウェアの企画・開発に従事。

ま え が き

ICT（Information and Communication Technology）インフラは，業務効率化とスピードアップを図るため，システムの規模を拡大してきた。しかしその反面で，サーバ台数の増加に伴い，運用・管理コストや消費電力の削減が課題となっている。

サーバ仮想化は，CPUやメモリなどのシステムリソースを稼働状況に応じて効率的に割り当てることで，初期投資を抑えて多数のサーバを無駄なく統合する技術である。ハードウェアの調達を待たずに，仮想サーバは即座に作り出せ，不要になったら簡単に破棄できる。サーバ仮想化は，システム最適化への第一歩となる。

富士通のUNIXサーバSPARC Enterpriseに搭載している Oracle Solaris Operating System（Solaris OS）[1]は，アップデートリリースごとに仮想化機能を強化している。

本稿では，SPARC Enterpriseで標準提供している仮想化機能のうち，サーバを仮想化[2],[3]するOracle VM Server for SPARCとSolarisコンテナを紹介し，さらにストレージを仮想化[4],[5]するSolaris ZFS（Solaris Zettabyte File System）を紹介する。

SPARC Enterpriseの仮想化機能

SPARC Enterpriseでは，仮想サーバに分割する階層により，三つのサーバ仮想化機能を標準提供している（**図-1**）。

● **ハードウェアパーティション**

物理システムボードを論理的に分割することで，独立したSolaris OSが動作するパーティションを構成できる。CPUやメモリは，業務拡張や新規業務の追加などの要求に応じて，システム運用を停止することなく動的に変更できる。

● Oracle VM Server for SPARC

Oracle VM Server for SPARC（旧名称：Sun Logical Domains）[6]は，物理サーバをファームウェア層のSPARCハイパーバイザにより仮想サーバに分割することで，独立したSolaris OSが動作する論理ドメインを構成できる。CPU，メモリ，I/Oデバイスは，Domain Managerで柔軟に割り当てられる。

● Solarisコンテナ

Solarisコンテナは，Solaris OSを仮想的に分割することで，独立した仮想OS環境であるゾーンを構築できる。CPUやメモリは，ゾーンの稼働状況に応じて柔軟に配分される。I/Oデバイスは，ゾーン構成時に割り当てる。

さらに，Solaris OSでは，ストレージ仮想化機能としてSolaris ZFSを標準提供している。ZFSファイルシステムは，複数の物理ディスクをストレージプールで管理する。ストレージプールから必要な領域を切り出すことで，仮想化されたボリュームを作成できる。ZFSは，堅ろう性と拡張性を実現しながら容易な管理を可能とするファイルシステムである。

以降では，Solaris OSの仮想化機能について説明する。

Oracle VM Server for SPARC

論理ドメインは，CPU，メモリおよびI/Oデバイスを論理的にグループ化した仮想サーバで，論理ド

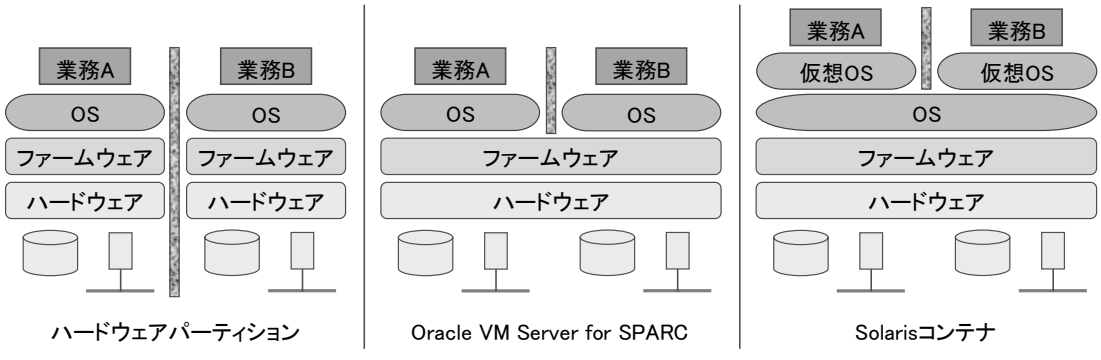


図-1　SPARC Enterpriseの仮想化機能
Fig.1-Virtualization feature on SPARC Enterprise.

メインごとに独立したSolaris OSが動作する。各論理ドメインは，個々に起動・停止でき，最大128ドメインを作成できる。論理ドメインは，ハイパーバイザの論理ドメインチャネル（LDC：Logical Domain Channel）で相互に通信する。ディスクやネットワークなどの仮想デバイスは，LDCで仮想サービスと通信して，物理デバイスにアクセスする。

● **論理ドメインの役割**

論理ドメインは，つぎの役割を持っている（**図-2**）。

(1) 制御ドメイン

制御ドメインは，Domain Managerが動作するドメインで，ほかの論理ドメインの作成，管理，仮想リソースの割当てを行う。また，仮想ディスクサーバや仮想スイッチなどの仮想デバイスサービスを提供する。

(2) I/Oドメイン

I/Oドメインは，物理I/Oデバイスに直接アクセスできる。また，制御ドメインと同様に，仮想デバイスサービスを提供できる。

(3) ゲストドメイン

ゲストドメインは，制御ドメインまたはI/Oドメインから仮想デバイスサービスを受け，仮想ディスクや仮想ネットワークなどの仮想I/Oデバイスを使用できる。

● **リソース管理**

仮想CPU，メモリ，仮想I/Oデバイスなどのリソースは，論理ドメインが動作中であっても動的に再構成できる。さらに，論理ドメインの仮想CPUを，動的資源管理ポリシーに従って自動的に増減させられる。動的資源管理ポリシーは，使用数，使用率，上限，下限，期間などを組み合わせて作成する。

また，CPU電源管理では，業務負荷が変化して未使用となったCPUの電源を切断することで消費電力を抑えられる。

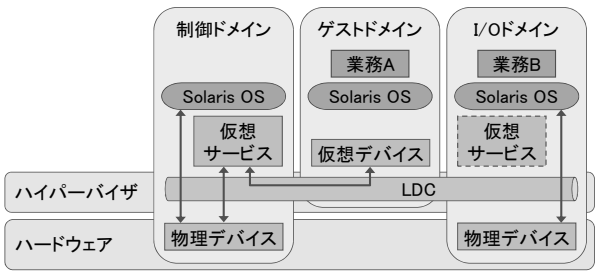


図-2　論理ドメインの役割
Fig.2-Roles for logical domains.

● **ドメインの複製**

新規にドメインを作成する場合，既存のドメインを複製することで，利用者の要求に応じて仮想サーバを迅速に提供（プロビジョニング）できる。ZFSファイルシステムにゲストドメインの起動ディスクイメージを格納すれば，後述のZFSクローン（clone）で瞬時に複製できる。複製した起動ディスクイメージを，別ゲストドメインに割り当てることで，Solaris OSのインストール作業が不要となる。

● **マイグレーション**

業務負荷の変動や段階的なサーバ増設などに伴い，ゲストドメインを別の物理サーバへ移行できる。ゲストドメインを一時停止させ，メモリ内容を圧縮して高速転送する。

また，サーバを統合する場合は，Physical-to-Virtual（P2V）移行ツールで物理サーバを論理ドメインに移行できる。物理サーバの構成情報を収集して，ファイルシステムイメージを作成する。収集した構成情報から論理ドメインを作成して，ファイルシステムイメージを仮想ディスクに復元する。

Solarisコンテナ

Solarisコンテナは，一つのOS空間を仮想的に分割して，複数のOSが動作しているように見せるSolarisゾーン（Solaris Zone）機能と，CPUやメモリなどのハードウェアリソースを柔軟に配分するSolarisリソースマネージャ（Solaris Resource Manager）機能で構成される。

Solarisゾーンは，仮想化されたOS環境で，アプリケーションの実行に適した安全で隔離された環境を実現する。ゾーンごとに実行されるプロセスは分離されるので，ほかのゾーンに影響を及ぼさない。

● **Solarisゾーン**

global zoneは，Solarisシステムに一つ存在するゾーンで，システム全体を管理する（**図-3**）。non-global zoneの作成，管理，物理I/Oデバイスの割当てなどはglobal zoneでのみ行える。

non-global zoneは，仮想Solaris環境のソフトウェアパーティションで，ほかのゾーンに影響することなくアプリケーションを実行できる。最大8191個のゾーンを作成でき，各ゾーンは許可されたファイルシステムおよび許可された物理I/Oデバ

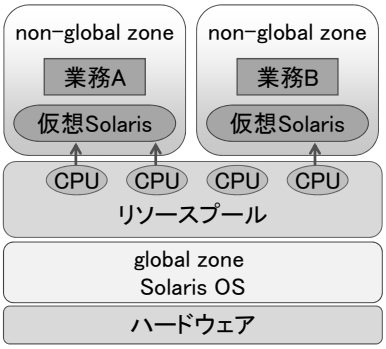


図-3　Solarisコンテナの構成
Fig.3-Solaris Containers' configuration.

イスのみ使用できる。

non-global zoneを構成するシステムファイルは，ゾーン作成時にglobal zoneからコピーされる。global zoneへのパッチ適用によりこれらのファイルが更新されると，すべてのnon-global zoneのファイルも同期して更新される。

● Solarisリソースマネージャ

Solarisリソースマネージャは，リソースの使用状況を定期的に監視して，ゾーンを停止せずに設定されたリソース量の範囲内で自動的にリソースを割り当てる。ゾーンに割り当てるリソースは，CPUを制御するリソースプールとメモリを制御する資源上限デーモンで管理する。さらに，リソースプールは，CPUをグループ化するプロセッサセットとCPUの配分を制御するスケジューリングクラス（タイムシェアリング，公平配分）で構成される。

● ゾーンの複製

既存のゾーンをコピーして新規のゾーンを簡単に作成できる。ゾーンがZFSファイルシステム上にある場合は，後述のZFSクローンを使用することで，ゾーンの複製を瞬時に行え，使用するディスク容量を節約できる。

● マイグレーション

CPUやメモリなどのリソースが不足する場合やゾーンの組合せを変更する場合は，ゾーンを別サーバに移行できる。移行元サーバでゾーンを切り離し（detach），移行先サーバに組み込む（attach）。両方のサーバでパッケージ構成などの環境が異なっても，組込み時にゾーンを構成するシステムファイルが移行先サーバのglobal zoneと同期される。

さらに，稼働中のSolaris10 OSシステムは，P2V機能で，フラッシュアーカイブをゾーンに展開することで移行できる。また，Solaris8/9 OSシステムは，Solaris 8/9 Containersで，アプリケーションを変更せずにSolaris10 OSのゾーンに移行できる。

Solaris ZFS

Solaris ZFSは，128ビットのファイルシステムで，実用上は無限大の容量を管理できる。ZFSファイルシステムを管理するメタデータは，必要に応じて動的に割り当てられるため，ファイルシステム数やファイル数の制限はない。従来のファイルシステムでは，ファイルシステムのサイズが物理デバイスのサイズに制限されていた。

しかし，Solaris ZFSでは，ZFSストレージプールで物理デバイスを隠ぺいするため，特定の物理デバイスに制限されない。ZFSファイルシステムは，初期化せずにファイルシステムの階層を簡単に作成でき，ZFSストレージプールに割り当てられたディスク容量の範囲で自動的に拡張する。

● ZFSストレージプール

ZFSストレージプールは，物理ディスクを集約して管理する仕組みで，非冗長，ミラー，RAID-Z（シングルパリティ），RAID-Z2（ダブルパリティ），RAID-Z3（トリプルパリティ）の冗長構成を選択できる。

ストレージプールへのデータ書込みは，使用可能なすべてのデバイス上にデータが動的にストライプ化される。また，冗長構成のストレージプールでは，不正なデータブロックを検出すると，別の冗長コピーから正しいデータを取得して自己修復する。

● ZFSストレージプールの管理

ストレージプールは，オンラインのままディスク増設・交換ができるので，初期導入時に将来使用分のディスク容量を確保する必要がない（**図-4**）。

(1)　接続・切離し

既存のミラー構成または非ミラー構成のストレージプールにディスクを接続（attach）することでミラー化できる。ディスクを追加すると，すぐに再同期化が開始される。ミラー化されたストレージプールからディスクを切り離す（detach）ことでミラー構成を変更できる。また，個別のディスクをオフラインにすると一時的にディスクを切断できる。オンラインにするとデータが再同期化される。

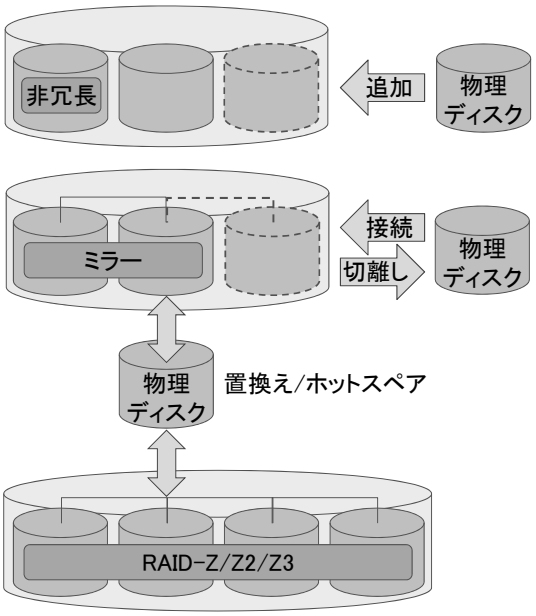


図-4　ZFSストレージプールの構成
Fig.4-ZFS Storage Pool's configuration.

(2)　置換え

冗長構成のストレージプールでは，動的にディスクの置換え（replace）ができる。ホットスペアのディスクを用意すれば，ストレージプールで障害が発生したディスクまたはエラー状態のディスクを自動的に置き換える。ホットスペアのディスクは，ストレージプール間で共有できる。

(3)　移行

ストレージプールは，他サーバに移行できる。ストレージプールをエクスポート（export）すると，未処理のデータがすべてディスクに書き込まれる。ストレージプールを構成するディスクをすべて移行先サーバに取り付けて，ストレージプールをインポート（import）するだけで使用可能になる。

● ZFSストレージプールの性能向上

性能向上と消費電力の削減を実現するため，メモリ，SSD（Solid State Drive），ディスクを組み合わせたZFSハイブリッドストレージプールを構成できる。

ZFSインテントログ（ZIL : ZFS Intent Log）は，通常ストレージプール内の領域を使用するが，SSDなどの高速デバイスを，個別のログデバイスに割り当てることで，同期書込みの性能を向上させる。

キャッシュデバイス（L2ARC : Level 2 Adaptive Replacement Cache）は，メモリとディスクの間にSSDなどの高速デバイスをキャッシュとして追加することで，静的データのランダム読込み性能を向上させる。

● ZFSファイルシステム

ZFSファイルシステムは，トランザクションファイルシステムである。データの書込みは，既存のデータを上書きせずに，元データをコピーして更新する（Copy-on-Write）。一連のデータ更新処理が完了した後で新旧データへのポインタを切り替えるため，ファイルシステムの整合性が常に保たれる。サーバの電源が突然切断されても，ファイルシステムが破壊されることがない。

ファイルシステムを作成するとマウントポイントを自動生成してマウントする。サーバ起動時には自動的にマウントされるので，マウントの管理が不要となる。ZFSボリュームは，ブロックデバイスとして識別される。ボリューム作成時に，初期サイズの領域が予約され，使用量に応じて自動拡張される。

● ZFSファイルシステムの信頼性

ZFSファイルシステムは，すべてのユーザデータおよび管理情報のメタデータについてEnd-to-Endチェックサムを持っている。データブロックのチェックサムは親ブロックに保持されており，トップの管理ブロック（Uber-block）まで順に続くので，データツリー全体の自己検証ができる（**図-5**）。異常を検出すると，冗長コピーからデータを回復する。ZFSファイルシステムのメタデータは，信頼性を高めるため，異なるディスクにまたがって何度か自動的に保存される（dittoブロック）。

さらに，ZFSファイルシステムでは，ユーザデータの複数コピーを保存することもできる。複数ディスクで冗長化できない場合でも，ディスクブロックの読取り障害から回復できる。

● ZFSスナップショットとZFSクローン

ZFSスナップショット（snapshot）は，ファイルシステムまたはボリュームの読取り専用コピーで，ディスク領域を消費せず瞬時に作成される。スナップショットは直接参照できないが，ロールバック（rollback），クローン，バックアップなどができる（**図-6**）。元のファイルシステムまたは元のボリュームを更新すると，データ変更した分だけディスク領域が消費される。

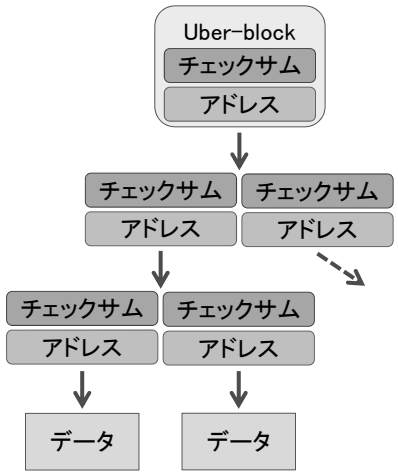


図-5　ZFSのEnd-to-Endチェックサム
Fig.5-ZFS' end-to-end checksum.

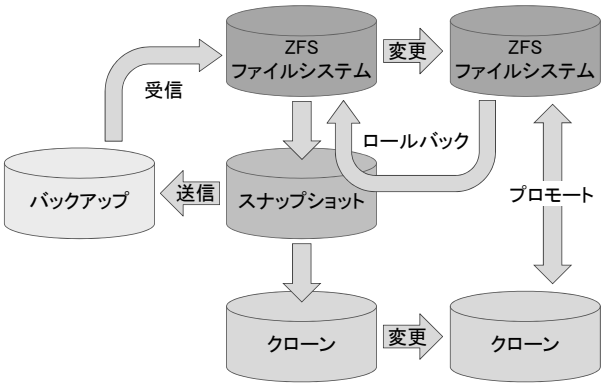


図-6　ZFSスナップショットとZFSクローン
Fig.6-Snapshots of ZFS and ZFS clones.

(1)　ロールバック

ロールバックすると，あるスナップショットを作成した後のデータ変更を破棄して，ファイルシステムをスナップショットが作成された時点の状態に戻せる。

(2)　クローン

ZFSクローンは，ファイルシステムまたはボリュームの書込み可能なコピーで，スナップショットから作成できる。また，スナップショットと同様に，ディスク領域を消費せず瞬時に作成され，データ変更した分だけディスク領域を消費する。プロモート（promote）すれば，ファイルシステムをそのファイルシステムのクローンで置き換えられる。

ZFSクローンを使用すると，ゾーンの複製，ゲストドメインの仮想ディスク複製およびSolaris Live Upgradeによるブート環境複製が容易である。

(3)　バックアップ

ファイルシステムをバックアップするには，スナップショットからストリームを作成して送信（send）する。このストリームを受信（receive）して，別ストレージプールでファイルシステムを作成する。変更分のデータだけ，増分スナップショットとして送信できる。また，別ストレージプールに展開せず，任意のアーカイブ形式で保存することもできる。

むすび

本稿では，Solaris OSの仮想化技術として，Oracle VM Server for SPARCおよびSolarisコンテナによるサーバ仮想化と，Solaris ZFSによるストレージ仮想化について紹介した。

Solaris OSは，仮想化機能の更なる進化を目指し，システム最適化に貢献する。[7]

参考文献

(1)　富士通：Solaris 10オペレーティングシステム. http://primeserver.fujitsu.com/unix/soft/opt/os-10/

(2)　湯原雅信：仮想マシン技術の展望. *FUJITSU*, Vol.60，No.3，p.221-227（2009）.

(3)　小口芳彦ほか：サーバ仮想化技術とその最新動向. *FUJITSU*, Vol.58，No.5，p.425-430（2007）.

(4)　熊沢忠志：ストレージ仮想化の展望. *FUJITSU*, Vol.60，No.3，p.241-246（2009）.

(5)　赤坂 勉ほか：ETERNUSディスクアレイの仮想化技術. *FUJITSU*, Vol.60，No.3，p.253-257（2009）.

(6)　富士通：Logical Domains Manager Software. http://primeserver.fujitsu.com/sparcenterprise/download/software/ldoms/

(7)　富士通：技術情報 Solaris Technical Park. http://primeserver.fujitsu.com/sparcenterprise/technical/