THE POSSIBILITIES ARE INFINITE FUJITSU

PRIMEQUEST 仮想マシン機能

ホワイトペーパー

２００９年　１２月

富士通株式会社

本書には、今後提供する予定の機能が含まれています。
本書に記載されている内容は、予告なく変更される場合があります。

目次

1. はじめに

ビジネスを取り巻く環境はますます厳しくなってきており、ビジネス環境の変化へ迅速に対応する「機敏性」、不要なコストを削減する「効率性」、365 日 24 時間ビジネスを提供する「継続性」が求められています。富士通では、お客様のビジネスの「機敏性」「効率性」「継続性」を支援するため、TRIOLE という IT 基盤のフレームワークを提唱し、仮想・自律・統合の 3 つの技術を軸に最先端の技術開発を続けております。このホワイトペーパーでは、基幹 IA サーバ PRIMEQUEST において仮想化を実現する PRIMEQUEST 仮想マシン機能について説明します。

2. サーバ仮想化とは

仮想化は、IT システムにおいてよく使われる考え方です。仮想化は、２つの層の固定的な依存関係を断ち切ることにより、新たな自由度を獲得するための技術です。たとえばストレージ仮想化では、サーバとストレージの固定的な結びつきを断ち切り、配線を変更せずにサーバとストレージの割り当て関係を自由に変更することが可能になります。ネットワーク仮想化では、サーバとネットワークの固定的な結びつきを断ち切り、物理的なネットワークを変更せずに、仮想的なネットワークを用意することが可能になります。

本書で扱うサーバの仮想化についても、同様のことが言えます。つまり、サーバの仮想化は、OS 層と物理マシン層（物理的に存在するサーバ）との１対１の結びつきを解放します（図 1）。これにより、１台の物理マシンの上で複数のマシン（仮想マシン＝Virtual Machine）を動作させたり、ある物理マシンの上で動作している仮想マシンを別の物理マシン上に移動させたりする自由度が得られます。

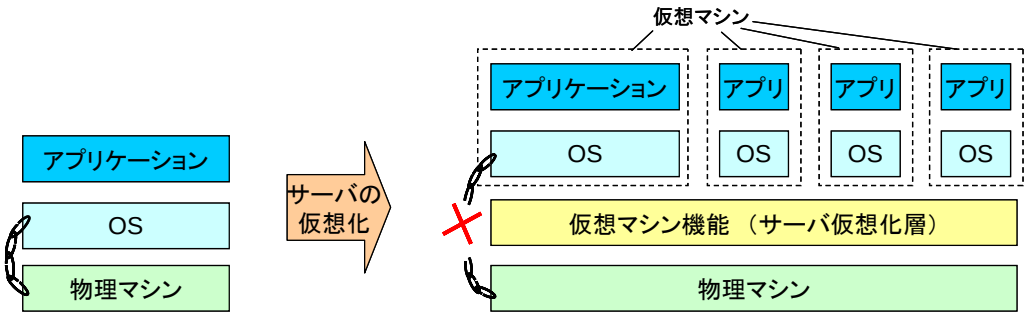


図 1. サーバの仮想化

最近サーバ仮想化が注目を集めています。その背景には、ビジネスの変化のスピードが速くなったり、企業内に散在するサーバの管理コストが増大したりしていることがあります。技術面でも、CPU のマルチコア化などによりサーバの性能が向上したり、インテルなどが提供する CPU に仮想化支援ハードウェアが導入されたり、オープンソースの仮想化ソ

フトウェアXen(ゼン)が登場するなど、オープンサーバのサーバ仮想化を推し進める素材がそろってきました。

オープンサーバの仮想化が最近注目を集めていますが、仮想マシンの概念は以前から存在しました。富士通は、1980 年にメインフレームコンピュータ向けの仮想マシン機能 AVM の提供を開始して以来、現在の AVM/EX にいたるまで、長年にわたり仮想マシンに関する技術を蓄積してまいりました。メインフレームでは仮想マシンは確立された技術であり、富士通の大型メインフレームでは、お客様の８割以上に仮想マシン機能をご利用いただいています。

また、富士通は Xen のオープンソースコミュニティに参加し、その開発に貢献してきました（図 2)。今後も引き続き、仮想マシンに関する富士通の豊富な経験を Xen の開発に活かしていく予定です。

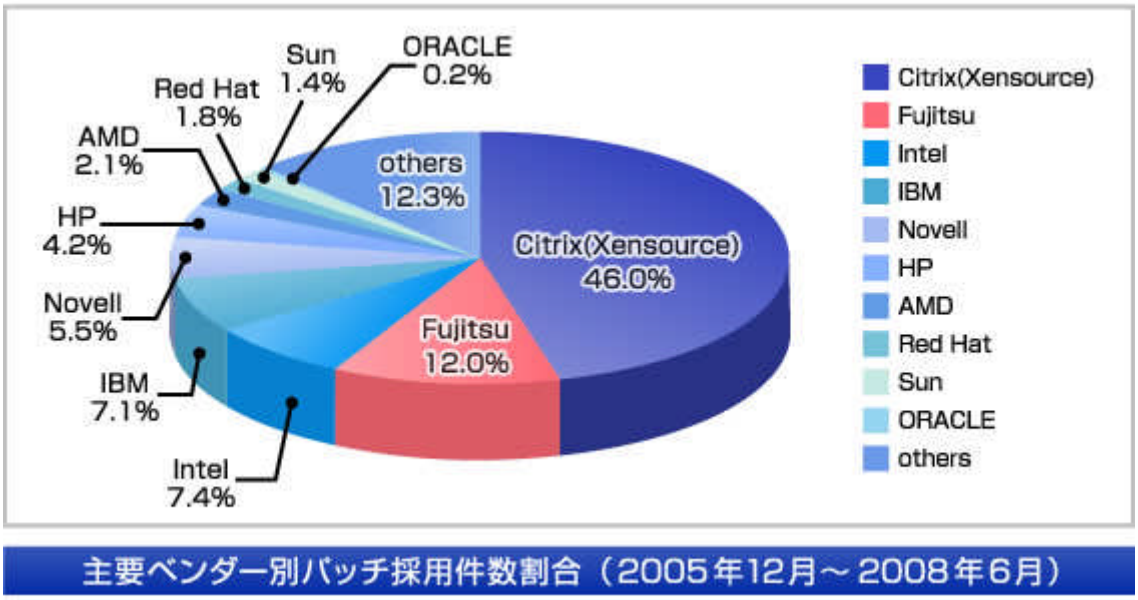


図 2. Xen コミュニティへの貢献（拡張・修正の投稿数、当社集計）

3. PRIMEQUEST 仮想マシン機能

PRIMEQUEST 仮想マシン機能（以降では、単に仮想マシン機能と記述することもあります）は、富士通の基幹 IA サーバ PRIMEQUEST 上でサーバ仮想化を実現します。PRIMEQUEST 仮想マシン機能により、高性能で高信頼な PRIMEQUEST のシステム資源を、柔軟かつ効率的に活用することができます。提供形態としては、Red Hat Enterprise Linux 5 に含まれる仮想化ソフトウェアと、それに対する富士通のサポートという形をとります。サポートには、性能や信頼性の向上、インストール支援のために提供する当社独自の付加ソフトウェアを含みます。PRIMEQUEST 仮想マシン機能は、PRIMEQUEST 500

シリーズ以降で利用可能です。

3.1 PRIMEQUEST 仮想マシン機能の特長

PRIMEQUEST 仮想マシン機能の特長は以下のとおりです。

■ Linux、Windows の複数 OS の同時動作

最大 60 個の仮想マシンを起動し、各仮想マシンでそれぞれ独立した OS（ゲスト OS）を動作させることができます。

■ 各 OS への柔軟できめ細かなシステム資源配分

PRIMEQUEST がハードウェアで提供しているパーティション機能である、PPAR(Physical PARtitioning)や XPAR(eXtended PARtitioning)と比べて、CPU やメモリ量、I/O 装置などのシステム資源を、細かい粒度で配分できます。配分量の変更も簡単に行えます。

■ 複数 OS からの FC／LAN カード共有

FC（Fibre Channel）カード、LAN カード、ケーブルなどを各 OS から共有することで、効率的で無駄のないシステム構築ができます。

3.2 統合運用管理の実現

PRIMEQUEST 仮想マシン機能を、富士通の統合運用管理ソフトウェア Systemwalker などと組み合わせることで、物理マシンと仮想マシンを一元的に運用管理することができます。

■ 自動運用

- Systemwalker Operation Manager と組み合わせることで、仮想マシンおよびゲスト OS の起動、停止、資源配分の変更をスケジュールにしたがって自動化することができます。

■ 監視

- Systemwalker Centric Manager と組み合わせることで、物理マシン上の仮想マシンおよびゲスト OS を視覚的に監視し、障害時には障害発生場所を明確に把握することができます。
- Systemwalker Service Quality Coordinator と組み合わせることで、仮想マシンやゲスト OS、ミドルウェアの性能を監視したり、各仮想マシンに配分された資源の利用状況を可視化したりすることができます。

■ 高速バックアップ

- 富士通のストレージシステム ETERNUS のディスクアレイ装置と、高速バックアップソフトウェア ETERNUS SF AdvancedCopy Manager を利用することにより、Disk-to-Disk の高速バックアップや、テープも含めた Disk-to-Disk-to-Tape の統合バックアップが行えます。

3.3 高可用性システムの構築

富士通では、PRIMEQUEST 仮想マシン機能を用いたシステムに対してさまざまな冗長化機能を提供し、システムの可用性を高めます。

■ PRIMEQUEST のハードウェアの冗長化

- システムミラー機構により、メモリやチップセットなど主要ハードウェアを二重化し、メインフレームクラスの信頼性を実現することができます。
- フローティング・システムボード（予備のシステムボード）により、万一システムボードが故障した場合でも、仮想マシン機能の再起動のみでシステムを短時間に復旧することが可能になります。

■ ディスクとネットワークの冗長化

- PRIMECLUSTER GDS により、ディスク装置のミラーリングが可能になります。
- ETERNUS マルチパスドライバにより、OS と ETERNUS ディスクアレイ装置の間のファイバチャネル接続パスを冗長化することができます。
- PRIMECLUSTER GLS により、OS と通信相手との間の伝送路を冗長化することができます。

以上の冗長化は、管理 OS において冗長化することも、ゲスト OS(Linux)において冗長化することも可能です。共有デバイスを管理 OS において冗長化すれば、各ゲスト OS では冗長化のための考慮が不要になり、運用が容易になります。

■ サーバの冗長化

- PRIMECLUSTER により、仮想マシン上のサーバ（ゲスト Linux）、物理マシン上のサーバ（ネイティブ Linux）にかかわらず、複数のサーバをあたかも 1 つのシステムのように運用し、サーバがダウンしても、残りのサーバで業務を継続することができます。

3.4 諸元

サポートハードウェア			PRIMEQUEST 500シリーズ以降
最大仮想マシン(ゲストOS)数			60
最大サポート物理CPU数			32CPU(64コア)
最大サポートメモリ量			各機種の最大搭載メモリ量 PRIMEQUEST520:256GB(2SB*128GB) PRIMEQUEST540:1TB(4SB*256GB) PRIMEQUEST580:2TB(8SB*256GB)
管理OSのメモリ量			1024MB以上を推奨
ハイパーバイザのメモリ量			64MB
サポートする仮想化方式			完全仮想化方式
サポートするゲストOS			•Red Hat Enterprise Linux 5.1 (for Intel Itanium) 以降 •Red Hat Enterprise Linux AS (v.4 for Itanium) Update4 •Red Hat Enterprise Linux AS (v.4 for Itanium) 4.5以降 •Microsoft Windows Server 2003, Enterprise Edition for Itanium-based Systems SP1以降 •Microsoft Windows Server 2003, Datacenter Edition for Itanium-base SP1以降 •Microsoft Windows Server 2008 for Itanium-Based Systems
仮想ネットワーク	仮想ネットワーク形態		仮想ブリッジ接続
	最大仮想ブリッジ数		128(ただし、外部に接続する仮想ブリッジ数は、搭載物理NIC数以下)
	仮想ブリッジに接続できる最大VNIF(*1)数		64
ディスク	サポートするブロックデバイス		ディスク、パーティション、GDS論理ボリューム、LVM論理ボリューム、 イメージファイル形式(*2)
接続可能周辺装置			RHEL5がサポートする周辺装置
可用性	クラスタ構成		PRIMECLUSTERによるゲストOSのクラスタ構成を実現(Linuxゲストのみ)
	冗長化(*3)	ディスクパス	ETERNUS マルチパスドライバにより可能
		ディスク装置	PRIMECLUSTER GDSにより可能
		LAN	PRIMECLUSTER GLS(*3)またはBondingにより可能
主な制限事項			Hyper-Threading(ハイパースレッディング・テクノロジー)は使用禁止

(*1)VNIF:仮想ネットワークインターフェース
(*2)イメージファイル形式:仮想マシン環境およびデータをファイルとして持つ形式
(*3)管理OSにおける冗長化。LinuxゲストOSにおける冗長化はPRIMECLUSTER GLSにより実現。

4. 仮想マシンの利用シーン

PRIMEQUEST 仮想マシン機能は、TRIOLE が掲げている、

効率性：コストの削減（ハードウェア、ソフトウェア、運用管理のコスト削減）

機敏性：時間の削減（変化へ即座の対応）

継続性：ダウン時間の削減（ソフトウェア障害やセキュリティへの対策、過負荷対策）

というメリットを多面的に提供します。PRIMEQUEST 仮想マシン機能の利用目的や利用形態に応じて、これらのメリットがいろいろな形で現れます。以下では、具体的な利用シーンを挙げて、それぞれの例でどのようなメリットが得られるかを説明します。

4.1 複数の業務システムの同時動作

効率性	機敏性	継続性
◎	―	―

１つの物理サーバを分割して複数の OS を動作させるのが、PRIMEQUEST 仮想マシン機能の最も基本的な使い方です。従来ならば複数のサーバを用いた場合でも、物理サーバを１台しか使わないことにより、サーバのハードウェアコストや運用管理コスト、消費電力

を削減します。これは、サーバハードウェアの性能向上と、ハードウェア性能を分割することのできる PRIMEQUEST 仮想マシン機能により可能となっています。PRIMEQUEST 仮想マシン機能を用いれば、複数の Windows と複数の Linux を 1 台のマシン上で同時に動作させることができます。

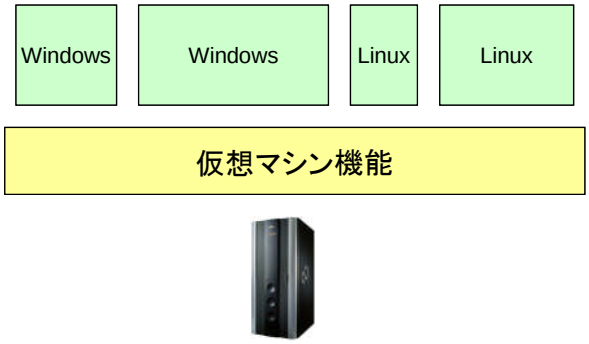


図 3. 1 台の物理サーバで複数 OS を動作

4.2 開発環境の提供や新 OS への移行

効率性	機敏性	継続性
◎	○	○

新しいアプリケーションを開発する場合には、いろいろな理由で複数の開発環境が必要になります。

- 開発チーム毎に、OS の設定を変える必要がある。
- 異なる版数、異なるパッチレベルの OS での開発が必要。
- 複数の独立したテストを同時に行う必要がある。
- 開発チーム毎に、自由なタイミングで OS を再起動したい。

PRIMEQUEST 仮想マシン機能を用いれば、複数の開発環境を必要なときに必要な規模で、短時間に低コストで用意することができます（図 4)。同一の物理マシン上に業務システムと開発環境を同居させることもできます。開発フェーズの進展に応じて、仮想マシンに割り当てる資源量を調整し、資源の利用効率を最適化することができます。

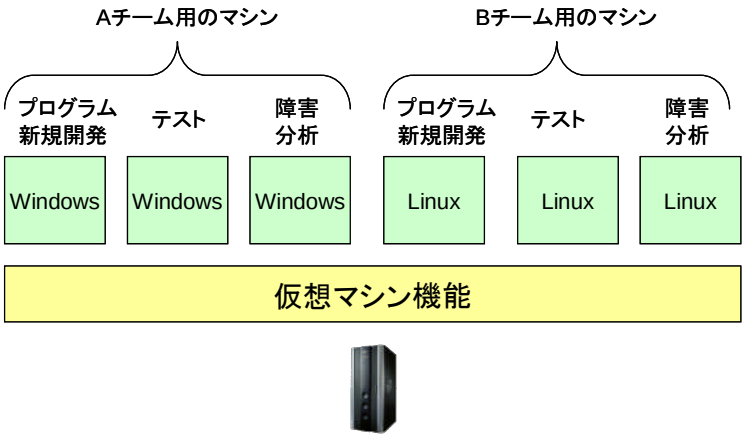
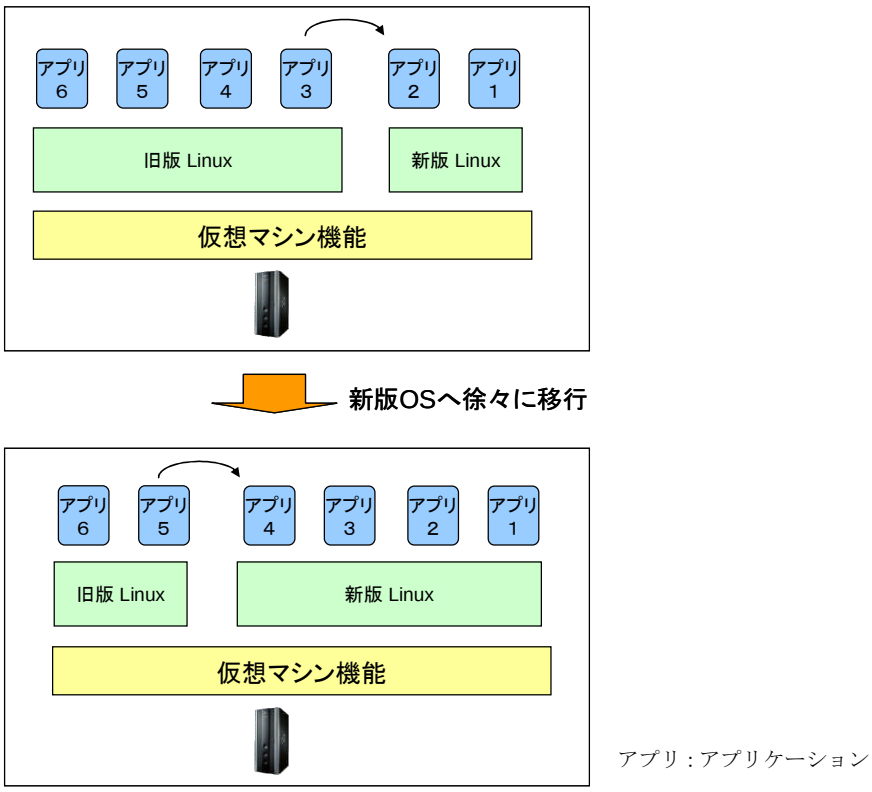


図 4. 開発環境の提供

旧 OS から新 OS へスムーズに移行させたいときにも、新 OS、旧 OS の両方の運用システムを、１つの物理マシンで動作させることができます。複数の業務アプリケーションのうち新 OS での検証が終わった業務から新 OS での運用に切り替えつつ、新 OS へ未対応の業務は旧 OS での運用を継続することができます。業務アプリケーションが新 OS に移行するのに応じて、新 OS の仮想マシンに割り当てる資源を増やして行くことで、資源を有効活用できます。



アプリ：アプリケーション

図 5. 新 OS へのスムーズな移行

4.3 新サーバの迅速な提供

効率性	機敏性	継続性
◎	◎	―

ビジネスの変化、組織変更などに対応するため、突発的に新しいサーバを立ち上げなければならないことがあります。しかし、ハードウェアの購入予算を確保するところから始めて、購入手続きを行い、ハードウェアの納入や現地調整を待っていると、実際にハードウェアが利用できるようになるまで数ヶ月かかることも珍しくありません。これではせっかくのビジネスチャンスを逃してしまいます。ハードウェアのサーバを手配する代わりに、PRIMEQUEST 仮想マシン機能を使って、ソフトウェアでサーバ（仮想マシン）を用意すれば、短時間にサーバマシンを起動できるようになります。状況にもよりますが、たとえば、数分から数十分でサーバを用意できます。従来の数ヶ月に比べれば、手配にかかる時間が数千分の１に短縮されるということです。

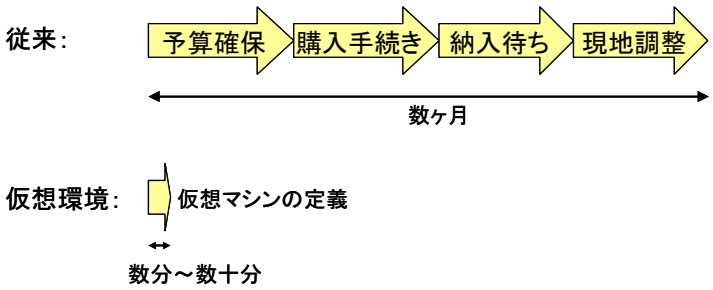


図 6. 物理／仮想サーバの手配

PRIMEQUEST 仮想マシン機能は、OS やアプリケーションのインストール作業の実質的な時間短縮にも貢献します。物理マシンとは違い、仮想マシンは簡単に手配できますので、仮想マシンを使ってあらかじめインストール作業を行うことができます。インストールの結果を仮想マシンイメージとして保存しておき、実際にサーバが必要になったときに、そのイメージを使って仮想マシンを再起動すれば、短時間にサーバの運用を開始することができます。仮想マシンを別の物理マシン上で運用する場合には、仮想マシン間の移行ツールを利用します。

同種のサーバが複数必要になる場合には、インストール済のイメージを原本として、仮想マシンイメージの複製（クローニング）を行います。クローニングツールを用いることにより、サーバ毎に手動でインストールする場合と比べて、時間を短縮できるだけでなく、インストール作業のミスを減らすこともできます。次のような用途には、特に有効です。

・負荷分散のため、同種のサーバを複数利用
・開発・テストのため、同種の環境を複数利用

・情報システムセンターから、各部門のサーバにソフトウェア環境を配布
・デモやトライアル、教育のためのソフトウェアスタックを配布

4.4 負荷に応じた動的資源配分

効率性	機敏性	継続性
◎	◎	○

サーバのコストを削減するには、サーバに割り当てるシステム資源を、必要最小限にしなければなりません。物理マシンでは、CPU などのシステム資源の変更には、専門知識を持った人間の介在が必要であるため、資源の変更を頻繁に行うことはできません。これに対し、PRIMEQUEST 仮想マシン機能では、仮想マシンに配分するシステム資源を動的に簡単に変更することができます。仮想マシン間で資源を融通することにより、全体としての資源を最小化することができます。

たとえば、オンライン業務とバッチ業務の２つの業務を運用する場合に、オンライン業務は昼間の負荷が高く、バッチ業務は夜間に負荷が高いと仮定しましょう。コストを抑えるためには、小規模と大規模の２台の物理サーバを用意して、オンライン業務とバッチ業務が利用するサーバを昼夜で交換する方法も考えられます。しかし、物理サーバではこのような頻繁な交換は現実的ではありません。結果として、２台ともに大規模なサーバを手配することになります（図 7 の上)。これに対し、PRIMEQUEST 仮想マシン機能を用いれば、１台の物理サーバで、昼間はオンライン業務のゲストに、夜間はバッチ業務のゲストに、多くの資源を配分することができます（図 7 の下)。このように、負荷がピークとなる時間が異なる場合、各負荷のピークの合計ではなく、負荷の合計のピークに対して物理資源を準備すればよいため、全体のコストを削減することができます。

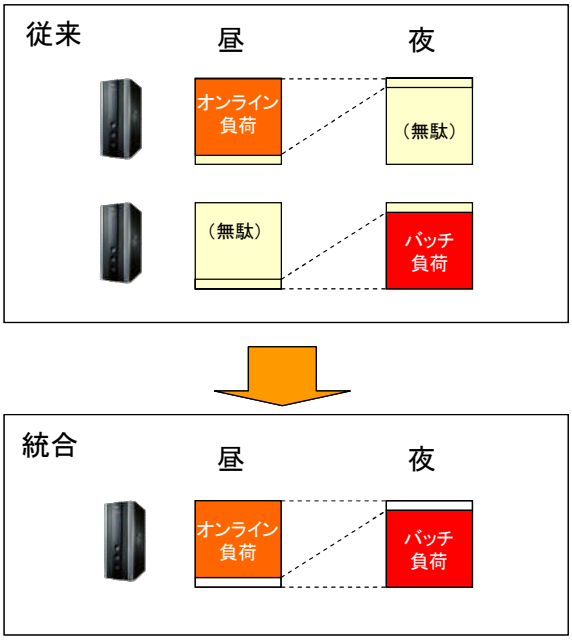


図 7.　動的資源配分

4.5　待機系の統合

効率性	機敏性	継続性
◎	―	◎

ビジネスの継続性を保つためには、サーバがダウンする万一の場合に備え、サーバの冗長化を行うことが必要です。そのため、高信頼性クラスタリングにより、現用系のサーバのほかに、待機系のサーバを用意しておき、現用系に障害が発生した場合には待機系に切り替える処置をとります。複数の現用系サーバがある場合、従来は同じ数の待機系サーバを物理的に用意しておく必要がありました。しかし、現用系に問題が発生することはめったにないため、経営の立場からは待機系サーバのコストが頭の痛い問題でした。

複数のサーバが同時に待機系へ切り替わる可能性が非常に小さいならば、PRIMEQUEST 仮想マシン機能を用いることにより、待機系を 1 台に集約できます。通常時は待機系の各仮想マシンに最小限のシステム資源しか割り当てず、待機系に切り替える必要が生じたときに、その仮想マシンに十分なシステム資源を割り当てます。このように運用することにより、少ない投資で高い可用性を得ることができます。

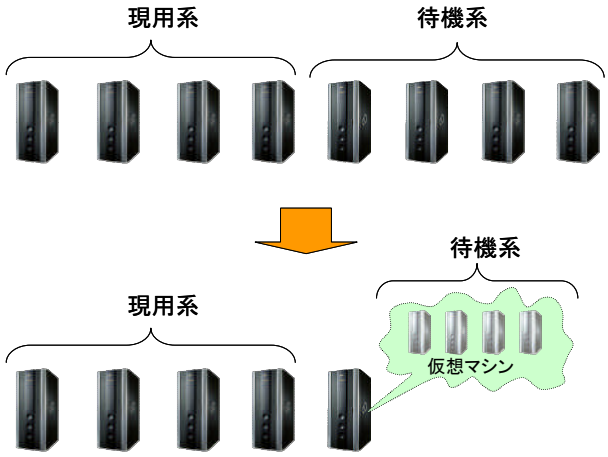


図 8. 待機系マシンの共有

4.6 物理マシン停止時のサービス継続

効率性	機敏性	継続性
―	○	◎

物理マシンは、マシン自身の保守などのために、定期的に電源を切断しなければならないことがあります。PRIMEQUEST 仮想マシン機能を用いれば、このようなときでも仮想マシンを他の物理マシンに移動することにより、ビジネスを継続することができます。

短時間のサービス停止が許される場合には、仮想マシンの OS を一時的に停止して、移行ツールを使って仮想マシンを別の物理マシンに移動してから OS を再起動します（静的マイグレーション）。短時間のサービス停止も許されない場合には、OS を動作させたまま仮想マシンを別の物理マシンに移動することができます（動的マイグレーション、計画中）。

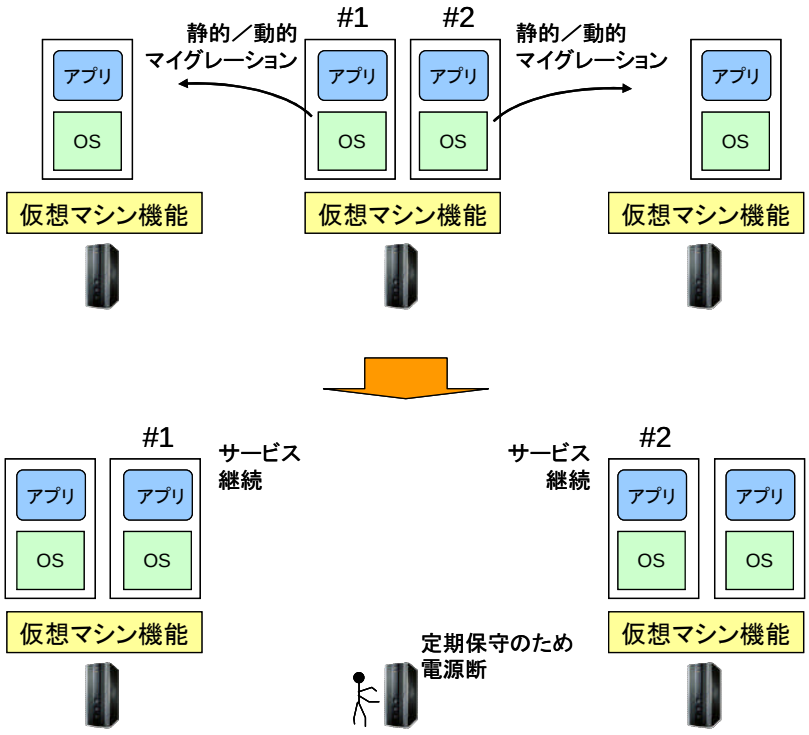


図 9. 定期保守時のサービス継続

5. 仮想マシン機能の実現技術

以下では、PRIMEQUEST 仮想マシン機能を利用するにあたり理解しておくべき実現技術を説明します。

5.1 全体構造

PRIMEQUEST 仮想マシン機能は、ハイパーバイザ方式を採用しています。ハイパーバイザとは、仮想マシン機能を実現する基本モジュールのことです。

PRIMEQUEST 仮想マシン機能は、ハイパーバイザと管理 OS から構成されます。管理 OS には、仮想マシン機能を管理するためのソフトウェアが含まれます。管理 OS では、一般の業務アプリケーションは動作させません。

一般の業務アプリケーションはゲスト OS 上で動作させます。ゲスト OS は、ハイパーバイザの上の仮想マシンで動作します（図 10）。複数のゲスト OS を同時に動作させることができます。

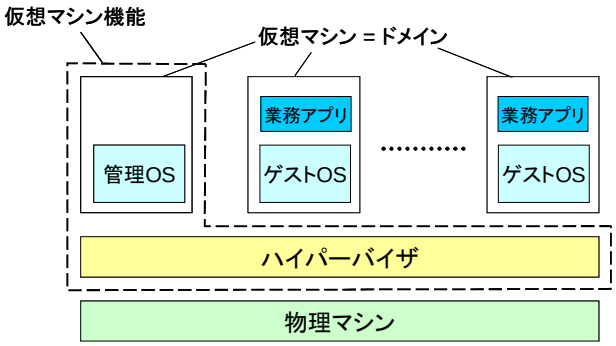


図 10. 仮想マシン機能の構成

各ゲスト OS には、仮想マシン機能により仮想化されたマシン（計算機）が提供されます。ゲスト OS から見ると、物理マシンと仮想マシンには、原則として違いがありません。PRIMEQUEST 仮想マシン機能では、仮想マシンのことをドメイン（domain）と呼びます。各ゲスト OS はドメインの壁により隔離されており、あるゲスト OS が動作不安定になったりハングアップしたりしても、他のゲスト OS はそのまま動作し続けることができます。仮想マシン機能上での動作と対比するため、OS を物理マシン上で直接動作させることを、ネイティブ環境で実行すると言います。また、物理マシン用の OS をネイティブ OS と呼びます。

5.2 パーティショニング技術としての仮想マシン機能

仮想マシン機能により 1 つの物理マシンの上で複数の OS を動作させるということは、物理マシンをソフトウェアで分割して利用していると捉えることもできます。PRIMEQUEST では、PPAR と XPAR というハードウェアによるパーティショニングも提供しています。表 1 に、これらの比較を示します。

表からわかるように、信頼性や性能のわずかな劣化も許されない用途には PPAR や XPAR が適しており、細かい分割粒度や運用管理の柔軟性が必要とされる場合には仮想マシン機能が適しています。両者を組み合わせて、PPAR や XPAR で分割したパーティションの中で仮想マシン機能を動作させることも可能です。図 11 に、4 つのシステムボード（SB）を持つ PRIMEQUEST を 3 つの PPAR に分割し、そのうち 1 つの PPAR を仮想マシン機能でさらに分割する例を示します。

表 1. ハードウェアパーティションとソフトウェアパーティションの比較

	実現手段	分割粒度	特長
PPAR、 XPAR	ハードウェア	大（PPAR: 1SB*、 XPAR: 1/2 SB）	ハードウェアによる信頼性と 強固な隔壁。 性能劣化なし。
仮想マシン	ソフトウェア	小（1 CPU コアを %単位で分割可能）	ソフトウェア管理のための柔軟性。 資源割り当ての動的変更が可能。 ゲスト間の資源共有が可能。

＊ SB: システムボード。PRIMEQUEST500 シリーズでは 1SB は、4CPU ソケット。

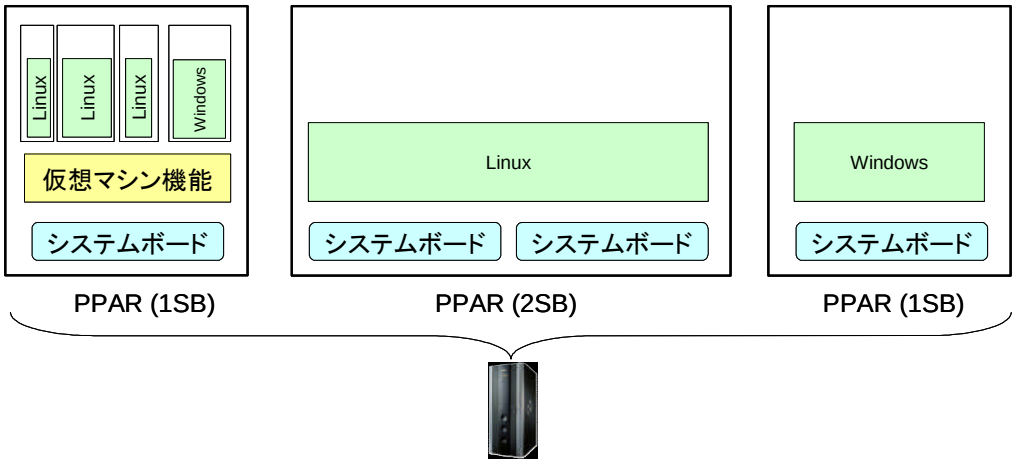


図 11. PPAR と仮想マシン機能の組み合わせ例

5.3 完全仮想化と準仮想化

マシンの仮想化には、準仮想化（para-virtualization）と完全仮想化（full virtualization）という２つの方式があります。物理マシン用の OS のカーネル（OS の中心部分）を、仮想マシン向けに一部変更するのが準仮想化方式で、変更しないのが完全仮想化方式です。PRIMEQUEST 仮想マシン機能の完全仮想化方式は、CPU ハードウェアの仮想化支援機能（Intel VT-i）を利用するため、完全仮想化方式を用いるドメインを HVM ドメイン(Hardware-assisted Virtual Machine Domain)と呼びます。これに対し、準仮想化を用いるドメインを、PV ドメイン（Para-Virtualized Domain）と呼びます。HVM ドメインと PV ドメインは、１つのハイパーバイザの上で共存することができます。

表 2. PV ドメイン（準仮想化）と HVM ドメイン（完全仮想化）

ドメイン種別	OS カーネル	注意点
PV ドメイン（準仮想化）	PV 用に修正したカーネル	PV 用カーネルが用意されている必要がある。管理 OS は常に PV。
HVM ドメイン（完全仮想化）	ネイティブ用と同じカーネル	CPU に仮想化支援機能(Intel VT-i)が必要。十分な I/O 性能を得るためには HVM ドメイン用 PV デバイスドライバが必要。

PRIMEQUEST 仮想マシン機能では、準仮想化に対応していない Windows Server 2003 を動作させるため、また、物理マシンと仮想マシンとの間でクラスタリングする際の OS カーネルの同一性を確保するために、ゲスト用には完全仮想化を用います。

5.4 CPU の仮想化

一般に計算機は、CPU、メモリ、I/O から構成されます。したがって、計算機を仮想化するには、CPU、メモリ、I/O をそれぞれ仮想化する必要があります。この節では、まず CPU の仮想化について説明します。

CPU の仮想化とは、各ドメインに対し、あたかも物理的な CPU（pCPU）を使っているかのように見せることです。ドメインから見える CPU は仮想 CPU（vCPU）と呼ばれます。

PRIMEQUEST 仮想マシン機能では、1 つの物理 CPU を複数のドメインの仮想 CPU から共有することができます。物理 CPU を共有するにあたり、仮想 CPU の CPU 能力を保証したり逆に制限したりして、共有の方法を制御することができます。また、1 つの物理 CPU を 1 つの仮想 CPU に占有させることもできます。

CPU の仮想化を実現しているのは、ハイパーバイザ内にある CPU スケジューラです。この CPU スケジューラが、各物理 CPU をどの仮想 CPU に使わせるかを決めます（図 12）。仮想 CPU 側から見ると、ある瞬間には、仮想 CPU がどれかの物理 CPU に対応づけられているか、あるいは物理 CPU に対応づけられていない（待ち状態）かのどちらかです。OS の中にも OS の CPU スケジューラがあり、仮想 CPU をどのプロセスに使わせるかを決めていますから、全体としては 2 階層のスケジューリングが行われることになります。

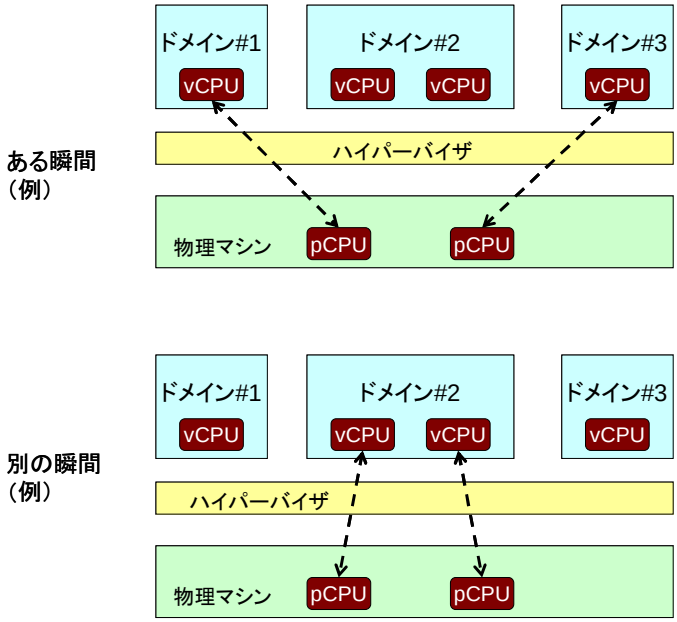


図 12. ハイパーバイザによる CPU スケジューリングの例

5.5 メモリの仮想化

PRIMEQUEST 仮想マシン機能のメモリの仮想化は、（物理マシンの）物理メモリを分割してドメインに割り当てます。ゲスト OS には連続した物理メモリ（ゲスト物理メモリ）が存在するように見えます。しかし、他のドメインのメモリは見えません。各ドメインに割り当てる物理メモリ量は、管理 OS から指定します。

ハイパーバイザ層では、ゲスト物理アドレスからマシン物理アドレスへのアドレス変換をページ単位で行います。OS もメモリの仮想化を行っていますので、全体としては、図 13 のように 2 段階のアドレス変換を行うことになります。

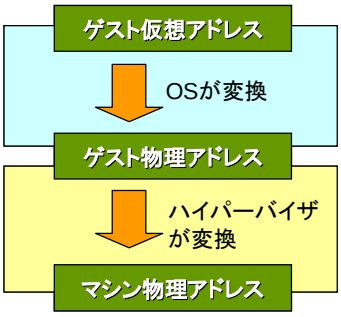


図 13. OS とハイパーバイザによる 2 段階のアドレス変換

PRIMEQUEST 仮想マシン機能のメモリの仮想化は、現在のところアドレス変換のみを行い、二次記憶（ディスク）への退避は行いません。二次記憶への退避は OS 層で行います。

5.6 I/O の仮想化

PRIMEQUEST 仮想マシン機能の I/O の仮想化は、物理マシンの I/O デバイスを使って、ゲスト OS から見える仮想的な I/O デバイスを実現します。仮想デバイスと物理デバイスとの対応づけは、管理 OS 内で行います。

物理的な I/O デバイスを、1 つのゲスト OS が占有するように対応づけることもできますが、ファイバチャネルのカード、ネットワークカード、ディスクボリュームなどを、複数のゲスト OS から共有するように設定することもできます。ただし、共有する場合も、仮想マシン機能でデータを共有することはありません。たとえば、複数のゲスト OS から 1 つのファイルを共有するための機能は、仮想マシン機能では提供しませんので、OS 層の分散ファイルシステムの機能を利用してください。物理カードの数は、サーバハードウェアのカードスロット数により上限がありますが、カードをゲスト間で共有すれば物理的な数の制約に拘束されなくなります。

また、I/O の仮想化により、ゲスト OS が最新式のデバイスに対応していなくても、ゲスト OS から最新の物理デバイスを利用することが可能になります。

PRIMEQUEST 仮想マシン機能における I/O の仮想化には、次のような方式があります。

- デバイスエミュレーション方式
- 仮想デバイスドライバ方式
- 直接 I/O 方式（計画中）

5.6.1 デバイスエミュレーション方式

デバイスエミュレーション方式は、CPU から見える I/O デバイスのハードウェアをソフトウェアによりエミュレート（模擬）する方式です。ゲスト OS では、ネイティブ OS で利用する場合と同じデバイスドライバを使用します。ゲストドメインの CPU からデバイスの制御レジスタやメモリにマシン命令レベルでアクセスされるたびに、制御を管理 OS 内のデバイスエミュレータに渡し、デバイスエミュレータで実際の I/O デバイスと同じ動きを実現します。古い仕様のデバイスを最新式のデバイスでエミュレートすることも、A 社仕様のデバイスを B 社仕様のデバイスでエミュレートすることも可能です。

デバイスエミュレーション方式は、ゲスト OS のデバイスドライバを変更する必要がないという利点がありますが、原理的に実行オーバヘッドの大きい方式です。したがって、低速デバイスや、ゲスト OS において次に紹介する仮想デバイスドライバが利用できない場合にのみ使用し、業務データ用のディスクやネットワークには使用しません。

PRIMEQUEST 仮想マシン機能のデバイスエミュレータは、準仮想化方式のゲストドメインからは利用できません。

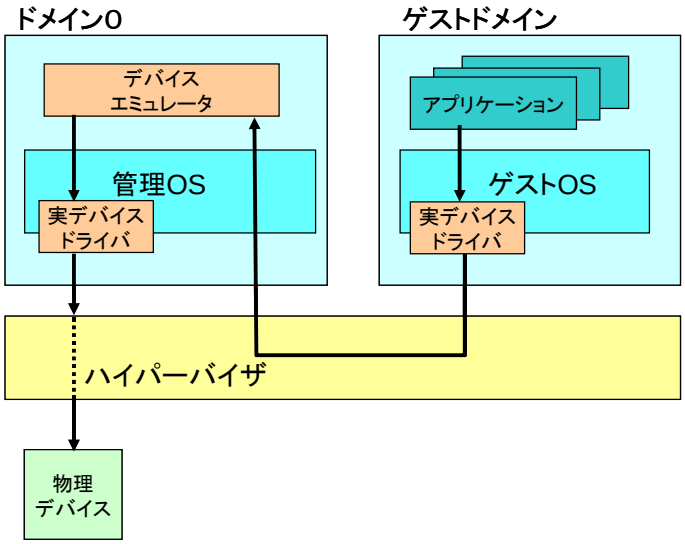


図 14. デバイスエミュレーション方式

5.6.2 仮想デバイス方式

仮想デバイス方式は、抽象的な仮想デバイスを定義し、仮想デバイス専用のデバイスドライバをゲスト OS にインストールして利用する方式です。このデバイスドライバは、仮想環境で動作していることを意識したドライバであるため、PV ドライバと呼びます。PV ドメイン用の PV ドライバと、HVM ドメイン用の PV ドライバがあります。いずれの PV ドライバも、ハイパーバイザ経由で管理 OS 内のバックエンドデバイスドライバと連携して動作します。

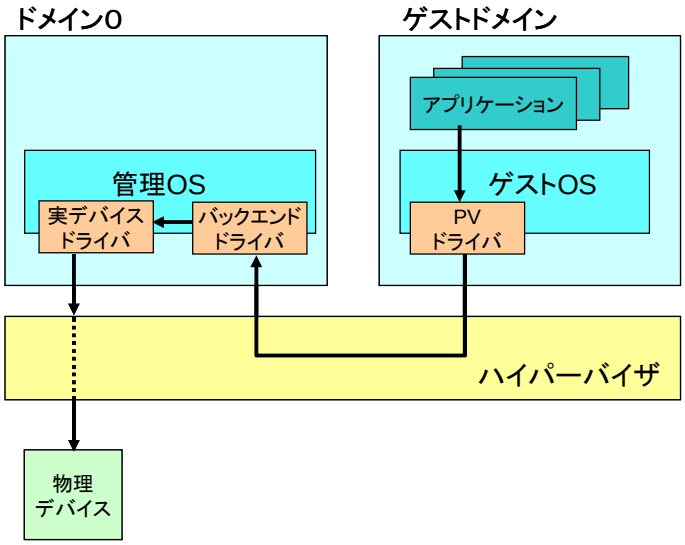


図 15. 仮想デバイス方式

仮想デバイスには、仮想ブロックデバイス（VBD: Virtual Block Device）、仮想ネットワ

ークインタフェース（VNIF: Virtual Network Interface）、仮想 SCSI（VSCSI: Virtual SCSI）、仮想フレームバッファ（VFB: Virtual Frame Buffer）などがあります。

仮想デバイス方式は、デバイスエミュレーション方式と比べて、管理 OS とのやりとりが少なく、ネイティブ環境に近い I/O スループットを得ることができます。また、管理 OS で複数のゲスト OS からのアクセス要求をまとめることにより、ゲスト OS 間で I/O デバイスを共有することができます。管理 OS のドメインには、I/O 量に応じた CPU 能力を割り当てておく必要があります。

5.6.3 直接 I/O 方式（計画中）

直接 I/O 方式は、ゲストドメインから I/O デバイスに直接アクセスする方式です。デバイスの割り当て・解放時のみ、管理 OS を経由します。ゲスト OS では、ネイティブ環境と同じ（または類似の）デバイスドライバを使用します。直接 I/O 方式では、データのやりとりに管理 OS が介在しないため、ネイティブ環境とほぼ同じ I/O 性能が得られます。直接 I/O 方式のためには、直接 I/O を支援するハードウェアが必要です。直接 I/O 方式では、一般にゲスト間でデバイスを共有することができません。共有するには、物理デバイス側に共有のための機能が必要になります。

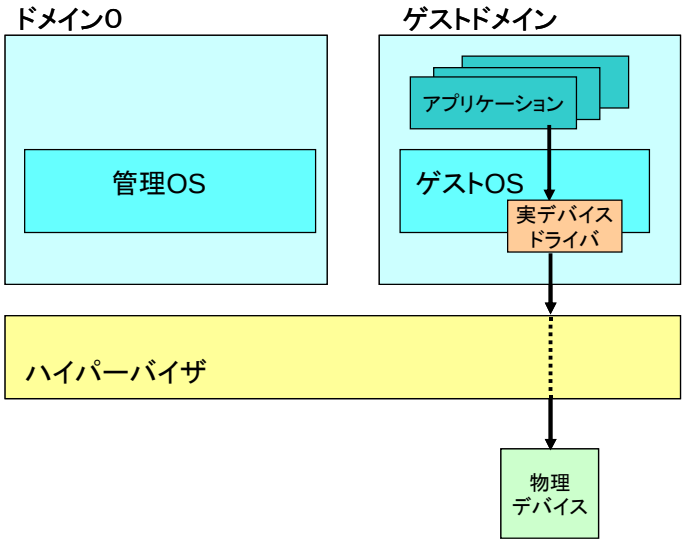


図 16. 直接 I/O 方式

5.7 ディスクの仮想化

I/O デバイスの 1 つであるディスク装置の仮想化は、I/O 仮想化の 1 つの例です。デバイスエミュレーション方式や仮想デバイス方式の場合、管理 OS 内でゲストのディスクと物理ディスクを次の図のように対応づけることができます。

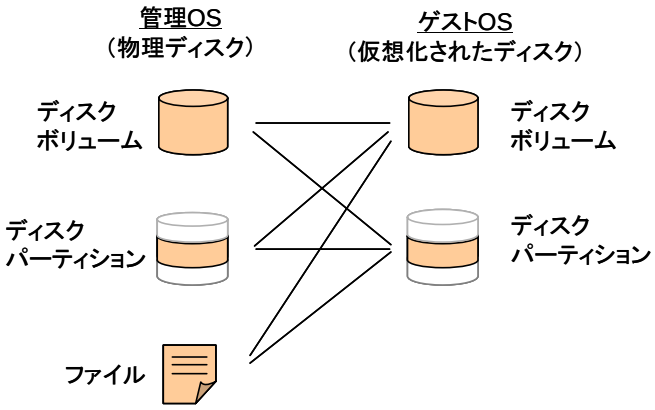


図 17. ディスクの仮想化

管理 OS 側のディスクボリュームをディスクパーティションに分割して、それぞれ別のゲスト OS のディスクボリュームに対応づけることにより、1 つのディスクボリュームを複数のゲストから共有することができます。

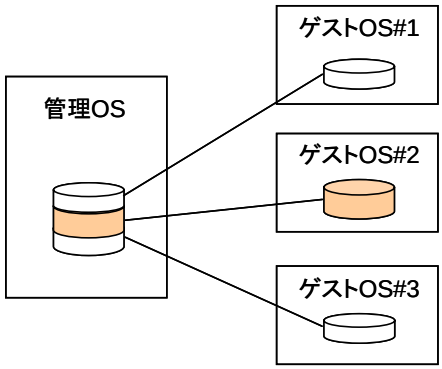


図 18. ディスクのゲスト間共有

性能や高信頼性を求める場合には、管理 OS 側のディスクまたはディスクパーティションに対応させます。管理 OS 側のディスクやパーティションとしては、管理 OS の上でディスクやパーティションに見えるものならば、原則としてどのようなものでも利用可能です。たとえば、富士通のストレージシステム ETERNUS などディスクアレイ装置内で仮想化したディスクを利用することも可能です。

管理 OS 側をファイルに対応させる方法は、管理が容易であるという利点があります。ゲスト側でディスクが必要になったときに、管理 OS 側ではファイルを 1 つ作成するだけで済みます。ただし、ファイルに対応させると、ディスクやディスクパーティションに対応させる場合に比べて、性能の点で不利です。

5.8 ネットワークの仮想化

ネットワークインタフェース（NIF）の仮想化により、物理的な LAN カード数の制限に

拘束されずに、仮想ネットワークインタフェース（VNIF）をゲスト OS に提供することができます。複数のゲストの VNIF から 1 つの物理 NIF を共有することができます。そのためには、管理 OS 側に、ソフトウェアで実現されたイーサネットスイッチ（仮想ブリッジ）を用意します（図 19）。VNIF と仮想ブリッジ、物理 NIF の対応づけは、管理 OS で行います。

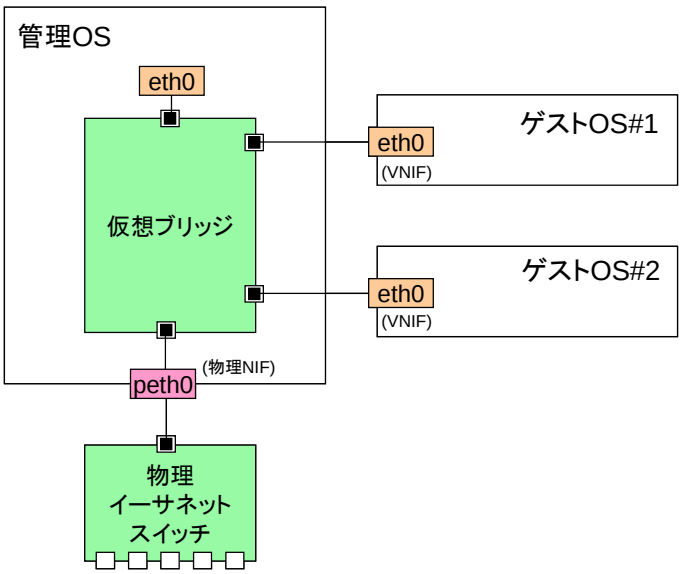


図 19. ネットワークカードのゲスト間共有

冗長化や性能分離の目的で、2 つの VNIF を別の物理 NIF に対応させたい場合には、2 つの物理 NIF に対応した 2 つの仮想ブリッジを用意して、それぞれに VNIF を接続します。

性能を重視する場合は、ゲストドメインに見える VNIF と物理 NIF を 1 対 1 に対応させて、ゲストに物理 NIF を占有させます。

VNIF には、LAN 上で重複しない MAC(Media Access Control)アドレスを割り当てる必要があります。

6. 仮想マシンの複製と移動

6.1 クローニング（仮想マシンの複製）

運用管理ツールのクローニング機能を使えば、既存の仮想マシンのソフトウェアスタック（OS からアプリケーションまでの同じ組み合わせ）を別の仮想マシンにコピーして、マシン固有の設定を施すだけで、同じ実行環境を短時間に用意することができます。

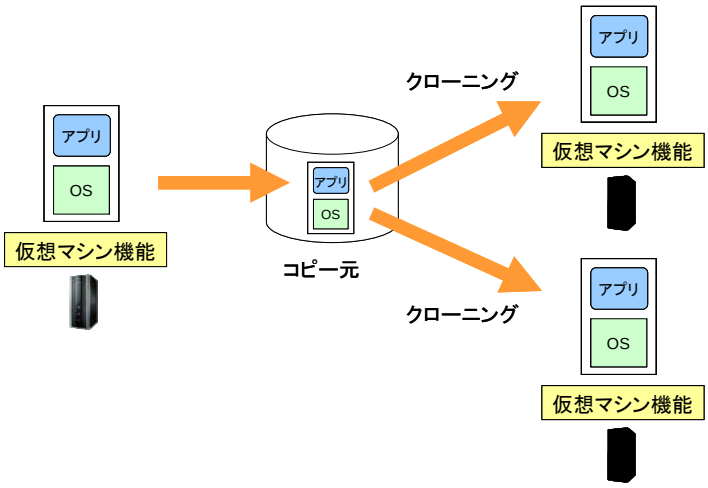


図 20. クローニング

6.2 静的マイグレーション（仮想マシンの静的移動、計画中）

仮想マシンと物理マシンの間で、あるいは、物理マシンが異なる２つの仮想マシンの間で、OS とその上のアプリケーション（ソフトウェアスタック）を移動したい場合があります。移動の際に一旦 OS を停止する方法を、静的マイグレーションと言います。富士通では、次のような静的マイグレーションを支援する移行ツールを提供する予定です。

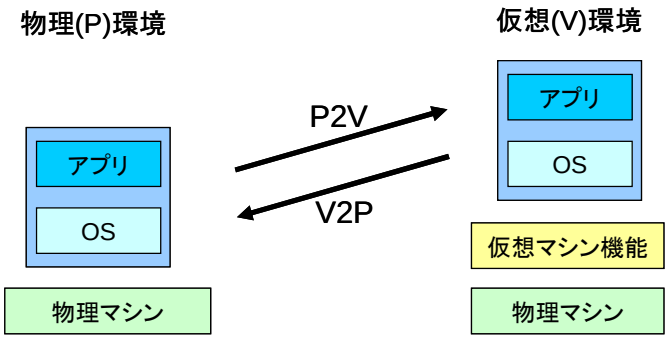


図 21. P2V と V2P

(1) 物理マシンから仮想マシンへ（P2V）

物理マシン(P)で動作させていたソフトウェアスタックを、仮想マシン(V)に移動させる P2V ツールです。たとえば、複数のサーバで行っていた業務を１台の物理マシンの上に仮想化して統合するときに使用します。

(2) 仮想マシンから物理マシンへ（V2P）

仮想マシン(V)として動作させていたソフトウェアスタックを、物理マシン(P)に移動させる V2P ツールです。開発を仮想マシン上で行い、実運用を物理マシン上で行う場合などに使用します。

(3) 仮想マシンから仮想マシンへ

同じ PRIMEQUEST 仮想マシン機能を使う２つの物理マシン間で、OS を一旦停止させてソフトウェアスタックを移動させるツールです。たとえば、中長期的な負荷分散や、計画保守時のサービス継続のために使用します。

6.3 動的マイグレーション（仮想マシンの動的移動、計画中）

仮想マシンから仮想マシンへの移動を、OS を動作させたまま行うことを、動的マイグレーション、または、ライブマイグレーションと言います。動的マイグレーションにより、次のようなことが可能になります。

- 動的負荷分散： 運用を停止することなく、高負荷サーバから低負荷サーバへゲストドメインを移動して負荷の最適化を行う。
- ハードウェア保守時のサービス継続： 計画保守のため物理マシンを停止する必要がある場合に、ゲストドメインを他の物理マシンに移動してサービスを継続する。

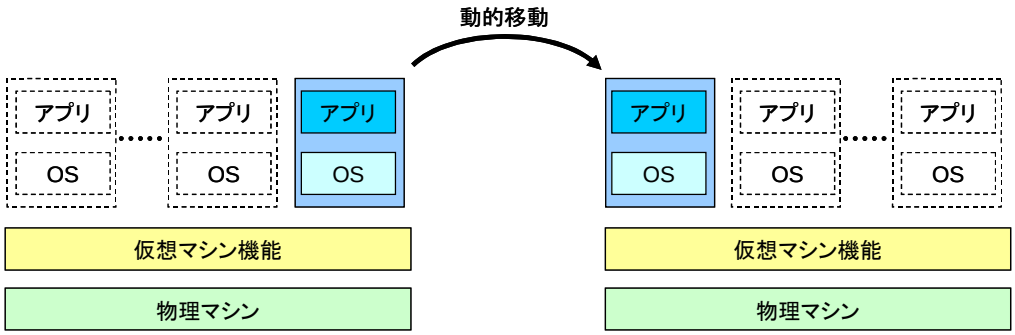


図 22. 動的マイグレーション

7. おわりに

このホワイトペーパーでは、Red Hat Enterprise Linux 5 で提供される PRIMEQUEST 仮想マシン機能について説明しました。

富士通では、お客様に対して本仮想マシン機能に対する導入支援、性能向上、信頼性向上のための当社独自付加ソフトウェアを含むサポートサービスを提供してまいります。また、Xen や Linux 等のオープンソースコミュニティに対する積極的な貢献および、Red Hat 社との協業を通じて、仮想マシン機能の信頼性と使いやすさの向上に引き続き努力してまいります。