Oracle Solarisコンテナを使ってみよう ～概要/設計編～

2011年10月(第2版)

富士通株式会社

はじめに

FUJITSU

- 本書は、Oracle Solaris 10 9/10で提供される機能をベースに作成しています。
- 最新のOracle Solaris の情報については、マニュアルにてご確認ください。
  - **Oracle Solaris 10 Documentation** (Oracle社webサイトへリンク)
    http://www.oracle.com/technetwork/documentation/solaris-10-192992.html
- 本書では Oracle Solaris コンテナをSolaris コンテナと記載することがあります。

Oracle Solarisコンテナ機能のサポート

FUJITSU

■Solarisコンテナ機能はSPARC Enterprise全モデルでサポートしています（Solaris 10 OS）

ミッションクリティカル

- SPARC64 VII/VII+搭載
- 高い処理性能とスケーラビリティ
- メインフレーム並の信頼性

～幅広い業務に最適～

M3000

SPARC64 VII+
1CPU(2コア/4コア)
2.86GHz

M4000

SPARC64 VII+
最大4CPU
(8～16コア)
2.66GHz

M5000

SPARC64 VII+
最大8CPU
(8～32コア)
2.66GHz

M8000

SPARC64 VII/VII+
最大16CPU
(64コア)
2.88/3.0GHz

M9000

SPARC64 VII/VII+
最大64CPU
(256コア)
2.88/3.0GHz

スループット コンピューティング

- SPARC T3搭載
- 高いスループット性能
- 省電力、省スペース

～特にWebフロント業務、
アプリケーションサーバ等に最適～

T3-1

SPARC T3
1CPU (16コア)
1.65GHz

T3-2

SPARC T3
2CPU (32コア)
1.65GHz

T3-4

SPARC T3
4CPU (64コア)
1.65GHz

FUJITSU

Oracle Solarisコンテナ概要

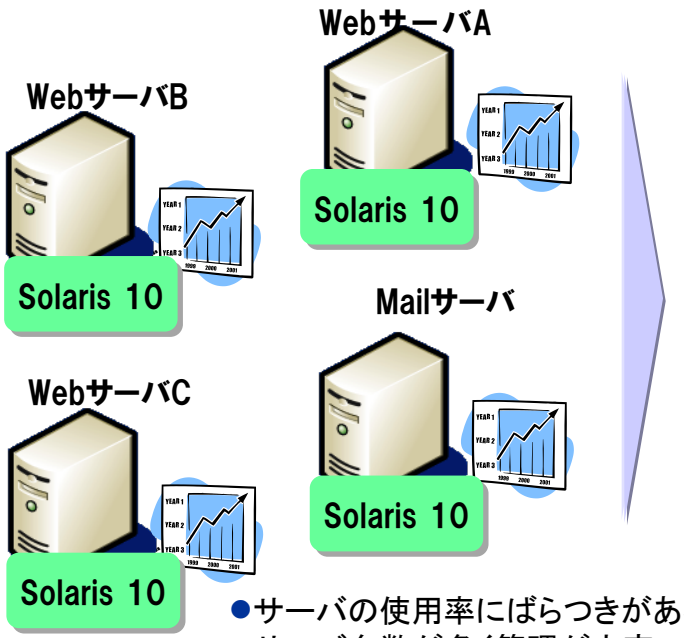
OSの仮想化機能：Oracle Solarisコンテナ

FUJITSU

■1つのSolaris 10環境上に、複数の仮想Solaris環境（Solarisコンテナ）を構築可能

- ハードウェア構成に依存せず、最大8191個の仮想OSを構築可能
- 仮想OSの追加・削除は簡単な作業で短時間に行うことが可能
- 仮想OS毎のOSインストール、パッチ適用は不要
- CPU、メモリなどのハードウェアリソースを柔軟に配分可能

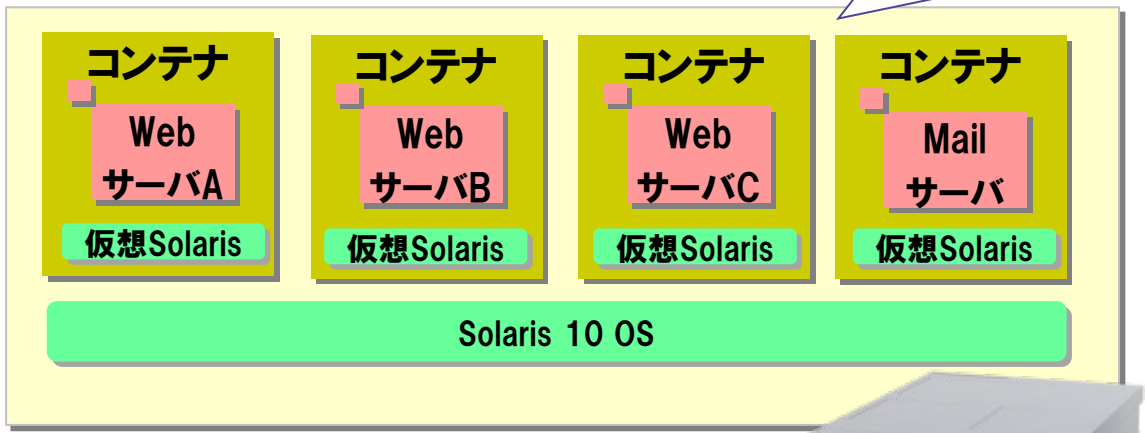
従来



- サーバの使用率にばらつきがある
- サーバ台数が多く管理が大変
- 電力やスペースを削減したい

Solarisコンテナで集約

業務の負荷状況に応じたリソース配分で、リソースの利用効率を向上



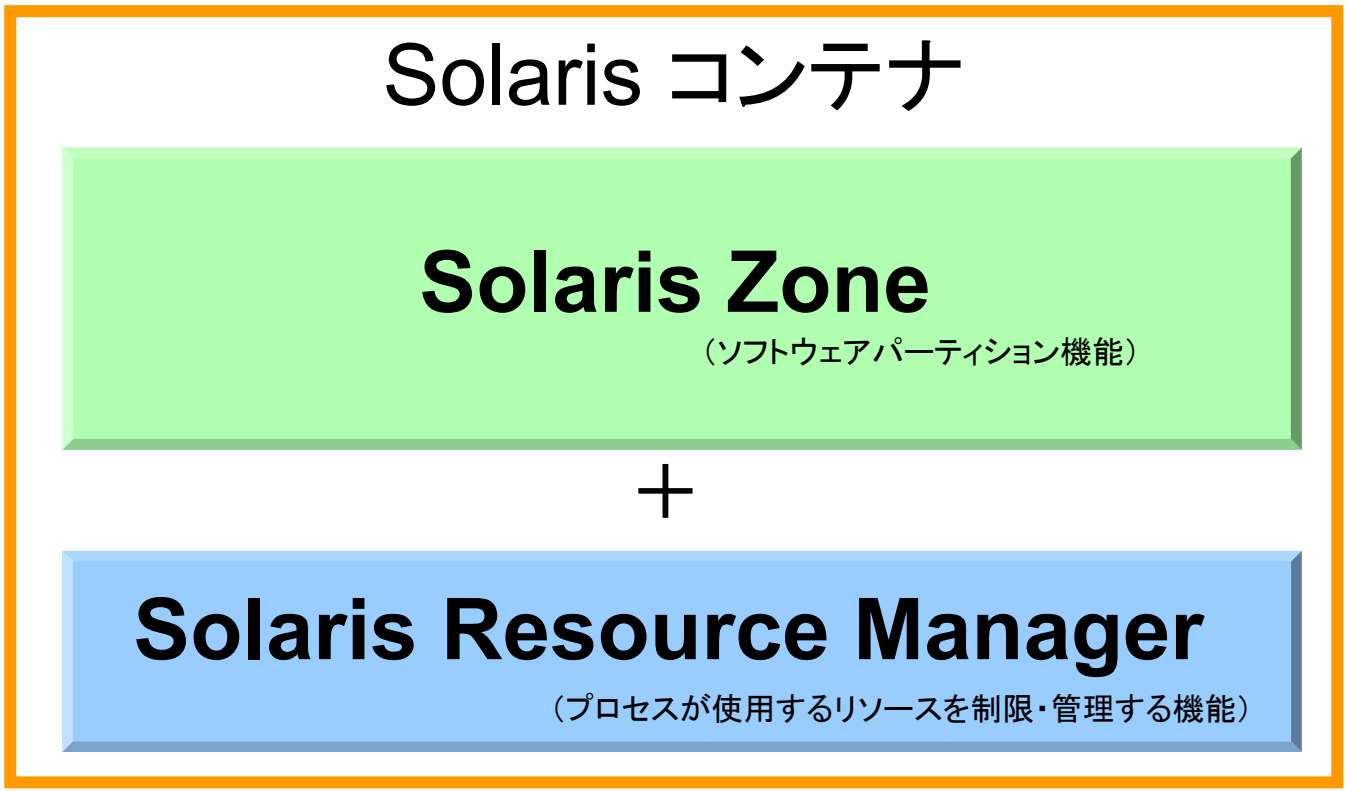
OSのメンテナンスやバックアップが一回に運用負荷を低減

簡単な設定変更でスピーディにSolaris環境を構築可能

Oracle Solarisコンテナの定義

FUJITSU

SolarisコンテナはSolaris**ゾーン機能と**Solaris**リソースマネージャ機能を組み合わせて構成します**

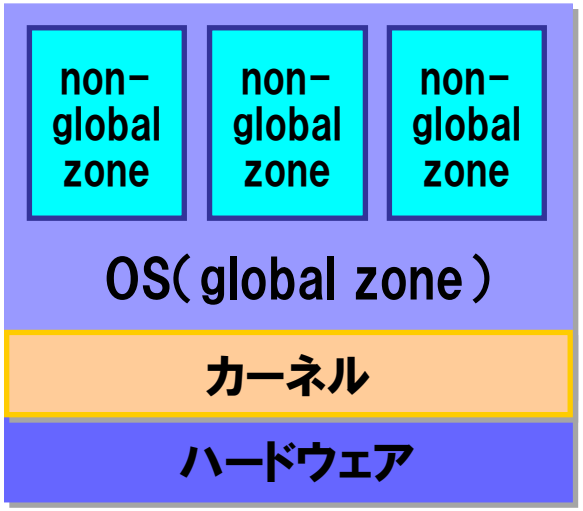


２種類のSolaris zone

FUJITSU

Solaris zoneは global zoneとnon-global zoneの２種類

- global zone:　**従来のOSに相当**
  - ✓OBPからbootするOS
  - ✓すべての物理デバイスにアクセス可能
  - ✓ハードウェア情報を取得可能
  - ✓ソフトウェアパーティション(non-global zone)の設定/制御が可能

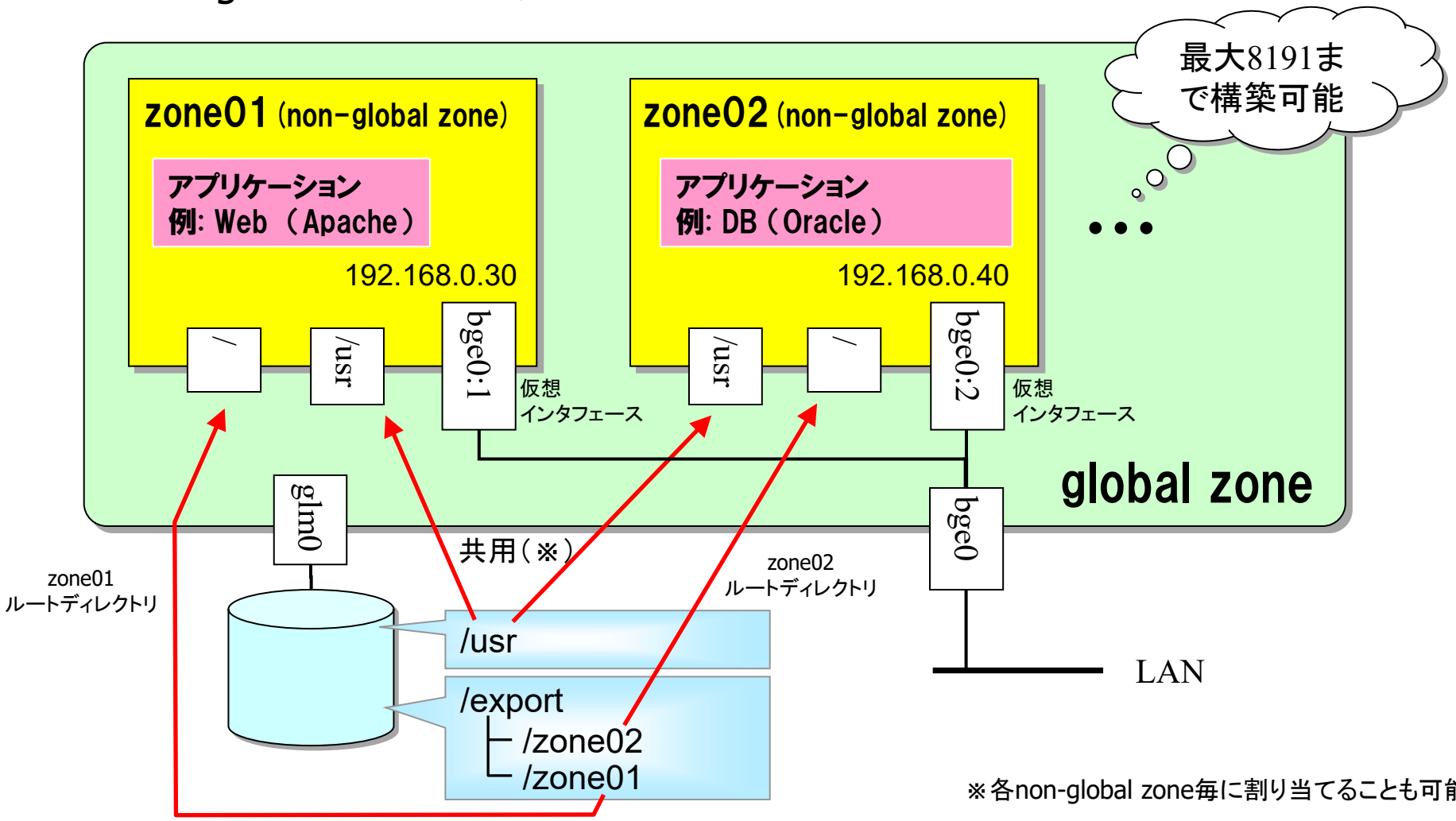


- non-global zone(以降、zoneとも表記):
  global zone上に構築されたソフトウェアパーティション
  - ✓固有のIPアドレスを持つ（zone間はネットワーク通信のみ）
  - ✓zone毎にroot権限を設定
  - ✓zone毎にboot、reboot、shutdown 可
  - ✓一つの zoneがクラックされても、他の zoneには影響なし
  - ✓許可された物理デバイスのみアクセス可能

global zoneとnon-global zoneの関係概念図

FUJITSU

non-global zoneはOS環境として必要なファイルシステムやネットワーク、その他デバイスはglobal zoneから設定、共有され構成されています



※各non-global zone毎に割り当てることも可能

FUJITSU

Oracle Solarisコンテナ適用シーン

Oracle Solarisコンテナ適用による効果

FUJITSU

①システム環境構築のスピードアップ

- ✓Solaris OS上にインストールされた一部のアプリケーションは、自動的にSolarisコンテナへインストールされる。
- ✓用途別に複数のSolarisコンテナ環境を即時に用意することが可能。

②必要リソースの最小化によるコストダウン

- ✓Solaris OSとSolaris コンテナ間、またはSolarisコンテナ同士でCPUやファイルシステム、物理デバイスを共有することが可能。
- ✓CPUリソースの細分化や動的変更により、リソースの有効活用が可能。
- ✓CPU、メモリリソースのキャッピングにより、必要資源の確保が可能。

③システム運用の効率化

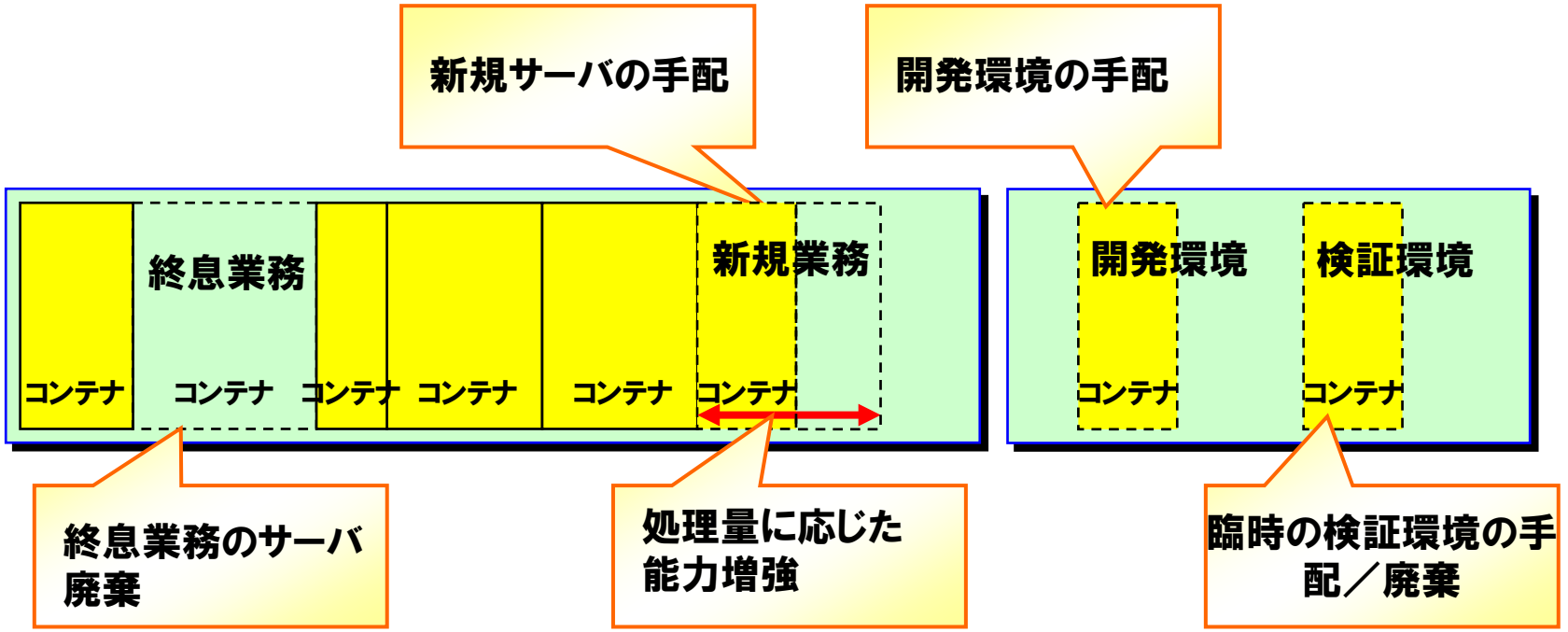
- ✓迅速な起動/停止/再起動が可能。
- ✓Solaris OSに適用した修正適用を全Solarisコンテナへ自動適用することが可能。
- ✓並列パッチの機能でコンテナへのパッチの適用時間を短縮することが可能。

Solarisコンテナ適用効果 ①スピードアップ

FUJITSU

コマンド投入後、数10分で独立したサーバ環境の提供／廃棄が可能

- サーバ構成の設計／手配の手間がなくなり、新規業務立ち上げがスピードアップ
- 終息業務に対するサーバの廃棄が不要
- 臨時の検証が必要な際も、環境の手配／廃棄がスムーズに可能

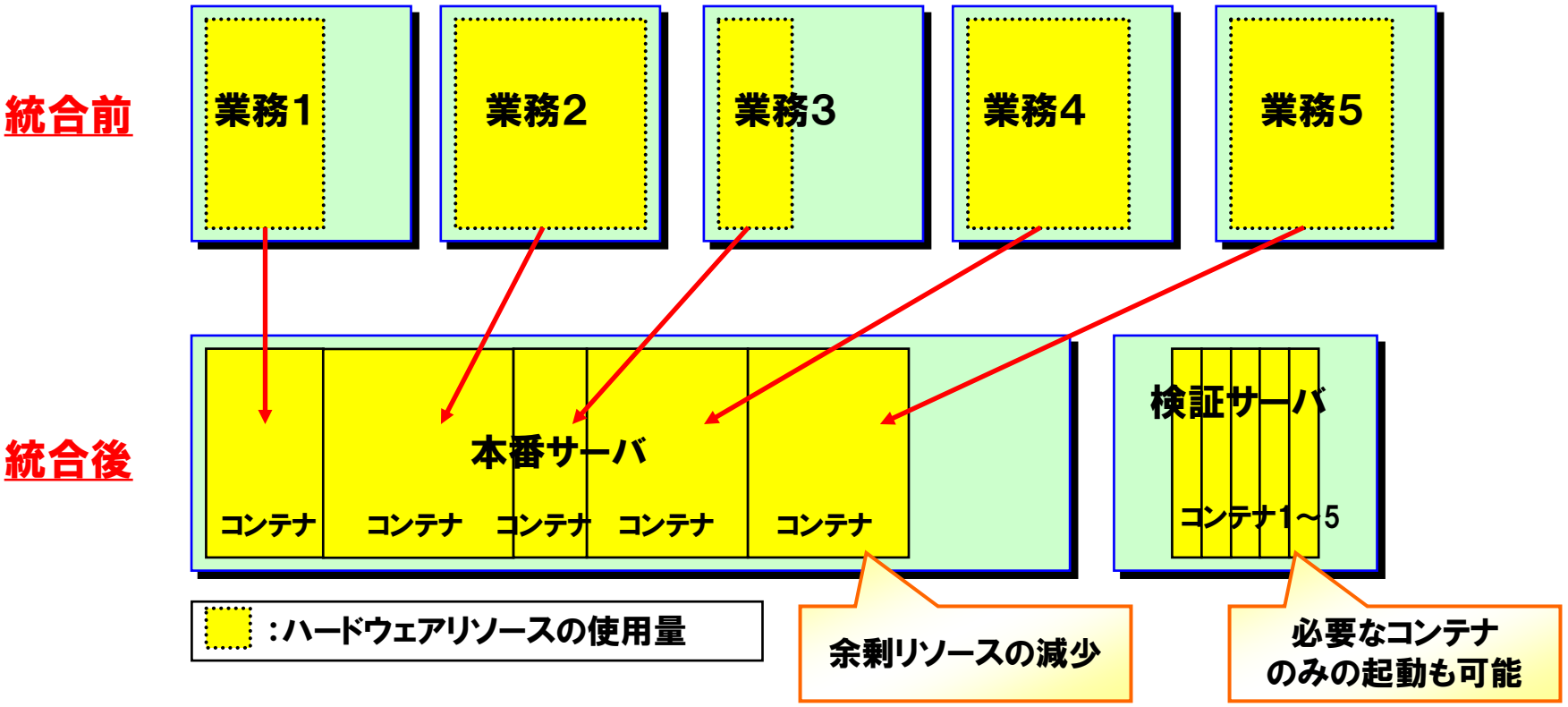


Solarisコンテナ適用効果 ②コストダウン

FUJITSU

サーバの仮想化により、業務間の独立性を維持したままサーバ統合が可能

- ハードウェアリソース（CPU/メモリ/ディスク容量）の有効利用による導入コストの削減
- 管理対象のサーバ、ネットワーク機器の削減による運用コストの削減
- 複数台構成システムの検証環境を1台のサーバに構築可能

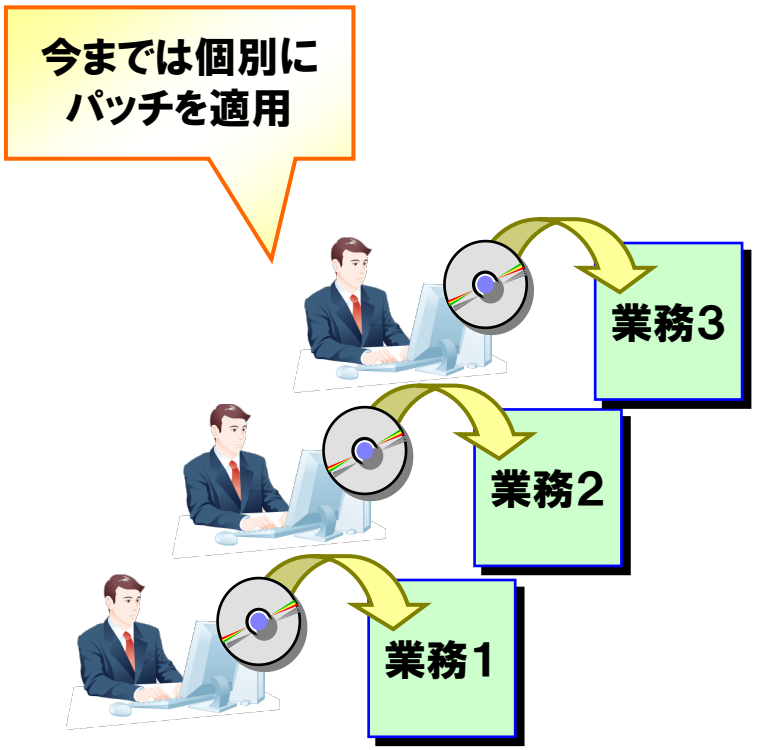


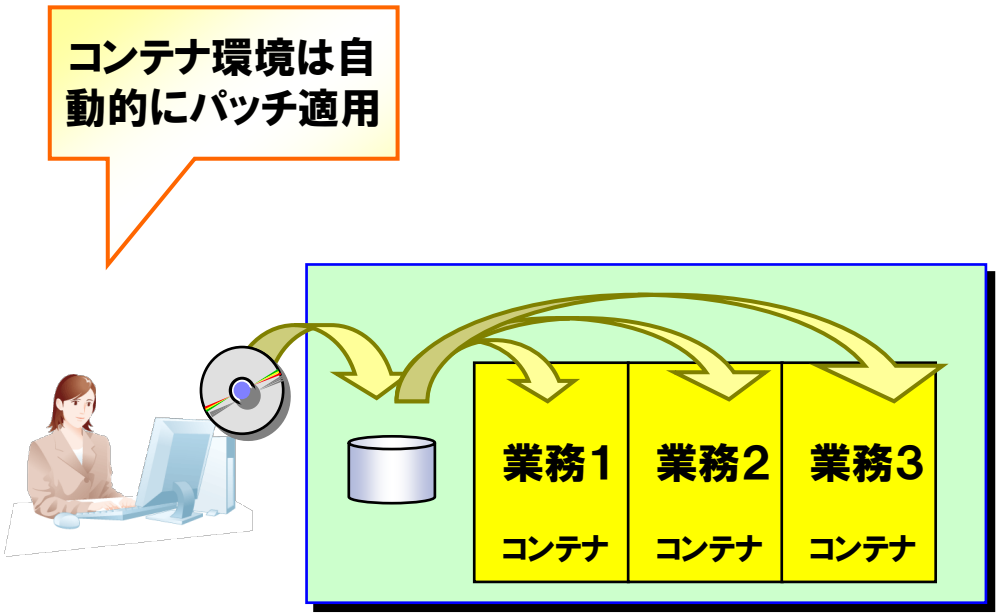
Solarisコンテナ適用効果 ③運用効率化

FUJITSU

コンテナの特性を活かした効率的なシステム運用を実現

- コンテナの迅速な起動・停止処理。（起動時間：約10秒、停止時間：約5秒、再起動時間：約15秒）
- パッチはglobal zoneに適用するだけで自動的にコンテナへも適用。（但し、コンテナ内のミドルウェアパッチは個別適用）
- 並列パッチの機能でコンテナへのパッチの適用時間を短縮。（Solaris 10 10/09以降）





FUJITSU

Oracle Solarisコンテナ統合

Solarisコンテナ統合適用シーン(1)

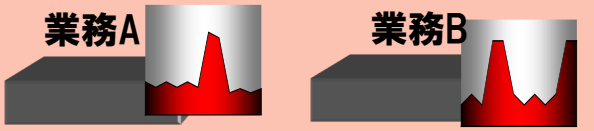
FUJITSU

- **業務間の隔離性を維持したまま、CPUリソースの有効活用（増強、配分）が可能**
  - 業務毎のピーク時間差を活用
  - 業務の負荷状況に応じて、柔軟かつ動的にCPUリソースの再配分が可能

サーバ稼働率を大幅向上

集約前

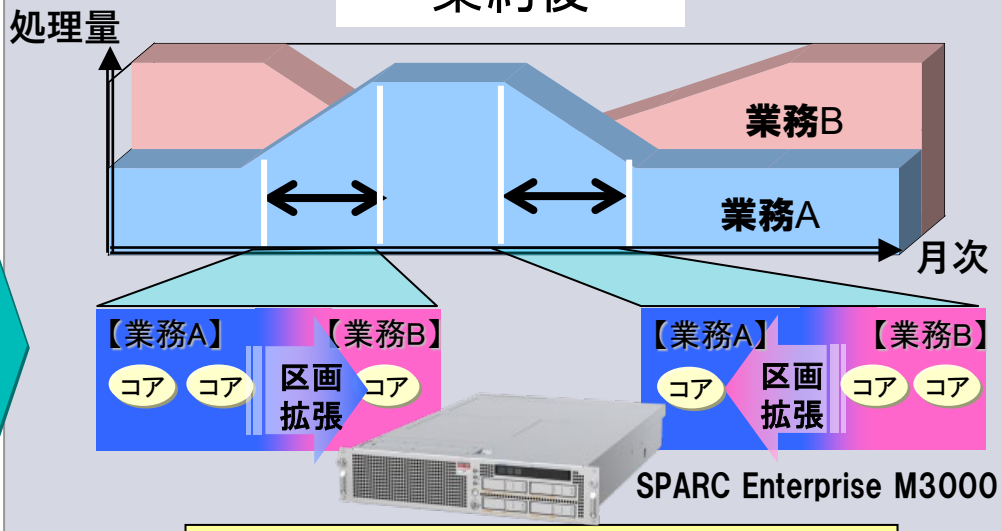
ピーク時を想定



業務A	業務B
通常：2CPU, 1GB 最大：4CPU, 2GB	通常：2CPU, 1GB 最大：4CPU, 2GB
4CPU + 2GB	4CPU + 2GB

平均稼働率
30%

集約後



業務間の独立性を維持したまま、
動的にCPUリソースを移動

- DBごとのCPUリソース優先度の設定
  （オンライン業務のレスポンスを確保し、バックグラウンド業務を統合できる）
- 時間帯別にCPUリソース量を最適化
  （バッチ業務の終了時間を守り、オンライン業務も止めない）

Solarisコンテナ統合適用シーン(2)

FUJITSU

- **スケールアウトで同一のサーバ環境を複数台構成にしているシステムのリプレースに適用**
  - ミドルウェアのライセンス数削減によりコスト削減を実現
  - 業務環境は変えずにシステムを統合しサーバリプレース

集約によりTCOを大幅削減

集約前

PRIMEPOWER250
1.98GHz x2CPU(2コア)
4GBメモリ

Interstage Web Server
×4ライセンス

Interstage
ライセンス
削減(▲50%減)

集約後

SPARC Enterprise M3000
2.86GHz x 1CPU(4コア) 8GBメモリ

Interstage Web Server
×2ライセンス(※)

※ 4コア × 0.5(マルチコア係数) = 2プロセッサライセンス

サーバ集約による費用/エコロジー効果

	PW250 x 2台	M3000 x 1台	効果
性能(相対値)	1	1.2	20%性能向上
Interstage	4プロセッサライセンス	2プロセッサライセンス	50%削減
消費電力料金	154千円/年	54千円/年	65%削減
CO2排出量	625.6Kg-CO_2	220.8Kg-CO_2	65%削減
質量(重量)	134Kg	23Kg	83%削減
占有ピッチ数	4U	2U	50%削減

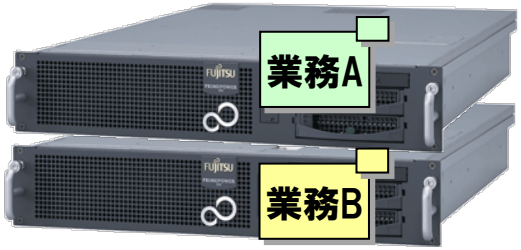
杉の木
30本/年 分

最新サーバは高性能で省電力

- 柔軟なシステム運用（リソース最適化）
- 運用の効率化（バックアップ、保守性）
- グリーンIT（省電力、省スペース）

信頼性が高いから、
サーバ統合も安心！

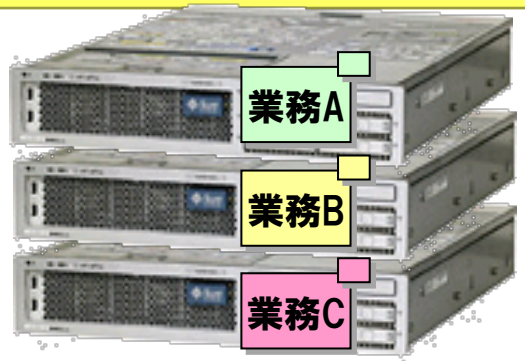
PRIMEPOWER 250を複数台運用



4U

総電力量: 最大 1,260W（630W/台 x 2）

Sun Fire V245を複数台運用



6U

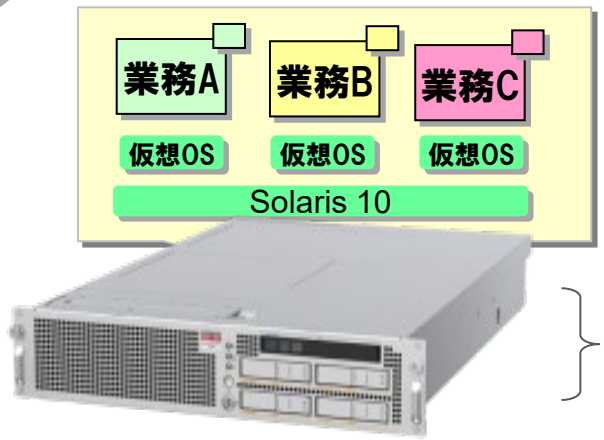
総電力量: 最大 2,280W（760W/台 x 3）

●消費電力、スペース1/2

同等性能

●消費電力1/4
●スペース1/3

SPARC Enterprise M3000に統合（Solarisコンテナ）



2U

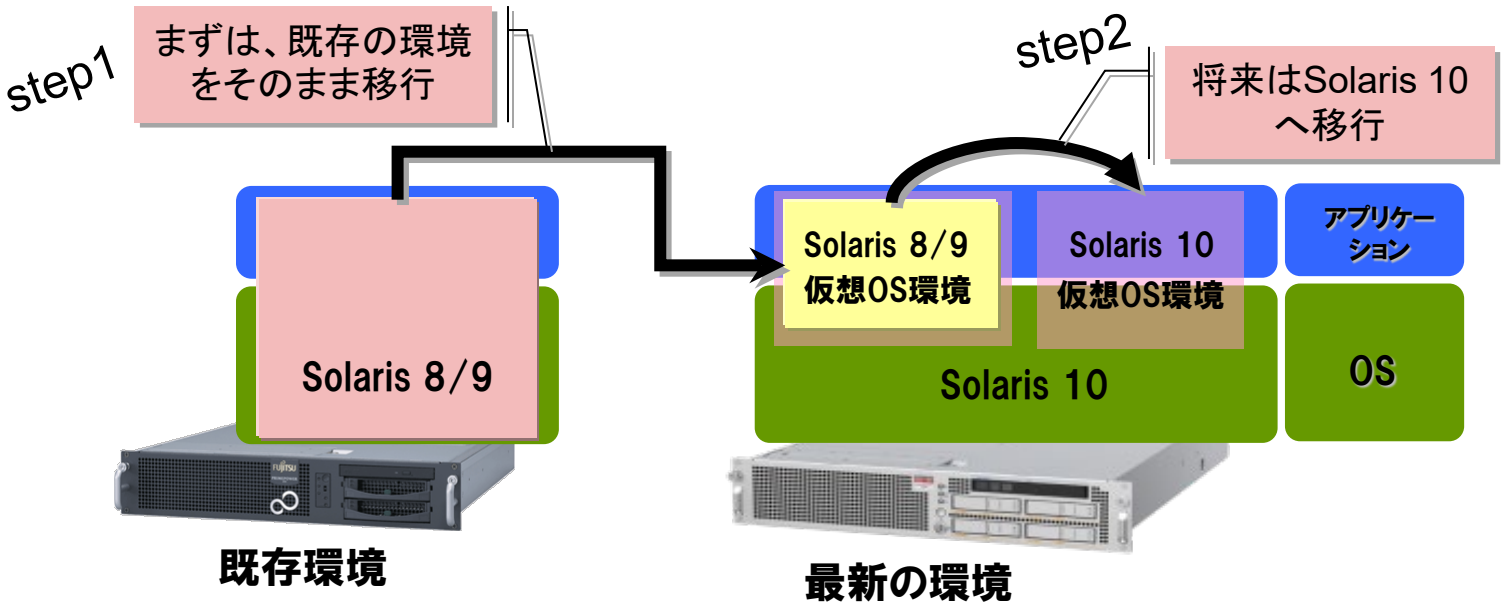
総電力量：最大 505W

＊200V接続時は500W

サーバ統合で、
スペース効率も改善！

【参考】Oracle Solaris Legacy Containers

- Oracle Solaris Legacy Containers(旧Solaris 8/9 コンテナ)はSolaris 10のコンテナ上でSolaris 8/9環境を動作させる仮想化機能(有償プロダクト)です。
- この機能によりSolaris 8/9の資産を一時的にSolaris 10環境で稼動させてSolaris 10への全面移行に向けた橋渡しを行います。



FUJITSU

Oracle Solarisコンテナの設計

Solarisコンテナの設計指針（１）

FUJITSU

■CPUリソースの設計指針

（１）容量見積もり

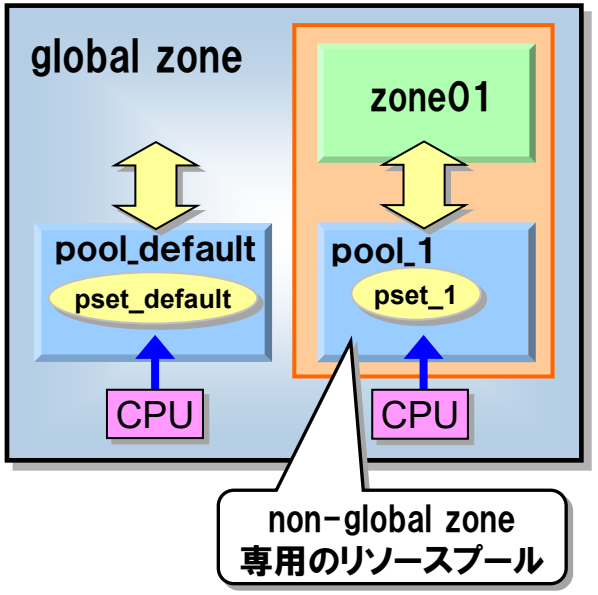
global zone 用のリソースプール（＝pool_default）には最低1CPUが割り当てられます。よって、CPU数を見積もるには、各non-global zone に必要なCPUの合計に1CPU分加算する必要があります。

※マルチコアCPUの場合はコア単位で換算します。

（２）non-global zone 用のリソースプール設定

non-global zone 用のリソースプールはglobal zone と同じpool_default に構成可能ですが、CPUリソースの競合が発生すると global zoneの動作に影響を与えます。

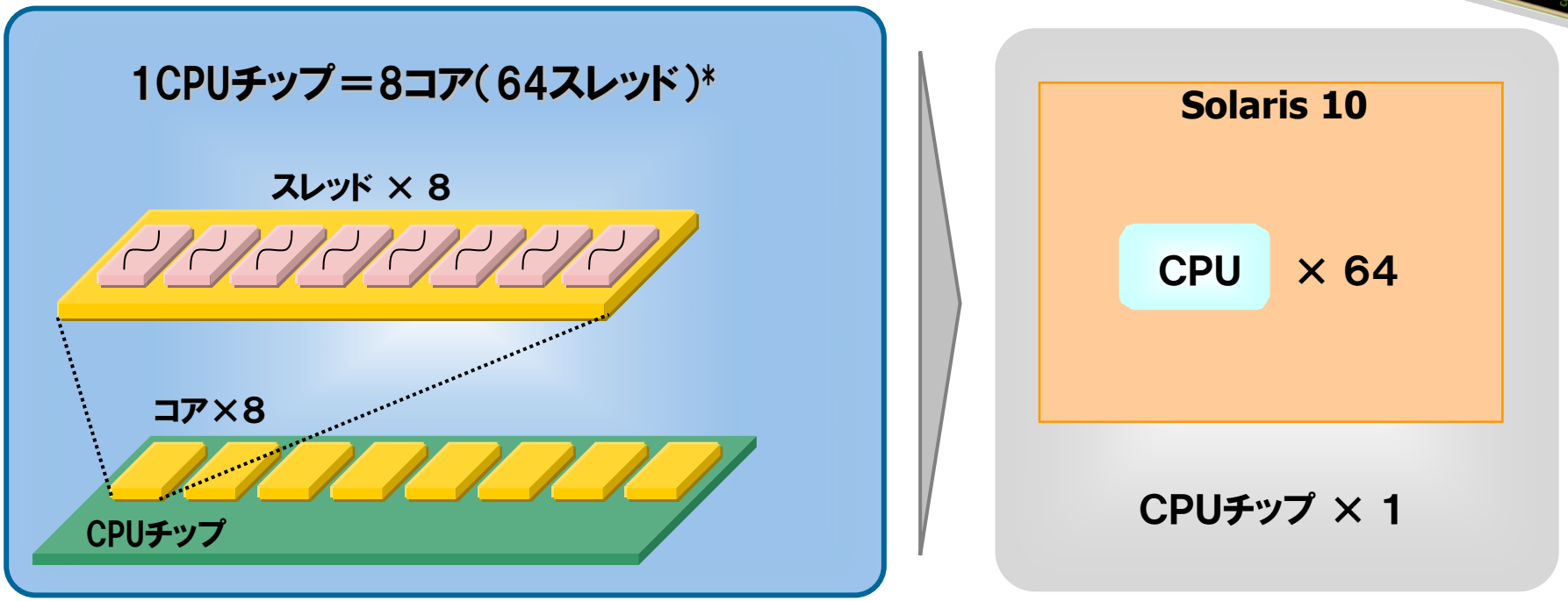
これを避けるため、non-global zoneには専用のリソースプールを設定することを推奨します。



<参考>コア搭載CPUのリソース認識（１）

FUJITSU

UltraSPARC T2（SPARC Enterprise T5120/T5220搭載）



Solaris OSがCPUリソースとして認識するのは、スレッド単位です。
UltraSPARC T2の場合１CPUチップ搭載で、OSからは64CPUと認識されます。

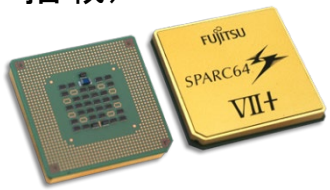
*:型名によって4コア/8コアモデルがあります

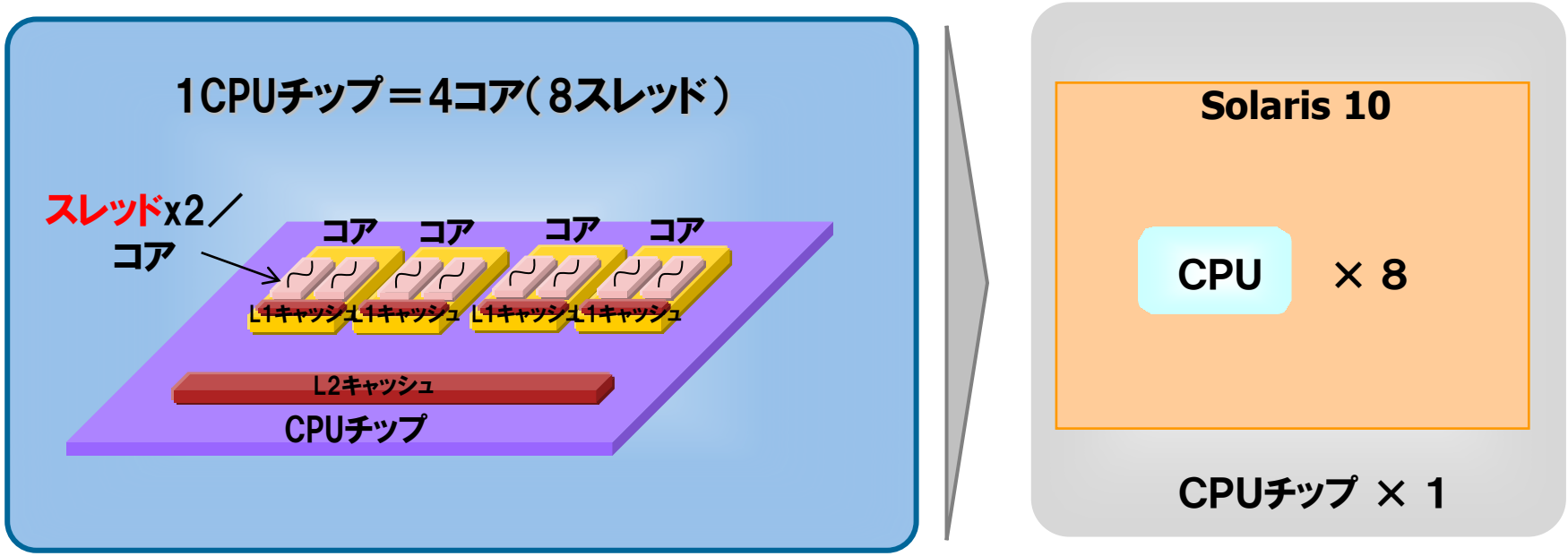
✓ T5140/T5240に搭載されるUltraSPARC T2 Plus は最大8コア/8スレッドのプロセッサを2チップ搭載します。よって、OSからは最大128CPUと認識されます。

<参考>コア搭載CPUのリソース認識（２）

FUJITSU

SPARC64　VII/VII+（SPARC Enterprise M3000/M4000/M5000/M8000/M9000 搭載）





Solaris OSがCPUリソースとして認識するのは、スレッド単位です。
SPARC64 VII/VII+の場合１CPUチップ搭載で、OSからは8CPUと認識されます。

＊:型名によって2コア/4コアモデルがあります

Solarisコンテナの設計指針（２）

FUJITSU

■メモリ容量の設計指針

（１）容量見積もり

global zone及び各non-global zone上で動作するアプリの使用メモリ量の総和から見積ります。

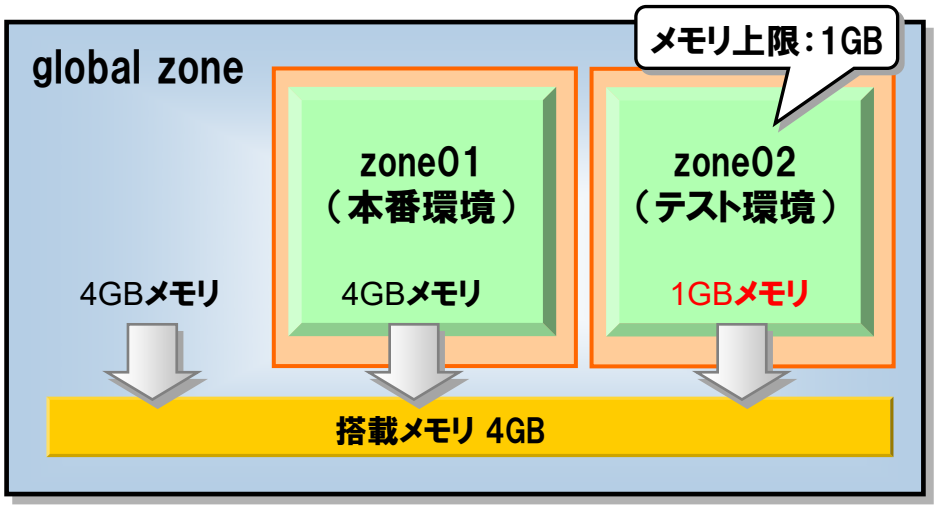
（２）物理メモリ使用量の設定

各non-global zoneが使用する物理メモリ使用量の上限設定のみ可能です。
特定のコンテナのメモリ占有によるシステム全体の性能劣化を回避したい場合に設定します。

但し、global zoneはシステム安定稼動のため未設定を推奨します。

✓ **rcapdによるメモリ上限設定は非推奨**

Solaris8 OS以降では、定期的にメモリ使用量を監視して閾値を超えた分をページアウトする、rcapd機能が提供されています。しかし、rcapdは、頻繁に発生するとディスクへのI/O負荷が高くなり、システム全体の性能に影響を与えることがありますので推奨されません。



Solarisコンテナの設計指針（３）

FUJITSU

■ディスクの設計指針

（１）容量見積もり

通常のOS領域に加えて、zone利用分のディスク容量を見積もります。

【zone01つ当りのディスク容量の目安】

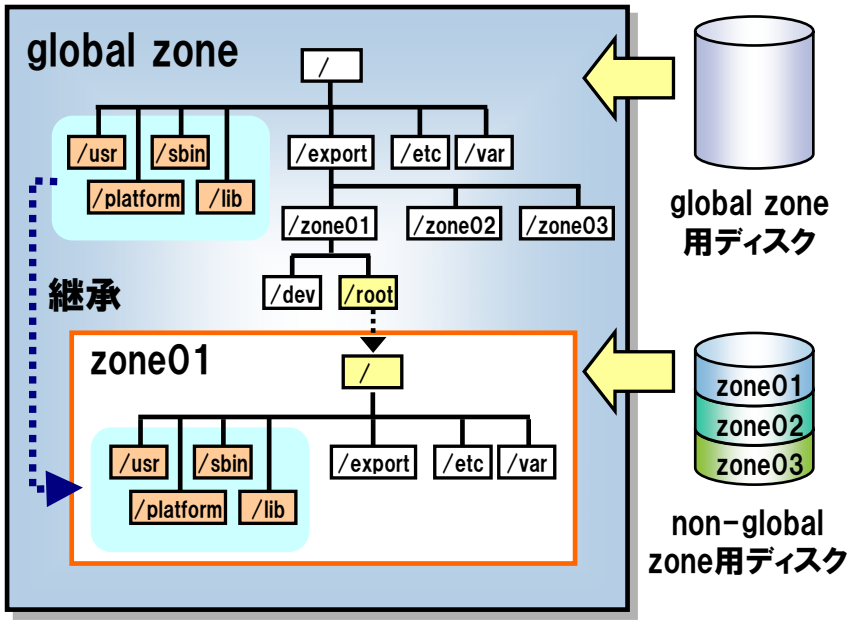
・ファイルシステム（/usr,/sbin,/lib,/platform）をglobal zoneと、
　共有（継承）する場合　⇒ **約220MB**
　共有（継承）しない場合 ⇒ **約4GB**

ファイルシステムの継承とは…

読み取り専用でglobal zoneのファイルシステムを共有すること。デフォルト設定で zoneを作成すると /usr、/sbin、/lib、/platform ディレクトリが継承されます。

アプリケーションやミドルウェアの中には、インストール時やパッチ適用時に継承ディレクトリへの書き込みが発生するため、継承設定を解除する必要があります。

注意!　**継承ディレクトリはzoneインストール後に変更不可のため事前設定が必要です。**



（２）ディスク領域の分割

non-global zoneはglobal zoneと同じディスク上に作成可能ですが、I/O性能や信頼性の面から、non-global zoneは別スライス、または別ディスク上に作成することを推奨します。

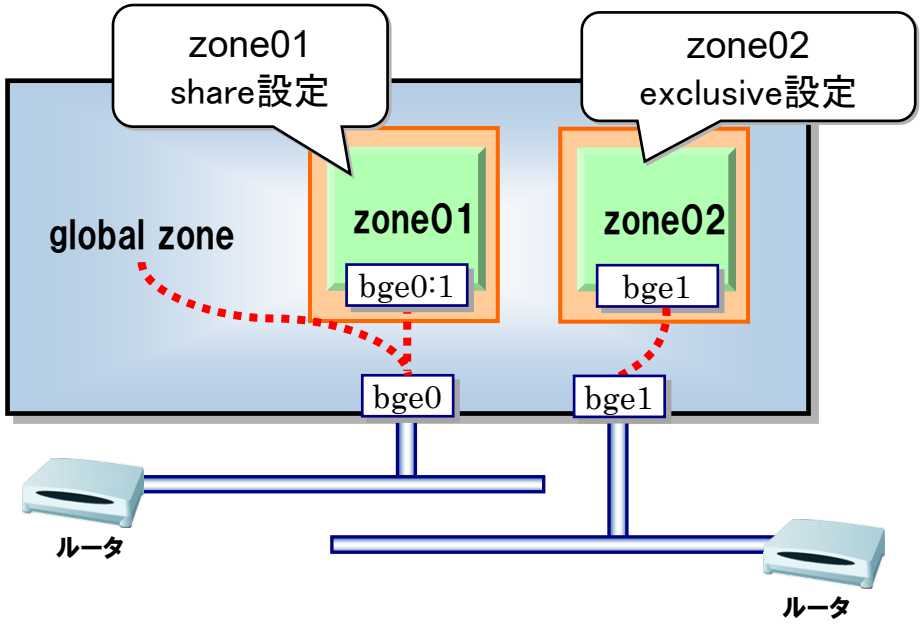
Solarisコンテナの設計指針（４）

■ネットワークの設計指針

（１）2種類のNIC設定

zoneのNIC設定はglobal zoneのNICを「共有(share)」するか、zoneで「占有(exclusive)」するかのどちらかに設定します。デフォルト設定は「共有」です。異なるセグメントのzoneを統合するなど、global zoneと異なるネットワーク構成にする場合は「占有」設定にします。

共有設定の場合、global zoneの物理NIC（例：bge0）から仮想NIC（例：bge0：1）がzoneに割り当てられます。占有設定の場合、zone専用に物理NIC(例：bge1)が割り当てられ、global zoneと異なるネットワーク体系を構成することが可能です。

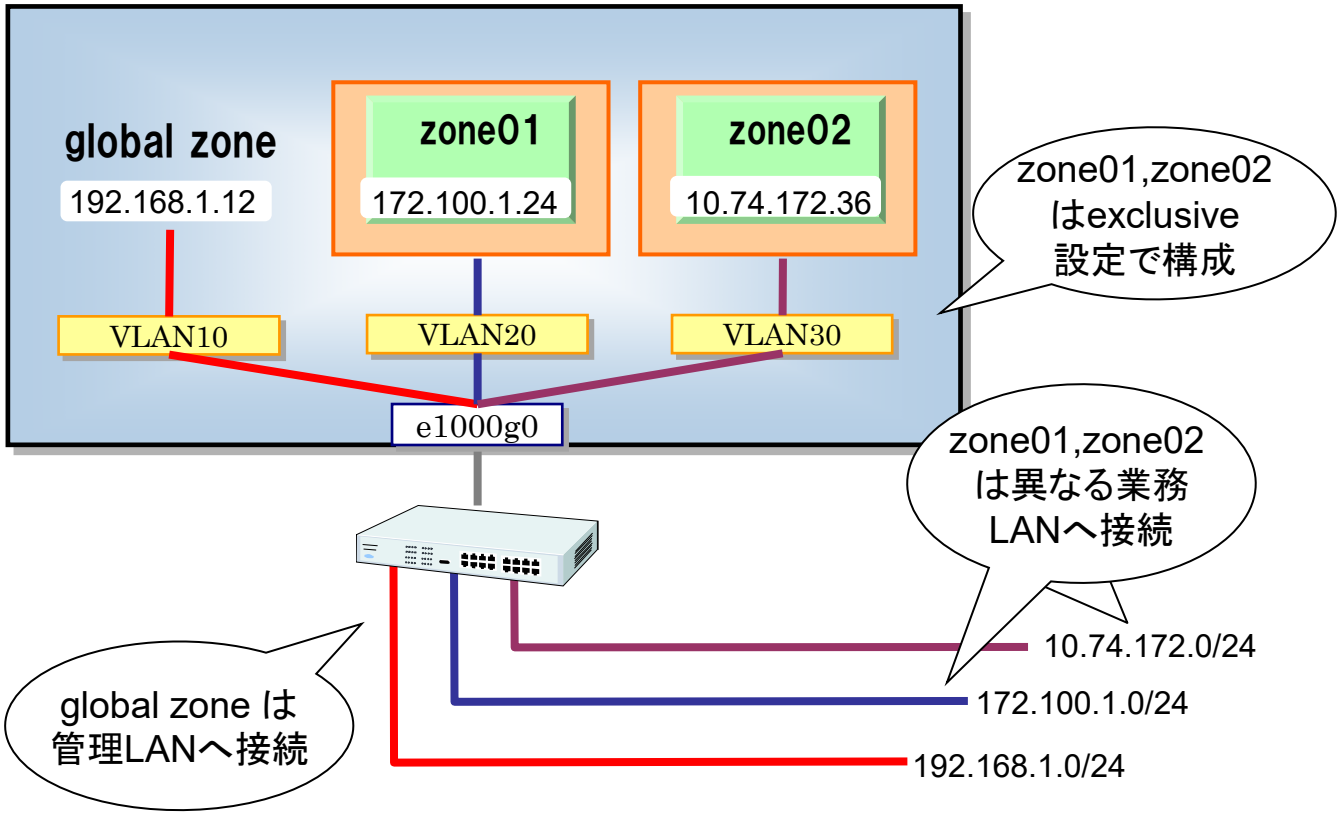


✓ ネットワークの占有設定は、Solaris 10 8/07以降、かつ
GLDv3対応のNICで構成する必要があります。
富士通製NICの場合「FUJITSU PCI Gigabit Ethernet 4.0」以降のドライバを適用します。

Solarisコンテナの設計指針（５）

（２）VLAN機能のコンテナ利用

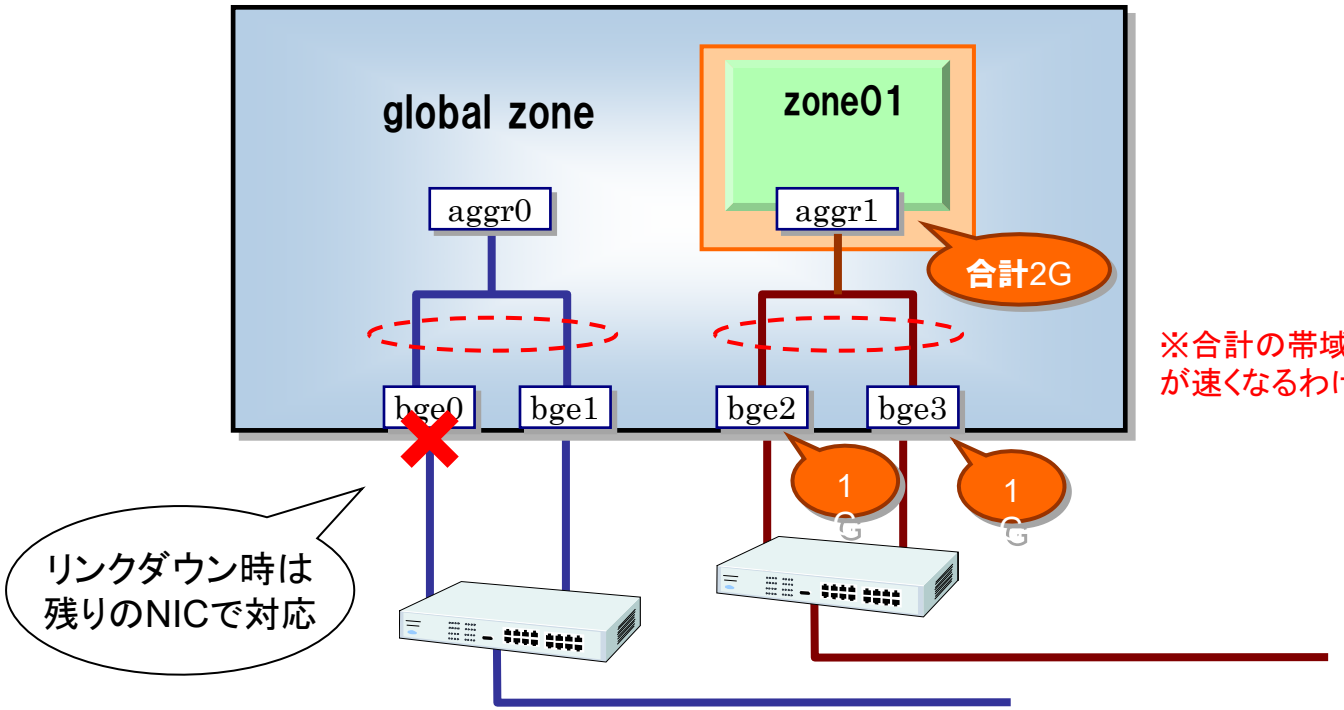
VLANを利用すると、物理NICを論理的に複数のネットワークに分割することが可能です。
複数のコンテナを統合した環境においても柔軟なネットワーク構成が可能となります。
物理NICの不足する場合などに有効です。



Solarisコンテナの設計指針（６）

（３）Link Aggregation機能のコンテナ利用

複数回線を束ねて、仮想的に1本のネットワークとして扱えるもの。束ねた回線の帯域を合計した量の帯域を使用可能です。障害発生時には、片系のNICのみダウンさせ業務の継続が可能です。



※合計の帯域幅が増えるだけであり速度が速くなるわけではないので注意

Solarisコンテナの設計指針（７）

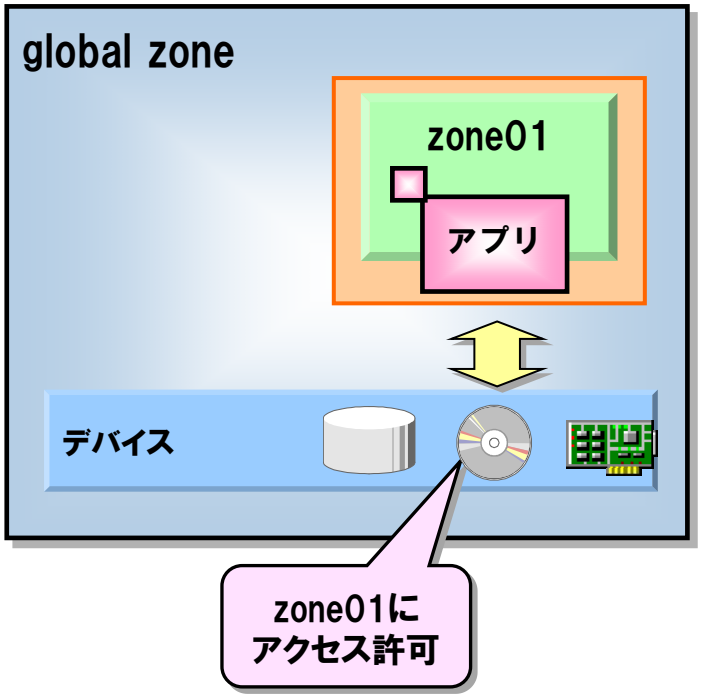
■その他のデバイス(DVD-ROM、DATなど)の設計指針

（１）non-global zone から物理デバイスへのアクセス

global zoneから許可したデバイスのみアクセスが許可されます。global zoneで仮想デバイス（/dev/lofi）を作成しアクセス許可することも可能です。

コンテナにミドルウェア等をインストールする場合には、CD-ROM装置の共有を行います。

※global zoneとnon-global zone間でのNFSマウントは不可



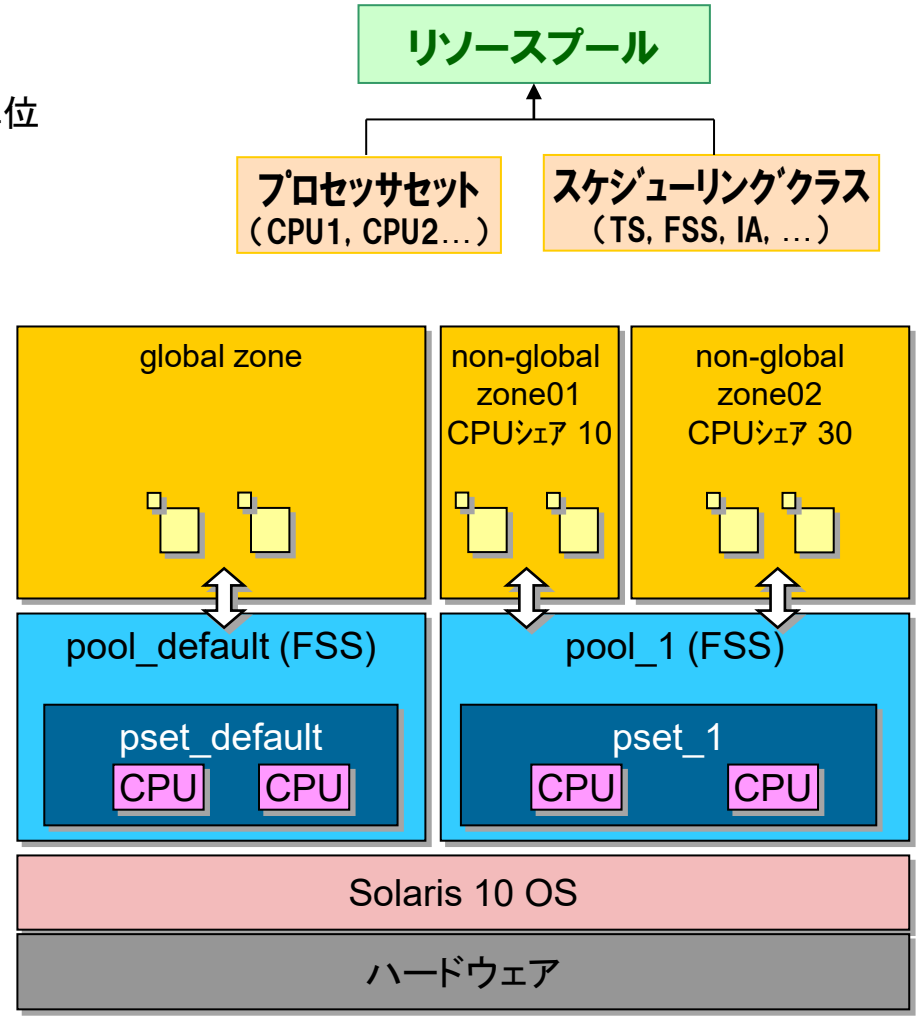
※デバイスを複数のzoneからアクセスさせることは、セキュリティ面で問題となるため利用アプリ側で特に規定しない限り非推奨です。

Solarisリソースマネージャ概要

FUJITSU

■zoneのプロセスが使用するCPU資源を制限・管理します

- **リソースプール（pool）**
  - システム資源をグループ化し、ゾーンに割当てる単位
  - プロセッサセットとスケジューリングクラスで構成
- **プロセッサセット（pset）**
  - CPUのグループ単位。格納するCPU数を設定。
- **スケジューリングクラス**
  - **FSS（Fair Share Scheduler）**
    - 比率（CPUシェア数）に従いプロセスへのCPUリソース配分を行なうスケジューリングクラス
    - ゾーン単位に比率（CPUシェア数）を設定
  - **TS（Time Sharing）**
    - 実行可能なプロセスに平等にCPUリソースを配分するスケジューリングクラス
    - Solarisデフォルトのスケジューリングクラス
- **Solaris Zone**
  - リソースプールを結合する単位
  - global zoneのリソースプールは固定（pool_default）



リソースマネージャによるリソース制御（１）

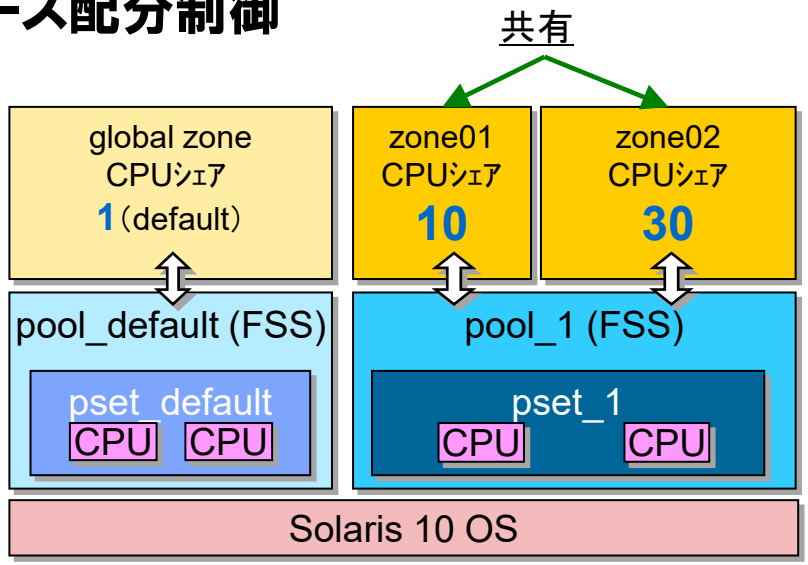
■リソースプールを共有するzone間のCPUリソース配分制御

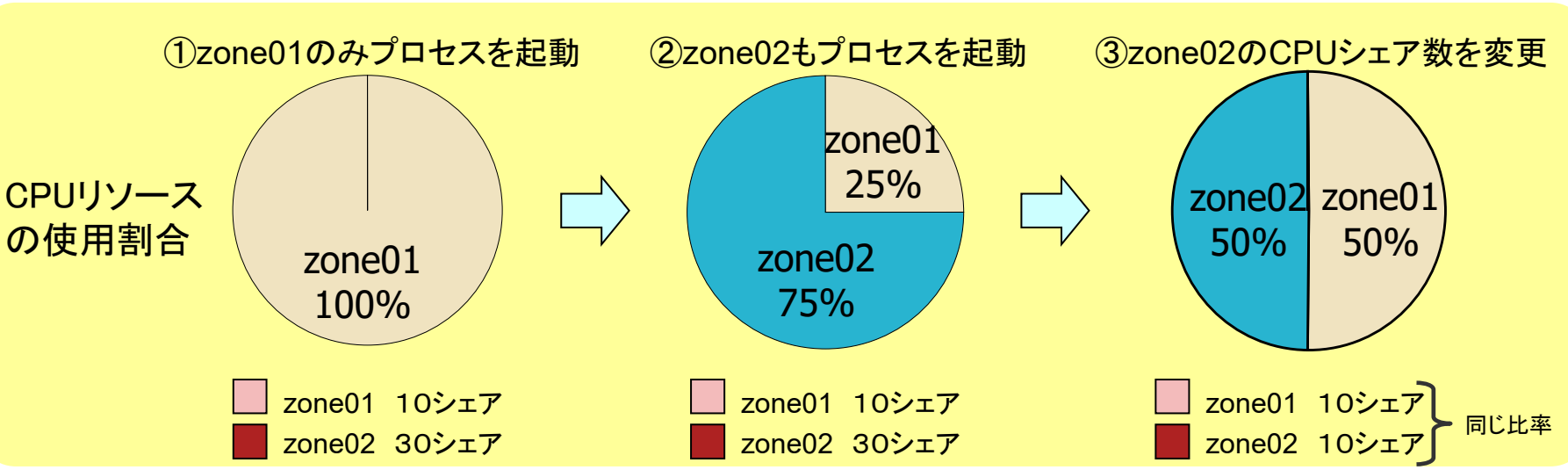
CPUシェア設定によるリソース配分のしくみ

① CPUリソースが空いている場合は利用可能な全てのリソースを利用可能

② zone間でCPUリソースの競合が発生した場は、CPUシェア数による比率配分を実行

③ CPUシェア数を動的に変更してリソース配分の制御を実行

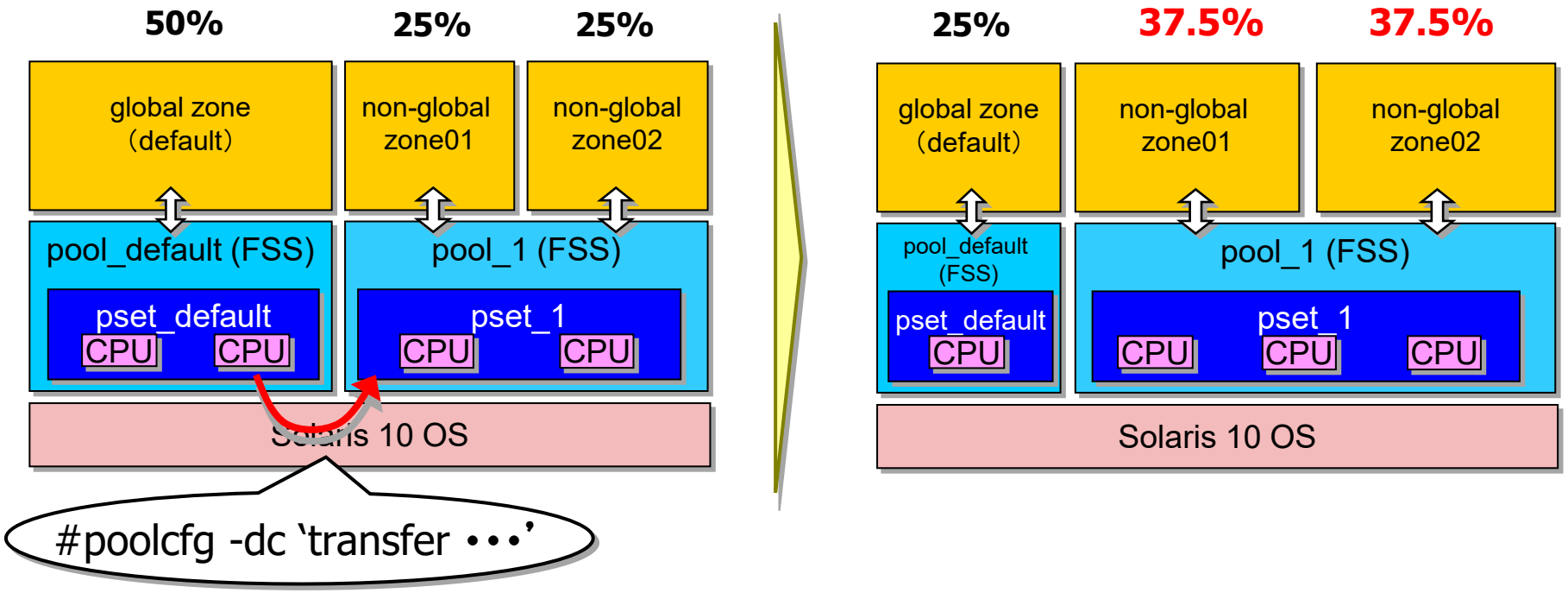




リソースマネージャによるリソース制御（２）

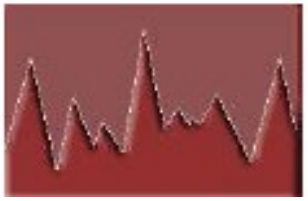
リソースプール間で動的にCPUリソースを移動する制御（手動）

✓コマンド1行で即時に構成変更が可能
✓コンテナ上の業務は停止せずに実行可能

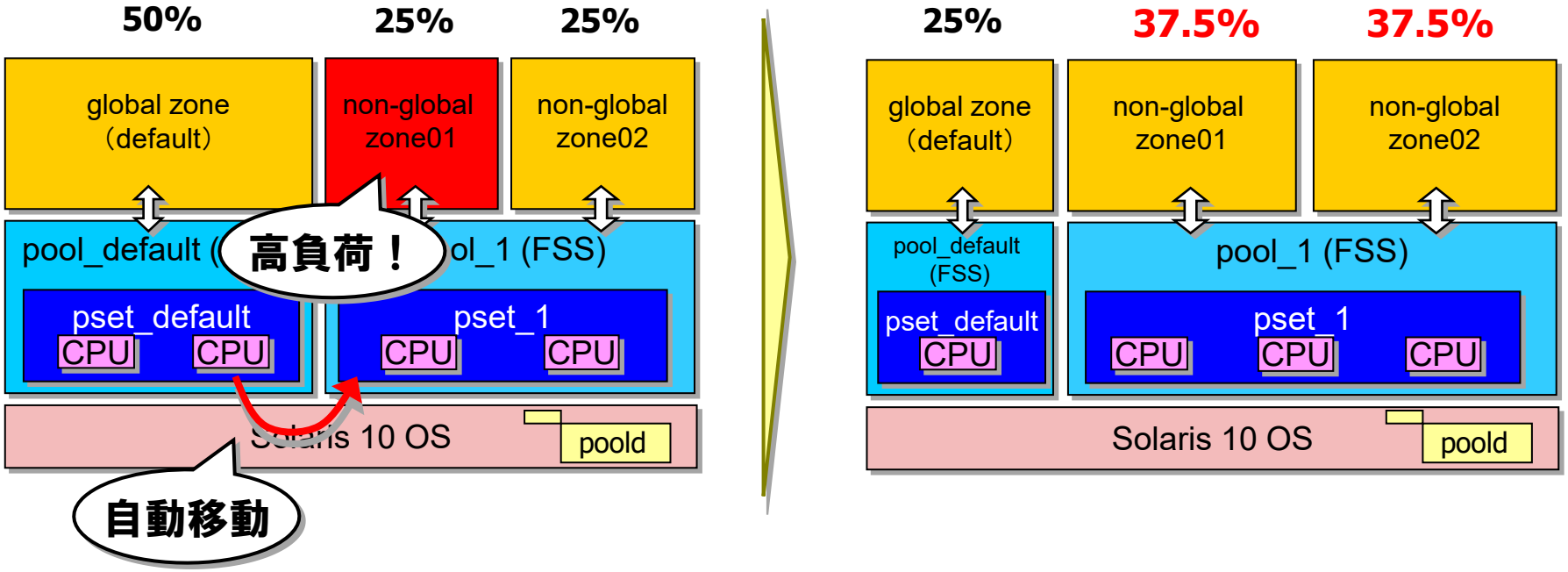


リソースマネージャによるリソース制御（３）

リソースプール間で動的にCPUリソースを移動する制御（自動）



✓pooldデーモンがリソースの使用状況を定期的に監視
✓予め設定したリソース使用範囲内でCPUを自動移動



Solaris 10 8/07**以降では、下記サービスを起動するだけで自動制御が有効になります（デフォルト：無効）**
サービス名「svc:/system/pools/dynamic:default」

リソースプールの設計指針

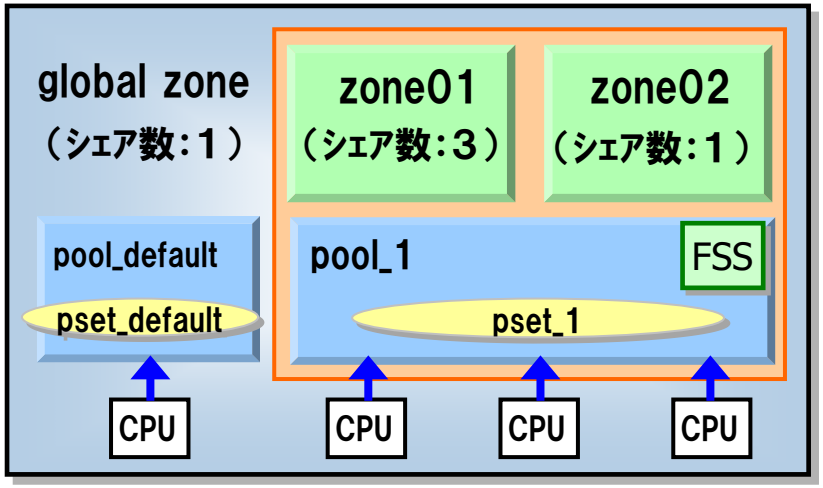
（１）global zoneとnon-global zoneのプールは独立させる

global zoneのCPUリソースを確保し安定稼動させるため、non-global zoneのプールは別に構成することを推奨します。

（２）プールのスケジューラは全てFSSに設定する

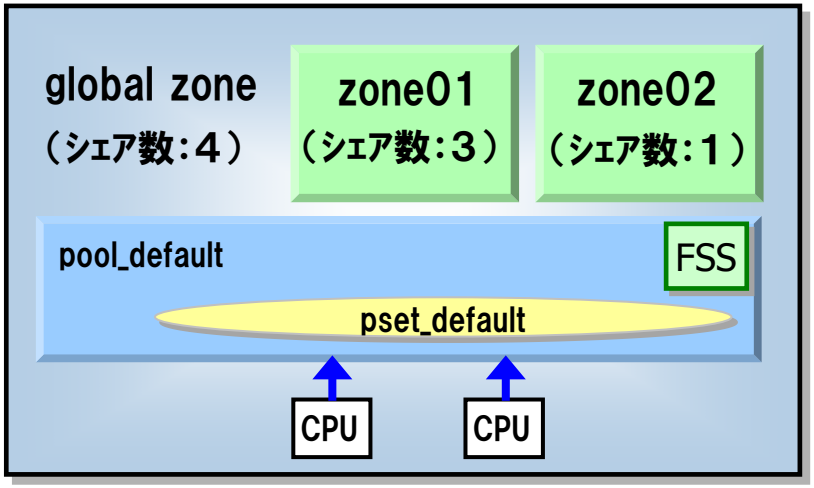
スケジューラをFSSに設定することで、CPUシェア数の指定によるリソースの比率配分が可能となります。zone用のリソースプールを作成しない（pool_defaultを共有する）場合は、global zoneのCPUシェア数を一番高く設定します。（zoneの負荷がglobal zoneに影響を与えないようにするため。）

①zone用のリソースプールを作成する場合



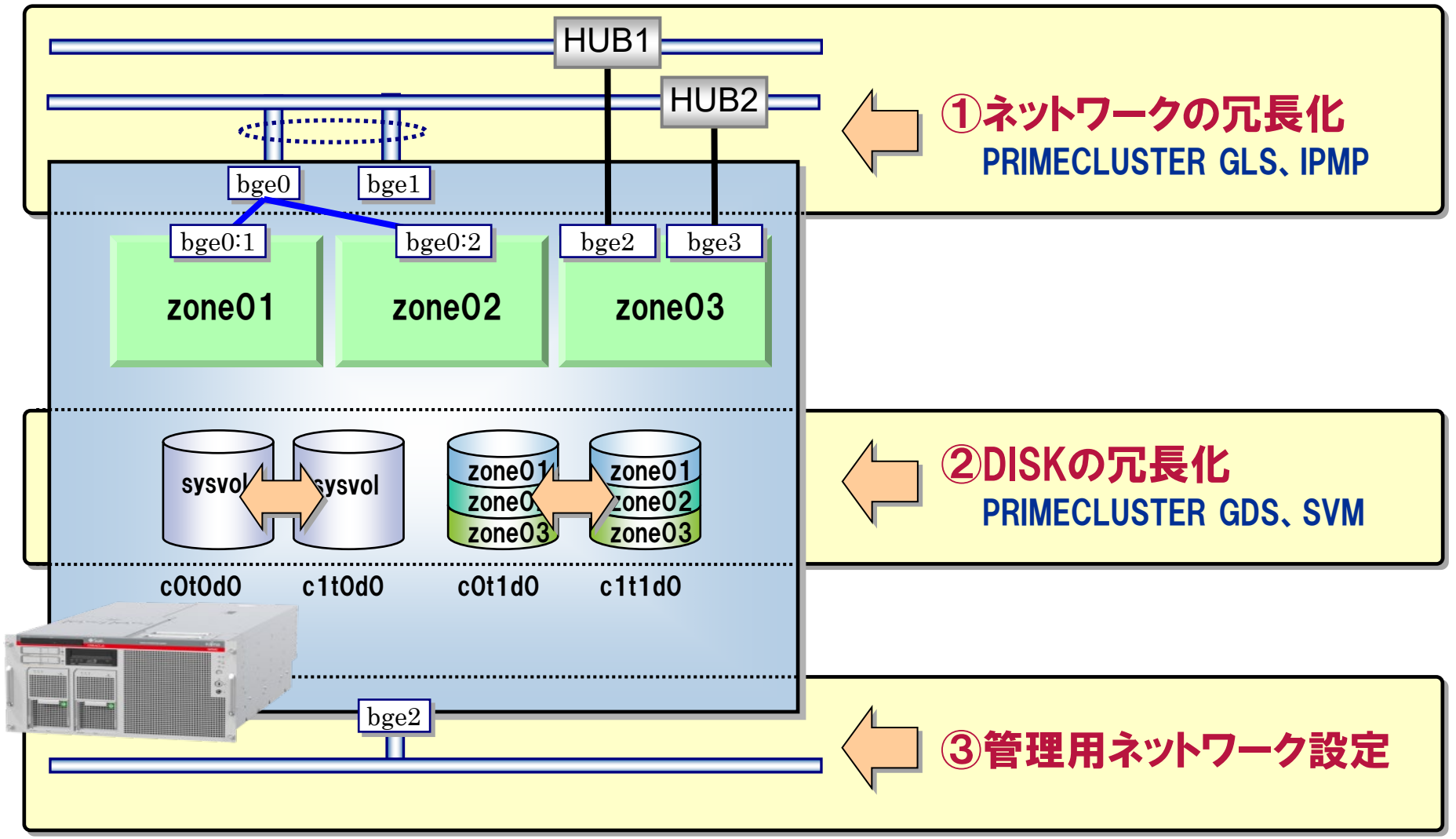
zone01とzone02は３：１のリソース配分となります

②zone用のリソースプールを作成しない場合



global zone、zone01、zone02は４：３：１のリソース配分となります

Solarisコンテナの高信頼化設計



①ネットワークの冗長化1/2

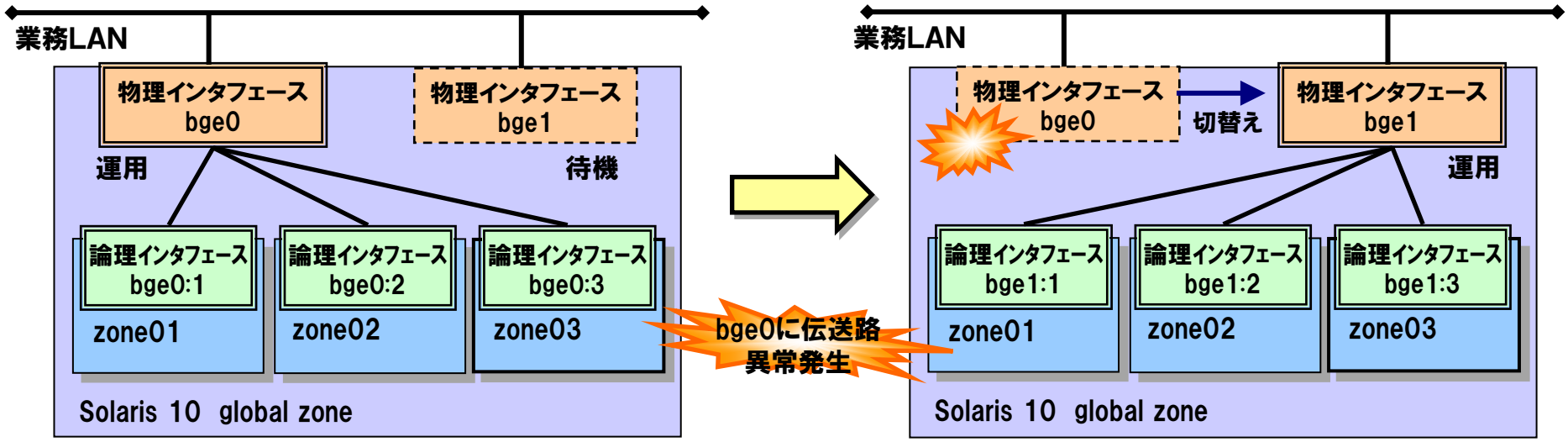
・PRIMECLUSTER GLSによりコンテナの仮想NICの冗長化が可能

<適用効果>

・物理NIC障害による業務停止時間を短縮することが可能です。

<留意事項>

・複数コンテナ環境の場合、一組のGLS構成を共有するためNIC障害発生時は、全てのSolarisコンテナの論理インタフェースが切替わります。



①ネットワークの冗長化2/2

FUJITSU

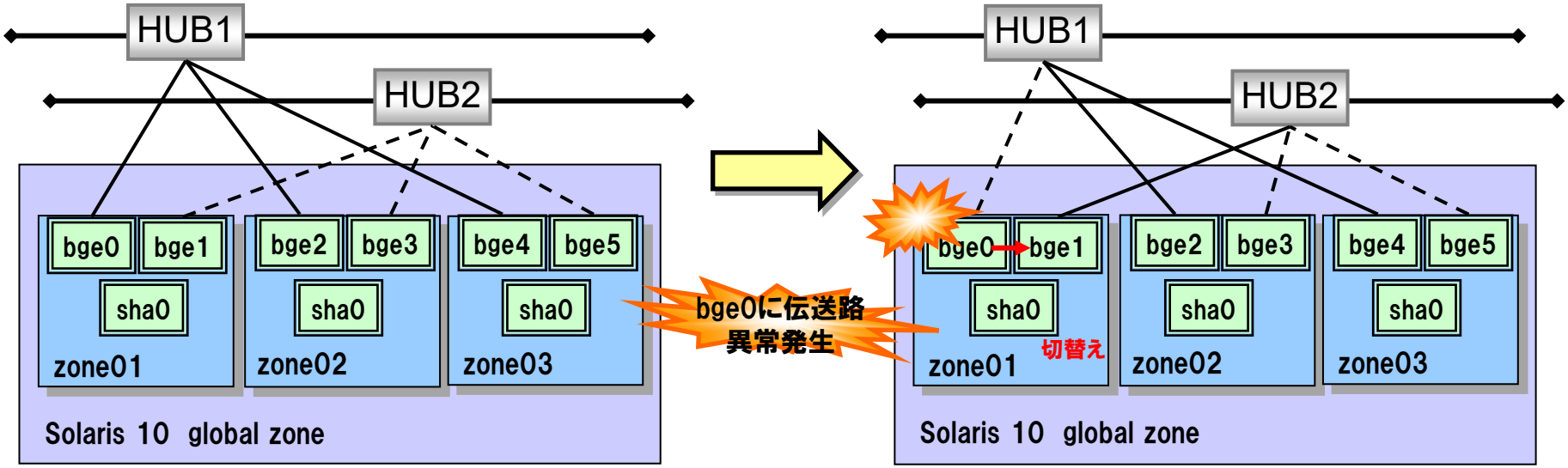
・PRIMECLUSTER GLSによりコンテナに割り当てた物理NICの冗長化が可能

<適用効果>

・物理NIC障害による業務停止時間を短縮することが可能です。
・コンテナ内の物理NICを切替えるため、他のコンテナには影響を与えません。

<留意事項>

・zoneのNIC設定にはexclusiveで設定するある必要があります。
・複数のコンテナ環境の場合、コンテナ環境数に応じて物理インターフェースが複数必要になります。



②ディスクの冗長化

・PRIMECLUSTER GDSによりコンテナのディスク領域の冗長化が可能

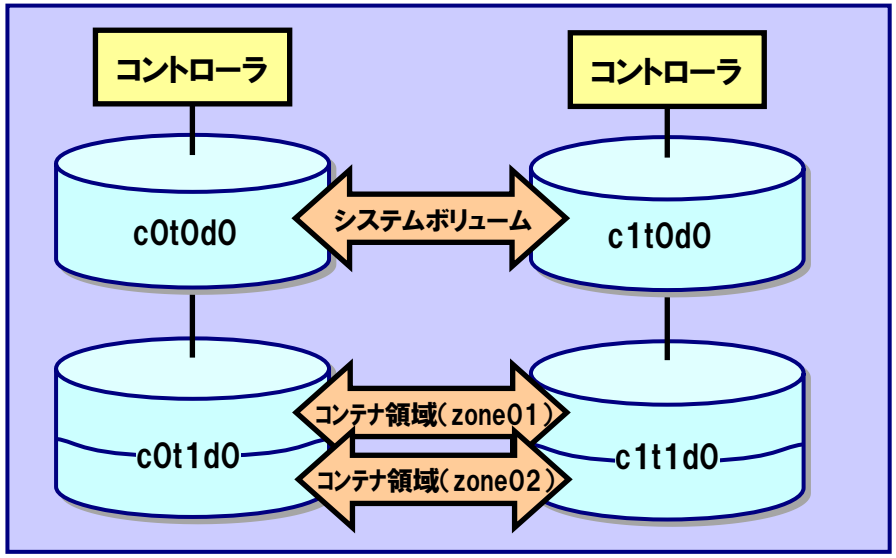
<適用効果>

・ディスク障害によるデータ損失および業務停止を回避することが可能です。

<留意事項>

・ミラーリングを行うディスクのペアは、コントローラ障害を考慮し、それぞれ異なるコントローラに接続されているディスクで構成してください。

・GDSで作成したボリュームをSolarisコンテナからデータベース領域（rawデバイス）として利用することも可能です。（※）



③管理用ネットワーク設定

FUJITSU

・global zoneとnon-global zoneを異なるネットワークアドレスに構成することが可能

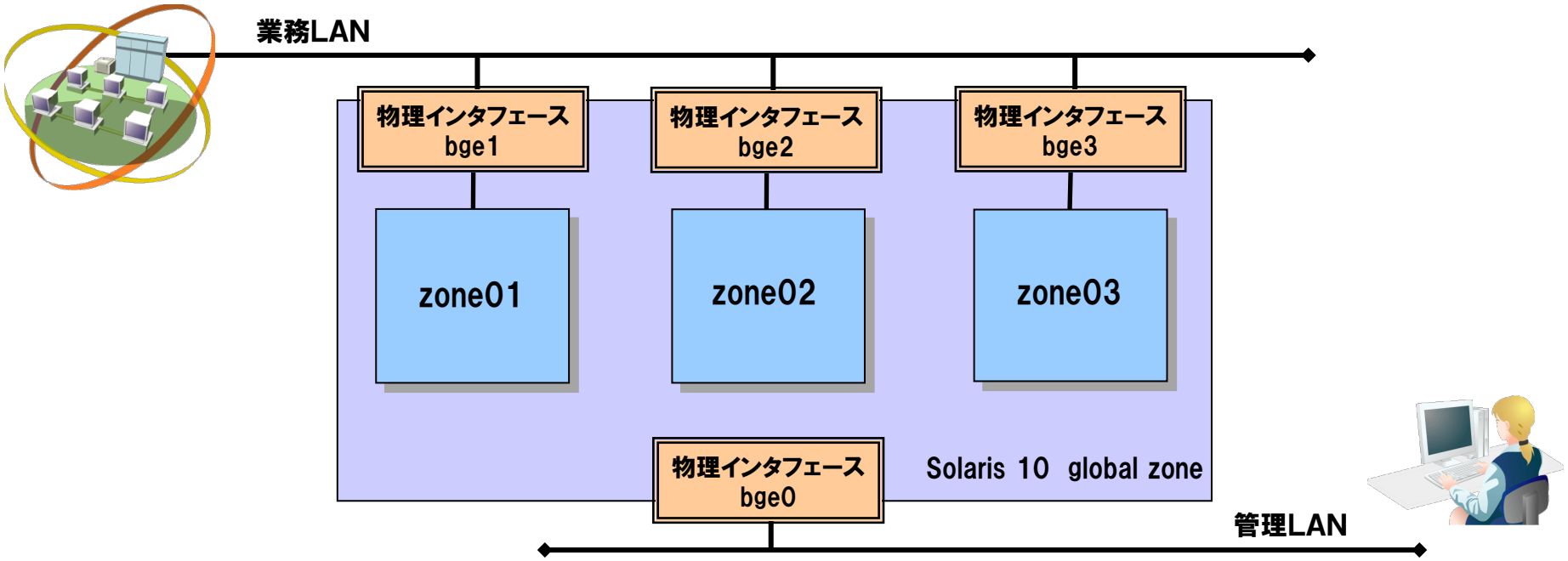
<適用効果>

・管理用ネットワークを構成することで、global zoneのセキュリティが高まります。

<留意事項>

・Solaris 10 8/07以降、かつGLDv3ﾃﾞｰﾀﾘﾝｸ対応のNICで構成する必要があります。

・VLAN構成の場合はzone毎に物理インタフェースは不要です。



商標について

FUJITSU

使用条件

- 著作権・商標権・その他の知的財産権について
  コンテンツ(文書・画像・音声等)は、著作権・商標権・その他の知的財産権で保護されています。本コンテンツは、個人的に使用する範囲でプリントアウトまたはダウンロードできます。ただし、これ以外の利用(ご自分のページへの再利用や他のサーバへのアップロード等)については、当社または権利者の許諾が必要となります。
- 保証の制限
  本コンテンツについて、当社は、その正確性、商品性、ご利用目的への適合性等に関して保証するものではなく、そのご利用により生じた損害について、当社は法律上のいかなる責任も負いかねます。本コンテンツは、予告なく変更・廃止されることがあります。

商標

- 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。
- SPARC Enterprise、SPARC64およびすべてのSPARC商標は、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。
- OracleとJavaは、Oracle Corporation およびその子会社、関連会社の米国およびその他の国における登録商標です。
- 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。

FUJITSU