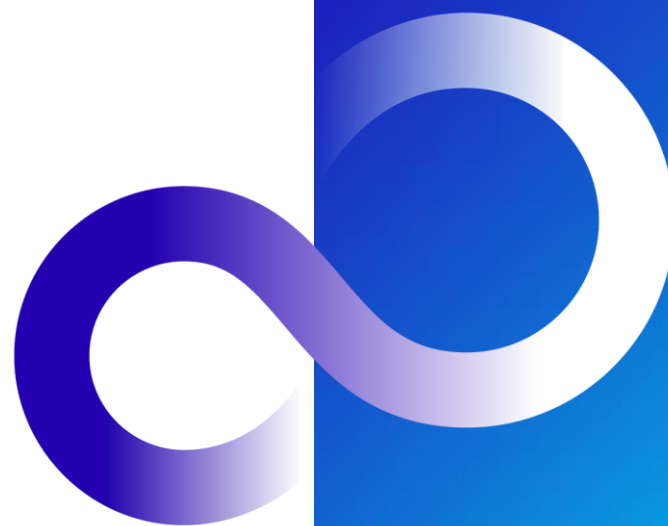


SPARC M12/M10 Building Blockを活用した 高可用性システム(概要編)

2023年5月

第5版

富士通株式会社



はじめに

1. 高可用性システムの構成パターン

2. 高可用性システム構成例 (*)

3. 故障発生時の動き (*)

4. 保守手順の概要 (*)

5. 留意事項

参考. 高可用性を実現するSPARC M12/M10の機能詳細

(*) 2、3、4項については、集約型の構成を代表例として挙げ説明しています。

■ 目的

- 本書は、Building Blockを活用した高可用性システム(BB High Availability: BB HA)の機能、メリット、設計指針について紹介します。

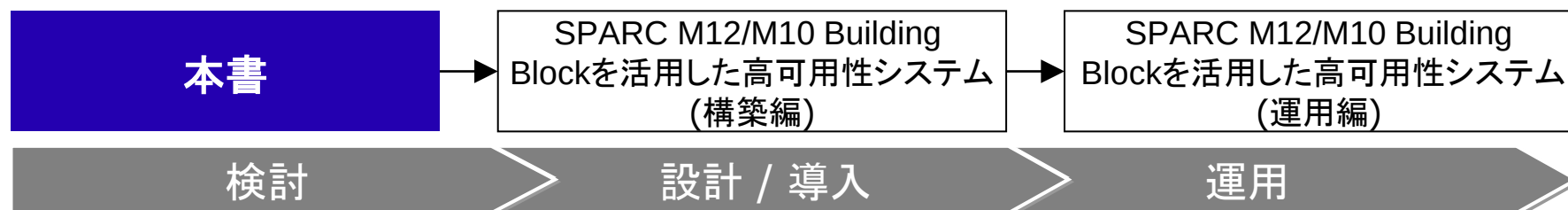
■ 対象読者

- SPARC M12/M10を使用して高可用性システムを構築したいと考えている方
- Oracle Solaris /Oracle VM Server for SPARCの基礎知識を有している方

■ 留意事項

- 本書では「Oracle Solaris」を「Solaris」,「Oracle VM Server for SPARC」を「Oracle VM」または「OVM」と記載することがあります。
- 本書ではBB HAシステムの概要を紹介します。具体的な設計や、導入手順に関しましては、「SPARC M12/M10 Building Blockを活用した高可用性システム(構築編)」をご参照ください。また、運用後の保守手順につきましては「SPARC M12/M10 Building Blockを活用した高可用性システム(運用編)」をご参照ください。

■ 本書の位置づけ



「Building Block機能を活用した高可用性システム(BB HAシステム)」の特長

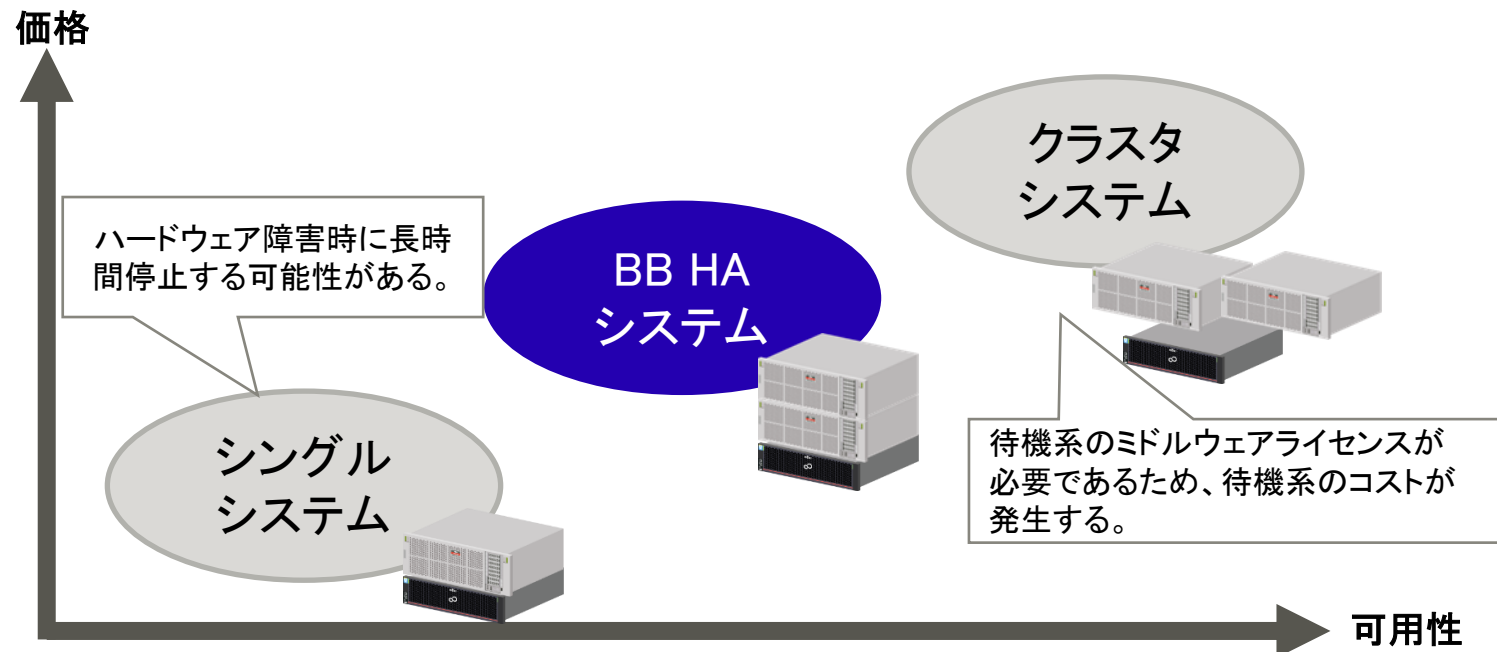
SPARC M12-2S/M10-4Sなら可用性を高めながらコスト削減が可能

ハードウェア障害時にリブート後、故障部分を切り離して
業務復旧。故障部分の活性保守が可能

可用性向上

待機系のミドルウェアライセンスが発生しない

コスト削減



■ 可用性を高めながらコスト削減が可能

- 可用性をより重視する場合はクラスタソフトを導入したシステムが優位
- 1時間程度のダウンタイムを許容できる場合はコスト面で本構成(BB HA)が優位

		シングルシステム	BB HAシステム	クラスタシステム
可用性	監視対象	× なし	○ ハードウェア異常	◎ ハードウェア異常、 OS異常、アプリ異常
	障害発生時の 業務復旧目安	× 手動復旧	○ 数分～1時間	◎ 数秒～数分
	ハード保守時の ダウンタイム	× 停止交換	◎ PPAR DR機能*により 停止時間は数分	◎ サーバーの切り替えにより 停止時間は数分
コスト	ミドルウェア	◎ 運用系のみ	◎ 運用系のみ	△ 運用・待機系
	クラスタソフト	◎ 不要	◎ 不要	× クラスタソフト必須

(*) PPAR Dynamic Reconfiguration(DR)機能: 詳細は後述(P 32)

■ 高可用性を実現するSPARC M12/M10の機能

■ Building Block

- ・ 複数の筐体を組み合わせ、ハードウェアを拡張
- ・ BB単位に切り離し、組み込みが可能
 - パーティションの構成変更が可能

■ CPUコア アクティベーション

- ・ 業務負荷に応じた動的なリソース増強
- ・ 未アクティベートなCPUコアを故障コアと自動交替
 - ハードウェア故障時の自動復旧

■ Dynamic Reconfiguration

- ・ システム無停止のリソース拡張
- ・ ハードウェア保守時のダウンタイム削減
 - 故障発生時のダウンタイムを極小化



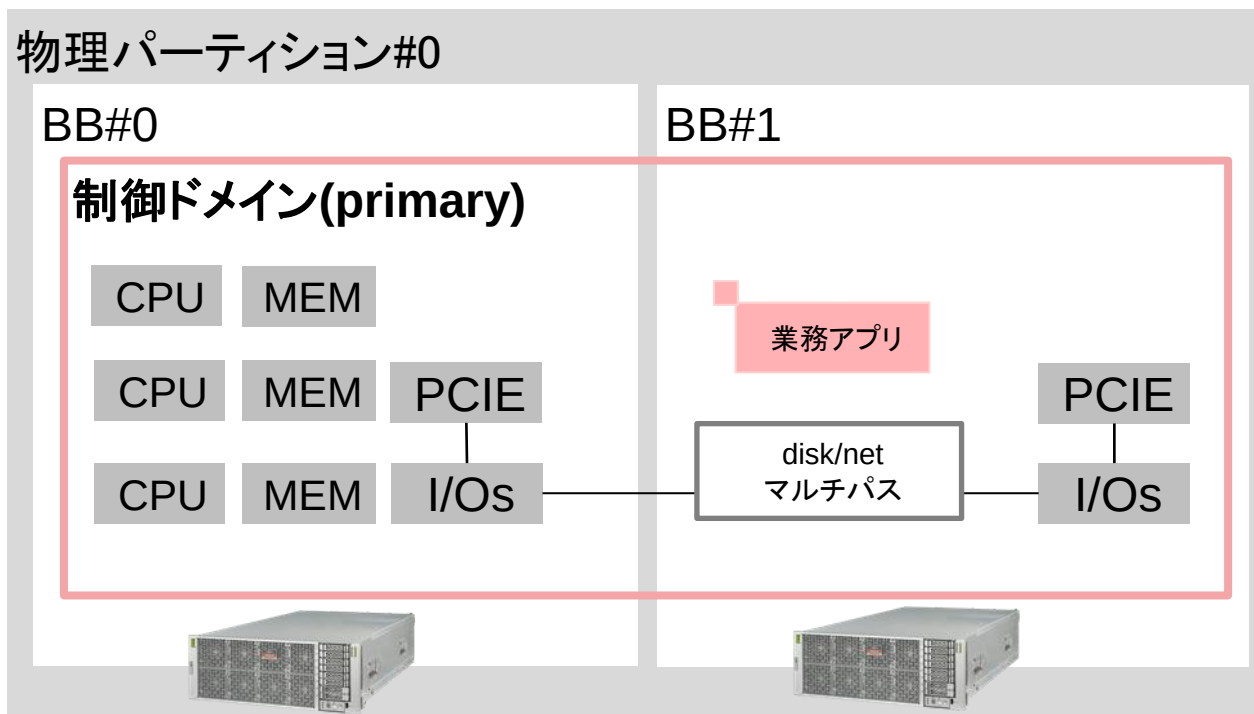
本書では、SPARC M12-2S/M10-4SのBuilding Block構成とPPAR DR機能を組み合わせることで、システムのダウンタイムを極小化させる構成について解説しています。

1. 高可用性システムの構成パターン

タイプ	概要	ドメイン構成	構築・運用の容易性
従来型	1つの物理パーティションで1つの業務を行う。	制御ドメインのみの構成。 Zoneを使用し複数業務の集約も可能。	易
集約型	OVMを利用し1つの物理パーティションに複数の業務を集約しつつ、高可用性を確保する。 物理IOを各ドメインで占有して使用するため、IO性能を確保したい場合に推奨。	制御ドメインと複数のI/O ルートドメインのみの構成。 SPARC M12-2S/M10-4S 2台の場合、最大4 I/Oルートドメインまで構築可能 (I/Oルートドメイン当たり4枚のPCIカードを使用)。	
高集約型	OVMを利用し1つの物理パーティションに多数の業務を集約しつつ、高可用性を確保する。 仮想I/Oを利用することで、集約型よりもさらに多くの業務を集約させることができる。	制御ドメインに加えて複数のゲストドメインを構成。	

■ 従来型の特長

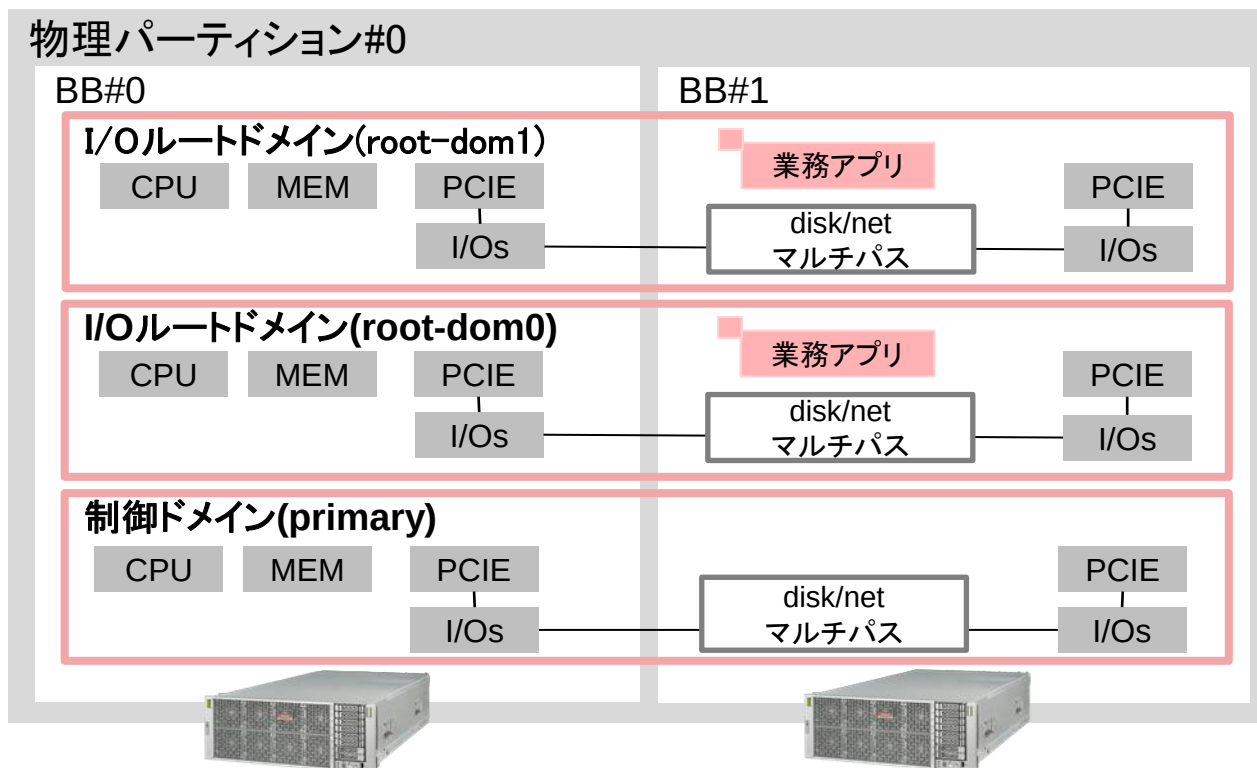
- ・ コア交替機能でCPU故障時もCPUリソースを確保
- ・ I/Oをマルチパスで冗長化
- ・ 万が一の故障発生時もBB単位で活性保守可能
- ・ **1つの物理パーティションで1つの業務を運用**



■ 集約型の特長

複数業務を1つの物理パーティションに集約しつつ可用性を確保

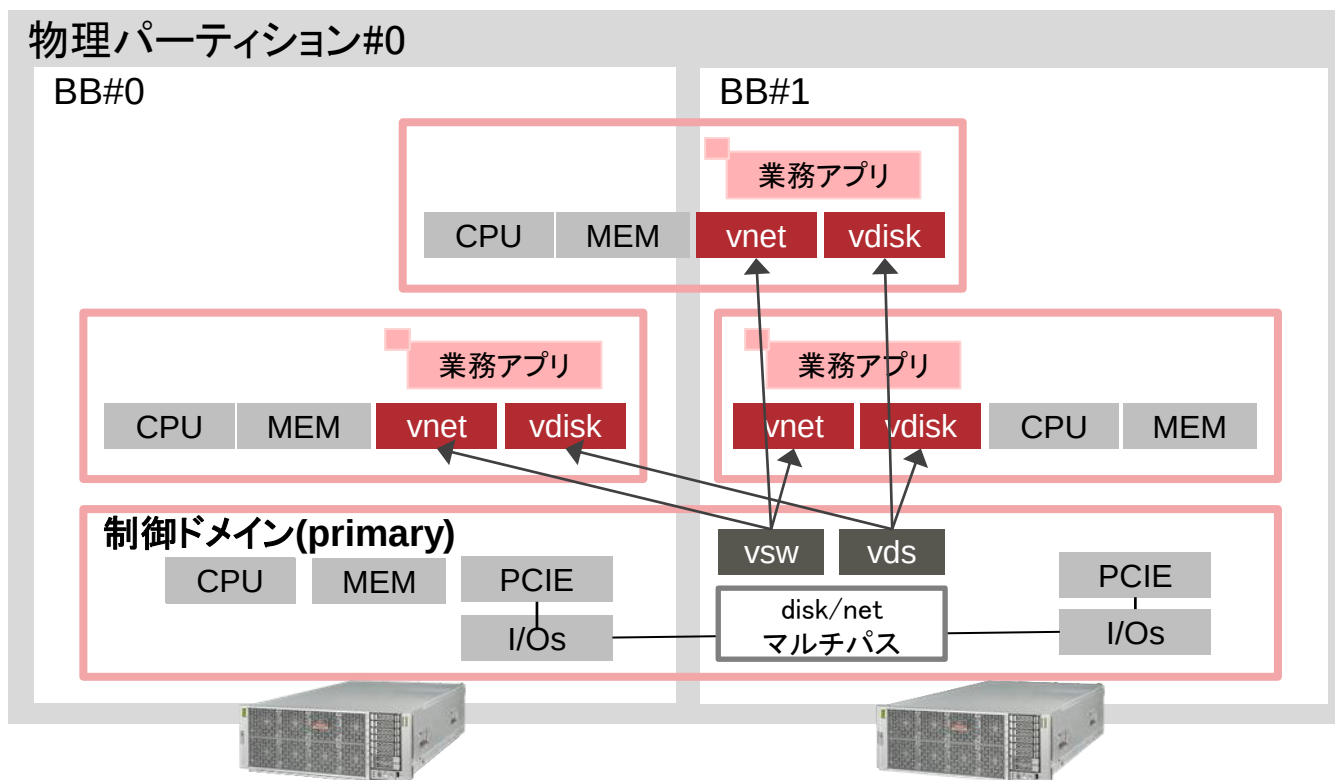
- コア交替機能でCPU故障時もCPUリソースを確保
- I/Oをマルチパスで冗長化
- 万が一の故障発生時もBB単位で活性保守可能
- I/O ルートドメイン上に業務アプリケーションを構築
- 複数業務を1つの物理パーティションに集約しつつ可用性を確保(最大4 I/Oルートドメイン)



■ 高集約型の特長

仮想I/Oを使用することで、高集約を実現しつつ可用性を確保

- コア交替機能でCPU故障時もCPUリソースを確保
- I/Oをマルチパスで冗長化
- 万が一の故障発生時もBB単位で活性保守可能
- ゲストドメイン上に業務アプリケーションを構築
- 複数業務を1つの物理パーティションに集約しつつ可用性を確保(4ドメイン以上集約可能)



■ ハードウェア条件

- 本体装置 : SPARC M12-2SまたはSPARC M10-4Sを2BBで構成
- I/Oカード : 2枚単位で冗長構成
- メモリ : 1.各BBの同じ位置には同じ容量のDIMMを搭載。
2.メモリグループの搭載容量は、以下の容量制限を満たす必要があります。
$$\text{CPU\#0 Group A} \leq \text{CPU\#0 Group B} \leq \text{CPU\#1 Group A} \leq \dots \leq \text{CPU\#3 Group B}$$

■ ソフトウェア条件

◆ サーバ:M12-2S(型番:SPNCCAA4xx)

- ファームウェア版数 : XCP 4020以降
- OS版数(全ドメイン) : Oracle Solaris 11.4 SRU22081(SRU11.4.48.126.1)以降

◆ サーバ:M12-2S(型番:SPNCCAA1xx~SPNCCAA2xx) M10-4S

- ファームウェア版数 : M12-2S: XCP 3021以降 M10-4S: XCP 2290以降
- OS版数(制御ドメイン) : Oracle Solaris 11.3 SRU17021(SRU11.3.17.5.0)以降
- OS版数(I/Oルートドメイン) : Oracle Solaris 11.3 SRU17021(SRU11.3.17.5.0)以降
- OS版数(ゲストドメイン) : Oracle Solaris 11.3 SRU17021(SRU11.3.17.5.0)以降
Oracle Solaris 10 1/13 パッチ:150400-07以降

最新情報は、マニュアル、プロダクトノートを確認してください。

<https://www.fujitsu.com/jp/products/computing/servers/unix/sparc/lineup/m12-2s/documents/>

<https://www.fujitsu.com/jp/products/computing/servers/unix/sparc/lineup/m10-4s/documents/>

■ 物理パーティション

- 物理パーティションは2台のBBで構成してください。

■ CPU

- CPUコア アクティベーション機能を利用し、物理パーティションに搭載されているCPUコアの半分以下を使用します。
 - これによりCPUコア故障時に自動交替機能が動作するようにします。
 - また活性交換時に業務に割り当てているCPU数を減らさずに保守できます。

■ メモリ

- 物理パーティションに搭載されているメモリの半分から、以下の表にあるHypervisorが使用するメモリ量を引いた容量をドメインに割り当てます。
 - これにより活性交換時に各ドメインのメモリ量を減らさずに保守できます。

	M12-2S	M10-4S
PPAR DR有効時	4GB	3.25GB
PPAR DR無効時	4GB	3GB

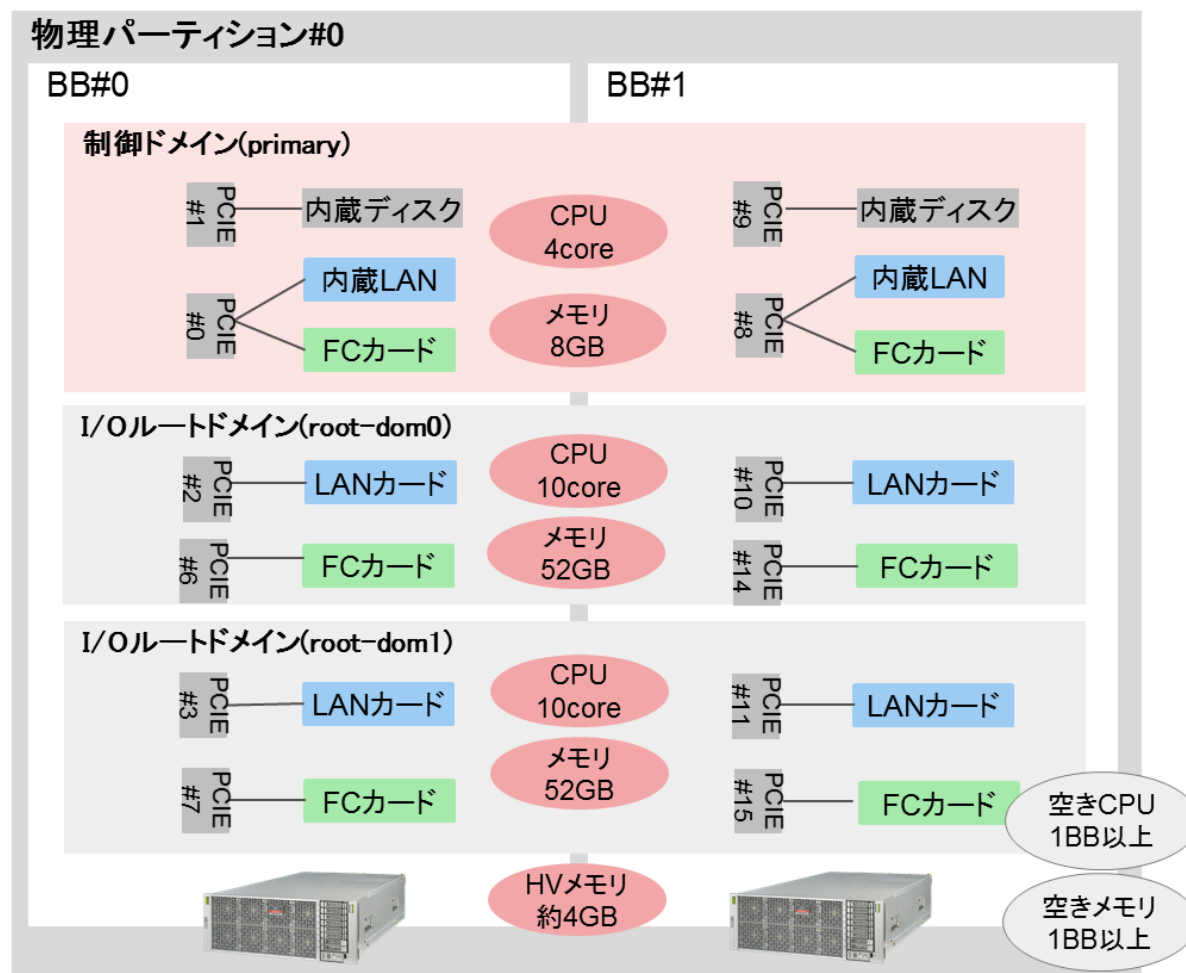
■ I/O

- 各ドメインのI/OをBBにまたがって冗長化します。
 - これにより片側のBBが起動しない場合でもI/Oの経路を確保できます。

2. 高可用性システム構成例 [集約型]

■ 複数業務を1つの物理パーティションに集約しつつ可用性を確保

- 保守時のCPU/メモリ維持のため、1BB分のCPU/メモリを予備リソースとして確保
- 各ドメインのI/Oをマルチパスで冗長化



■ 制御ドメイン

- CPU:1コア以上、メモリ:8GB以上 ※実際の環境に合わせて見積してください
- サービスドメイン設定
 - ・ 仮想コンソールサービス(vcc)のみ構築します。
- リソース割り当て
 - ・ CPUは割り当て時にコア数のみを指定します。
 - ・ メモリは割り当て時に容量のみ指定し物理アドレスは指定しません。
 - ・ 物理I/Oデバイスは各BBの内蔵ディスクを含むPCIe Bus(ルートコンプレックス) を割り当てます。
- システムボリューム
 - ・ 各BBのFCカードを冗長構成にして、制御ドメイン上で外部ディスクをマルチパス化します。
- NTPクライアント
 - ・ PPAR DR実行時にサスペンドが発生する場合があります。
この場合、時刻のずれが発生するため、そのずれをNTPにより補正します。

■ I/Oルートドメイン

■ CPU:1コア以上、メモリ:4GB以上 ※実際の環境に合わせて見積してください

■ リソース割り当て

- CPUは割り当て時にコア数のみを指定します。
- メモリは割り当て時に容量のみ指定し物理アドレスは指定しません。
- 物理I/OデバイスはPCIe Bus単位に指定します。

■ システムボリューム

- 各BBのFCカードを冗長構成にして、I/Oルートドメイン上で外部ディスクをマルチパス化します。

■ 業務アプリケーション

- I/O ルートドメインのOS環境で構築します。

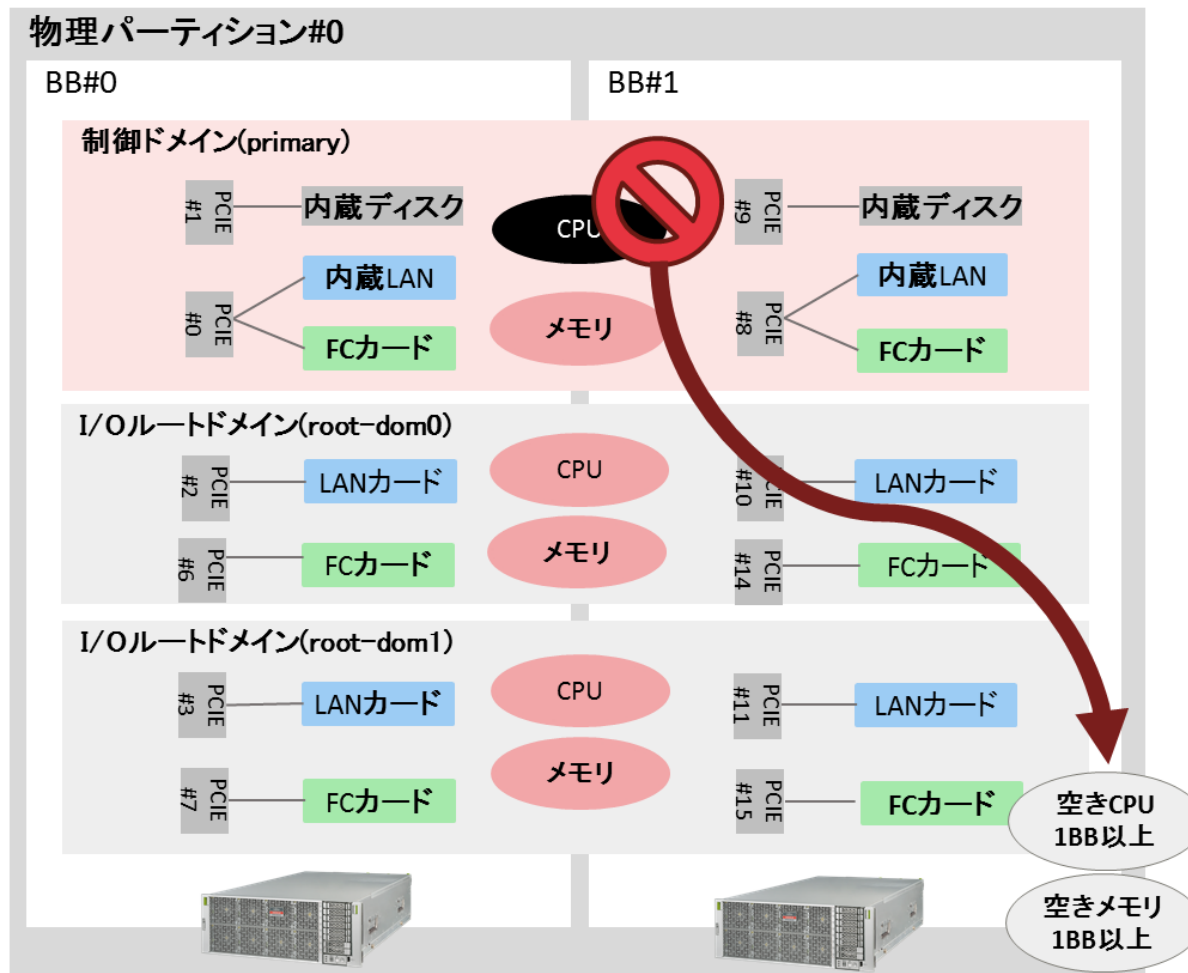
■ NTPクライアント

- PPAR DR実行時にサスペンドが発生する場合があります。この場合、時刻のずれが発生するため、そのずれをNTPにより補正します。

3. 故障発生時の動き[集約型]

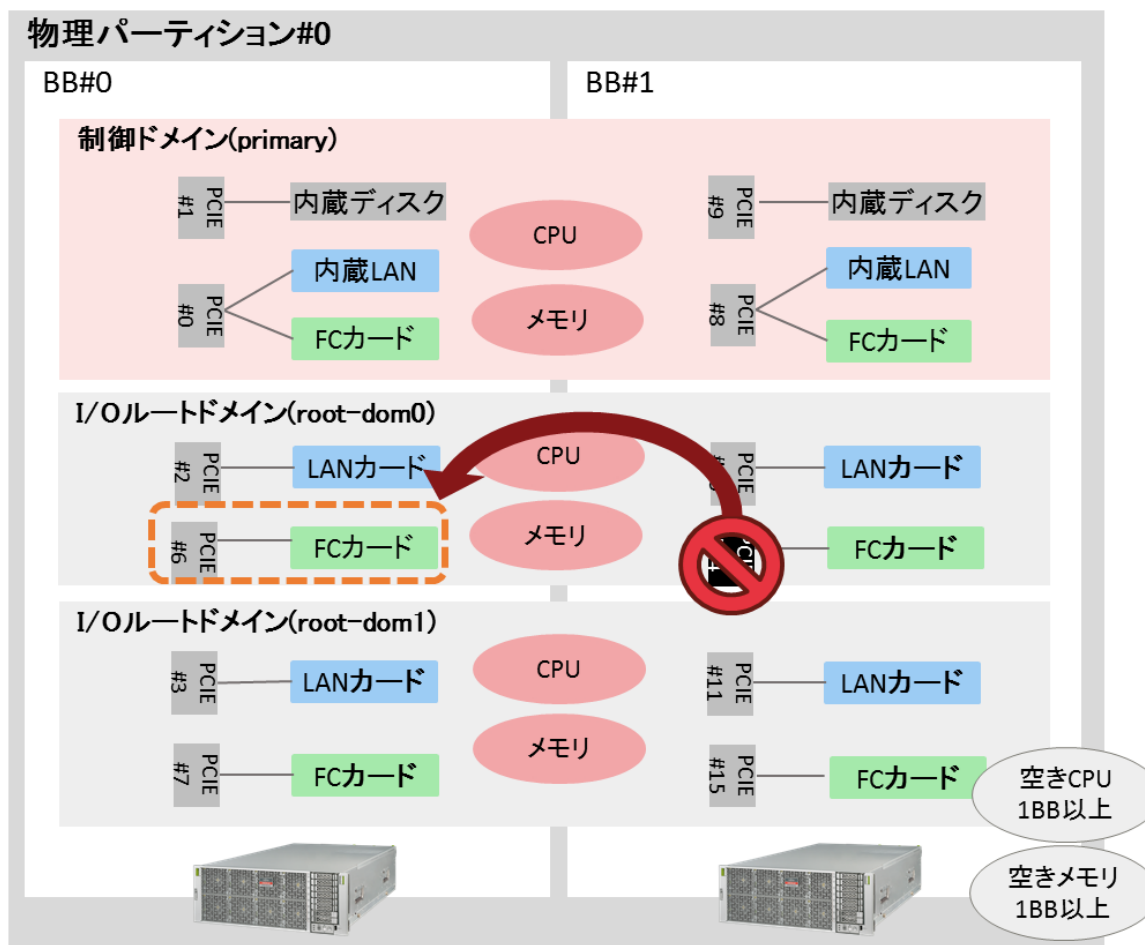
アクティベートされていないCPUコアがある場合、自動的に故障したCPUコアを正常なCPUコアと入れ替えます。入れ替えによるダウンタイムはありません。

(CPUコア故障によりOS Panicが発生した場合は、OSの再起動のため10分～15分のダウンタイムが発生します。)



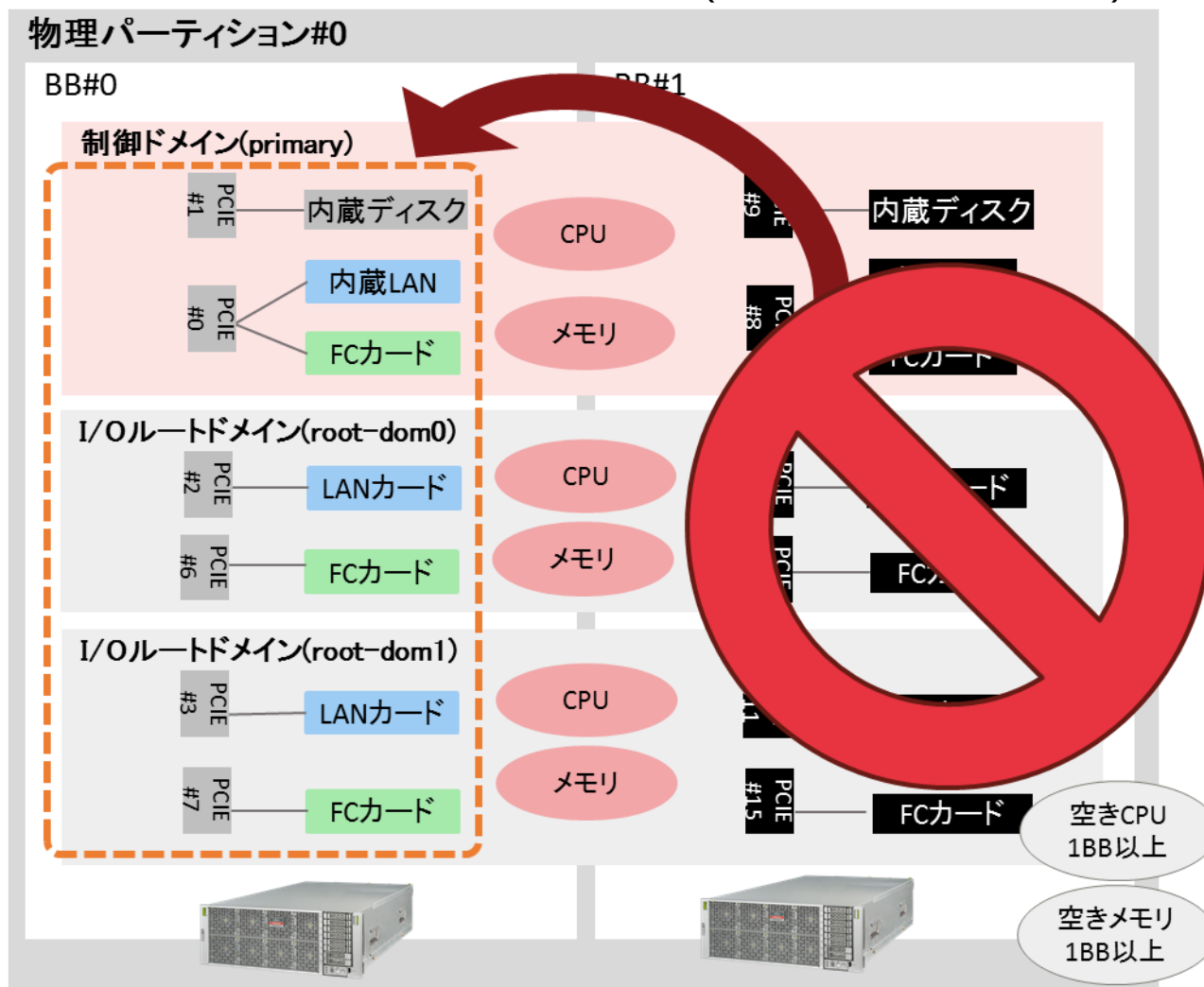
PCIe Bus故障発生時の動き

PCIe Busが故障した場合、OS Panicもしくは物理パーティションのリセットが発生します。
OS再起動後/物理パーティション再起動後に故障I/Oを切り離し、残りのI/Oを使用して
業務再開可能です。(ダウンタイム目安: OS再起動10分～15分/ 物理パーティション再起動 1時間)



筐体故障発生時の動き

1つのBBが使用不能となるような故障が発生した場合でも、物理パーティション全体が再起動した後に、recovery mode機能が働き、故障筐体側のI/Oを削除した構成(CPU数・メモリ量は維持)を自動生成し業務を再開させます。(ダウンタイム目安1時間)



4. 保守手順の概要[集約型]

■ 切り離し手順(2BB→1BB)の概要

- 制御ドメインの物理I/Oデバイスの冗長構成解除
 - (1) 切り離す側の物理I/Oデバイスのマルチパスを解除
 - (2) 切り離れた物理I/Oデバイスを未使用状態にする
- I/Oルートドメインの物理I/Oデバイスの冗長構成解除
 - (3) 切り離す側の物理I/Oデバイスのマルチパスを解除
 - (4) 切り離れた物理I/Oデバイスを未使用状態にする
- BBの切り離し
 - (5) XSCFのdeleteboardコマンドを実行し、BBを物理パーティションから切り離す

■ 組み込み手順(1BB→2BB)の概要

- BBの組み込み
 - (1) XSCFのaddboardコマンドを実行し、BBを物理パーティションに組み込む
- I/Oルートドメインの物理I/Oデバイスの冗長構成再設定
 - (2) I/Oルートドメインの物理I/Oデバイスを使用可能状態にする
 - (3) I/Oルートドメインの物理I/Oデバイスのマルチパスを再構成
- 制御ドメインの物理I/Oデバイスの冗長構成再設定
 - (4) 制御ドメインの物理I/Oデバイスを使用可能状態にする
 - (5) 制御ドメインの物理I/Oデバイスのマルチパスを再構成

詳細な手順につきましては「SPARC M12/M10 Building Blockを活用した高可用性システム(運用編)」を参照ください

5. 留意事項

■構築時の留意事項

- 各ドメインで使用するファイルシステムはすべてZFSにしてください。
- PPAR DR時のダウンタイム短縮のため、FCカードの全てのポートをLink Upできる状態にしてください。
- CPU自動交替およびPPAR DRを動作させるため、物理指定(CID/PA指定)でリソースを割り当てないでください。
- PPAR DRパラメータの設定変更後、有効にするためにはその物理パーティションをリセットしてください。
パラメータの変更方法は、XSCFリファレンスマニュアルの「setpparmode」を参照してください。
- 自動復旧を実施するためにはOVMのRecovery Modeの設定を行ってください。(デフォルトは無効)
パラメータの設定方法は「Oracle VM for SPARC Administration Guide」を参照してください。

■切り離すBBのXSCFについて

- 切り離すBBのXSCFはマスター以外のXSCFとなっている必要があります。マスタXSCFとして動作している場合は、事前にスタンバイ状態のXSCFに切り替えます。

※複数BB構成では各BBにXSCFユニットが存在します。その中の1つがマスタXSCFであり、それ以外は、マスタXSCFのバックアップとして動作するスタンバイ状態のXSCFが1つ、残りはスレーブXSCFです。

最新情報は、マニュアル、プロダクトノートを確認してください。

<https://www.fujitsu.com/jp/products/computing/servers/unix/sparc/lineup/m12-2s/documents/>

<https://www.fujitsu.com/jp/products/computing/servers/unix/sparc/lineup/m10-4s/documents/>

■PPAR DR実行時の留意事項

- PPAR DRでBBを切り離す処理は、完全に無停止で実施できる場合とメモリの再割り当てが発生して、各ドメインが一時停止することがあります。以下に記載の通り、2つの切り離すBBのメモリの状態により、ドメインの一時停止が発生します。
 - PPAR DRで切り離すBBに各ドメインのカーネルメモリが存在する場合、そのメモリを使用しているドメインが一時停止します。
 - PPAR DRで切り離すBBにHypervisorが使用するメモリが存在する場合、全ドメインが一時停止します。
- 一時停止する時間の最大値は、以下の時間の合計値になります。
 - メモリ移動時間
 - オンボードデバイスのサスペンド／レジューム時間
 - PCIカードのサスペンド／レジューム時間
- 各時間の見積り方法や詳細は、以下のマニュアルを参照してください。
 - 『SPARC M12/M10システムドメイン構築ガイド』の「2.5.2 運用時の留意点」
 - 『SPARC M12/M10システムPCIカード搭載ガイド』の「付録A PCI Hot PlugおよびDynamic Reconfiguration対応カード」
- 一時停止が発生した時間分、ゲストドメインの時刻遅延が発生するため、時刻同期の設定(NTPクライアント)を推奨します。
- 物理パーティション内にOBP状態の論理ドメインが存在している状態で、PPAR DRを実行した場合、エラーで終了します。PPAR DRは、各ドメインがOracle Solaris起動状態、または停止状態(inactive)のいずれかの状態のときに実行してください。
- PPAR DR時の停止時間削減のために以下設定を行ってください。
 - FCの各ポートは必ずLinkupさせてください

最新情報は、マニュアル、プロダクトノートを確認してください。

<https://www.fujitsu.com/jp/products/computing/servers/unix/sparc/lineup/m12-2s/documents/>

<https://www.fujitsu.com/jp/products/computing/servers/unix/sparc/lineup/m10-4s/documents/>

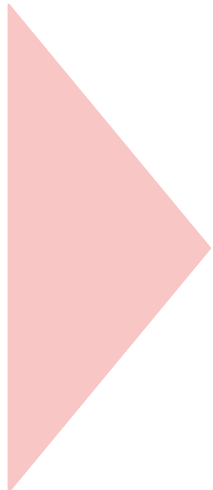
参考. 高可用性を実現するSPARC M12/M10の 機能詳細

■ 拡張性・可用性を合わせ持つ構成が可能

- SPARC M12-2S/M10-4SはBuilding Block(BB)方式を採用しており、複数のBBを高速インターコネクトで接続することで1つのシステム(物理パーティション)として取扱い、システムを拡張できます。
- システム拡張のためのBB追加や故障の際のBB交換はシステムを停止させずに活性で行うことができます。



複数筐体を接続した構成

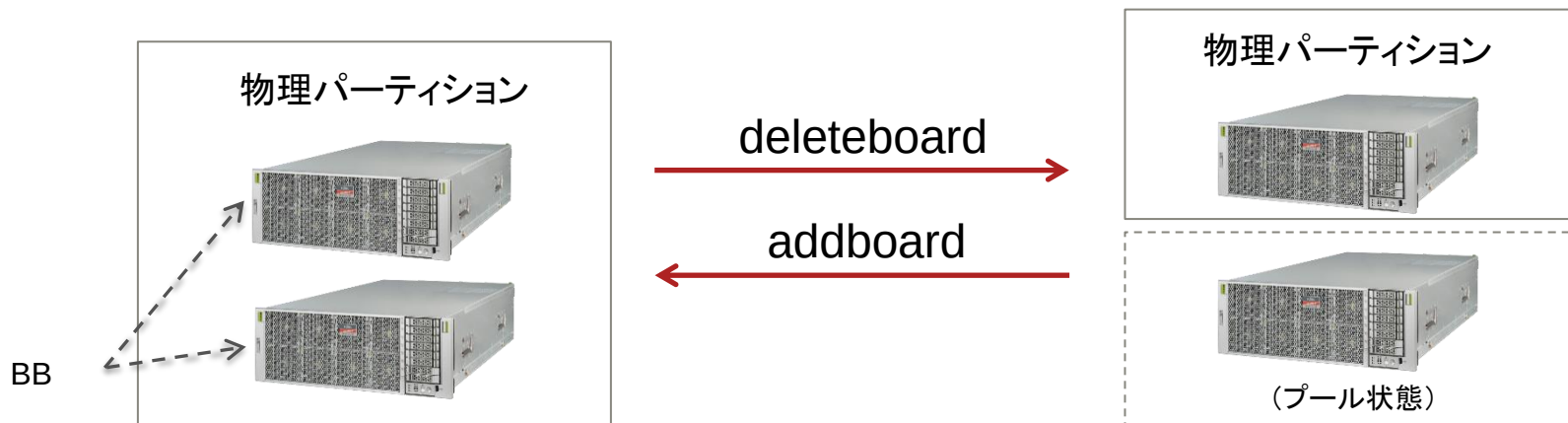
- 
- ✓ 最大で16BBまで接続可能
 - ✓ 複数のBBを組み合わせてハードウェアリソース(CPU、メモリ、PCIスロット)を拡張
 - ✓ BB単位に切り離し、組み込みが可能

■ BB単位に物理パーティションから切り離し・組み込みが可能

- 物理パーティションはSPARC M12-2S/M10-4SのBB単位で構成できます。
- 物理パーティションは、BB単位で切り離しや組み込みなどの構成変更が可能です。

■ ハードウェア構成変更(XSCFコマンド)

- BBの切り離し(deleteboard)
- BBの組み込み(addboard)



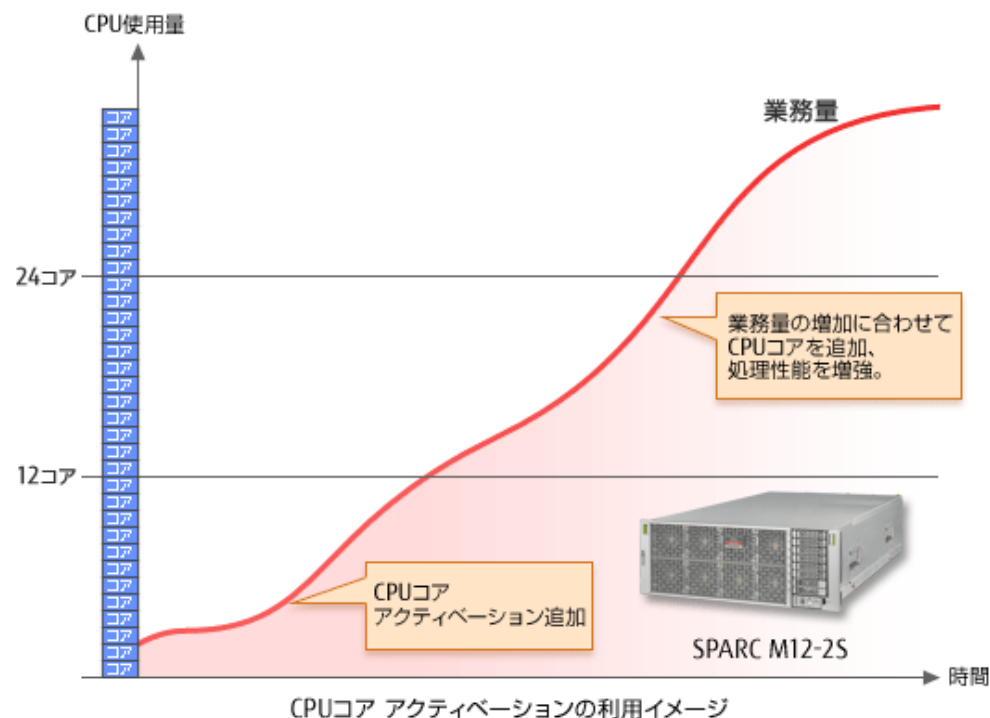
物理パーティションから切り離されたBBはプール状態となり、物理的にハードウェアを切り離したり、他の物理パーティションへ組み込んだりすることが可能です。

■ サーバの初期投資を抑えた導入が可能

- SPARC M12/M10は、業務量の増加に応じてシステムを停止せずにCPU能力を段階的に増強可能なCPUコア アクティベーションに対応しています。システム導入はSPARC M12-2/M12-2S/M10-1の場合2コア、SPARC M10-4/M10-4Sの場合4コアから可能であり、サーバの初期コストを削減できます。

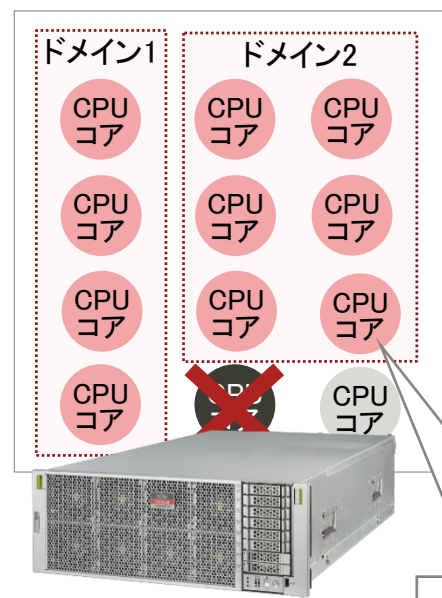
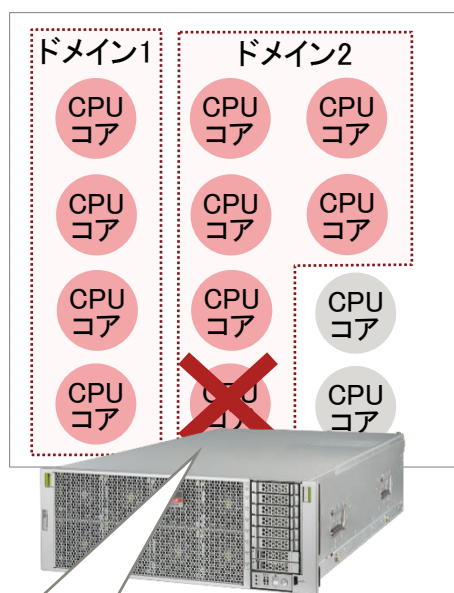
■ 業務負荷に応じた動的なリソース増強

- SPARC M12では1コア単位、SPARC M10では2コア単位での段階的なリソース追加が可能であり、業務の負荷に応じて、システム停止無しで、動的にCPUリソースを追加することが可能です。



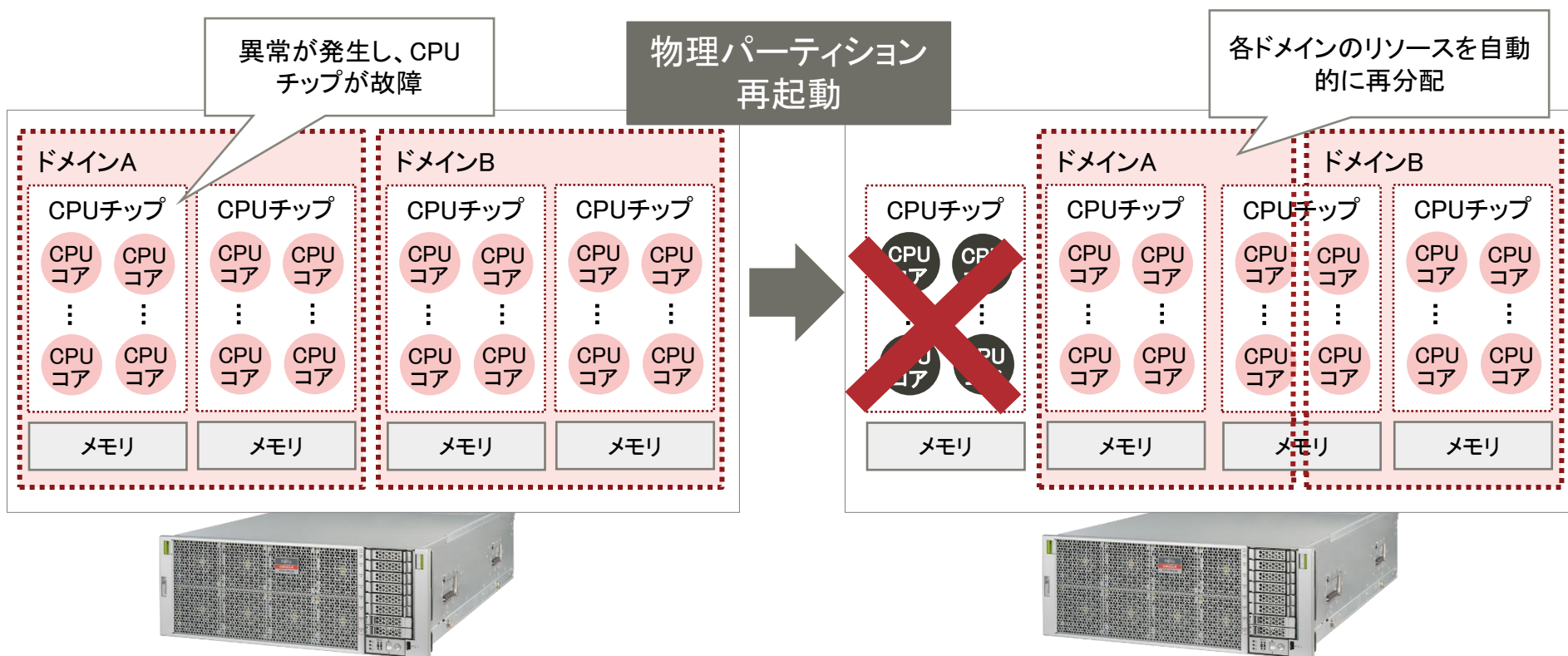
■ 自動CPUコア交替

- 物理パーティション内でアクティベートされていないCPUコアは故障発生時のスペアCPUコアとして扱われます。万が一CPUコアが故障した際には、故障したCPUコアと正常なCPUコアを自動的に交替し、瞬時に元通りの処理能力に戻すことが可能です。



■ Recovery mode(復旧モード)

- CPUチップ故障や筐体故障等によりドメイン構成を維持できなくなった場合でも、物理パーティション再起動後にリソースを自動的に再配分し、ドメイン構成を復元して業務を再開させます。



■ システム稼働中にハードウェア構成を変更

- 物理パーティションの構成変更は、システム稼働中に実行可能です。

■ PPAR DR機能のメリット

(1) システム無停止のリソース拡張

- 業務稼働中にCPU、メモリ、I/Oを追加できます。
- 業務ピークに合わせ、複数の物理パーティション間でリソースを移動できます。

(2) ハードウェア保守時のダウンタイム削減

- BB 交換でしか復旧できない故障が発生した場合に、システムを停止せずに活性交換ができます
- I/Oの冗長化と組み合わせることで、故障が発生した場合でも、故障発生から保守完了までの影響を最小限にできます。

■ ハード障害発生時の停止時間の最小化

- 2台のSPARC M12-2S/M10-4SサーバでI/Oを冗長構成にすると、I/O故障時に処理を継続します。
- ハード障害発生時、ドメインを自動的に再構築して復旧します。
- PPAR DRは、活性交換を提供します。



冗長構成

活性交換

自動復旧

SPARC M12-2S/M10-4SのBuilding BlockとPPAR Dynamic Reconfiguration(PPAR DR)を組み合わせることによって、システム停止時間を最小にすることができます。

改版日時	版数	更新内容
2015年6月	初版	新規作成
2015年9月	2版	記事修正(ミドルウェアの説明文一部変更)
		記事修正(制御、I/OルートドメインのOS版数を変更)
		記事修正(システムボリュームの説明文一部変更)
2017年4月	3版	SPARC M12-2Sサポートによる記事修正
		記事修正(ハードウェア条件、ソフトウェア条件)
		記事修正(BBとPPAR DRによる可用性向上を追加)
2021年4月	4版	各ドキュメントリンクアドレス最新化
2023年5月	5版	記事修正(ソフトウェア条件、留意事項)

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、SPARC64ロゴおよびすべての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。

