

IBM Power Systems 上での

CLUSTERPRO 動作検証結果

2010/01/15

(第2版: 2010/07/12 改訂)

1. はじめに

IT環境ではTCO削減が求められており、サーバーの設置スペースや消費電力の削減、ハードウェア資源の有効利用によりコスト削減を実現する手法の1つとしてサーバー仮想化技術を活用したサーバー統合が挙げられます。IBM Power Systemsは、複数のOSを1台の筐体で稼動させるサーバー仮想化機能(PowerVM)を標準搭載しており、数の少ないサーバーに業務を集約することができます。

一方、OSについてもコスト削減目的としてLinuxを採用するケースが増えています。企業や部門を最適化するための統廃合が進む中、Linuxはノウハウの再利用や技術者の確保を容易にしています。また、Linuxは様々なプラットフォームに対応しているため、導入するシステム毎に最適なプラットフォームを選択することができます。

Linuxの用途として旧来はWEBフロントエンドやメールサーバなどが中心でしたが、現在はシステムの基幹部であるリレーショナルデータベース(RDBMS)を安定稼働できるインフラとしての役割を担うまでになっています。

このように利用範囲が広まっているPower System上のLinuxですが、課題もありました。IBM Power Systemsはプロセッサー、メモリー、PCIアダプター等の各コンポーネントの高可用性を維持する機能を提供しますが、各論理区画(LPAR)上の業務アプリケーションの障害には対応できません。また、筐体の電源が全面的に停止する障害には対応できません。これはRDBMSのインフラとして使用するためには解決しなければならない課題です。RDBMSはデータ保全の要であるため、その可用性を高める事は極めて重要です。

この課題を解決するにはシステム異常を発見した場合に待機系に業務を引き継ぐHAクラスターソフトウェアを利用する事が一般的ですが、これまでIBM Power Systemsの機能を完全に引き出せるHAクラスターソフトウェアは存在しませんでした。

IBM Power Systemsが提供する仮想化機能(PowerVM)の1つにVirtual I/O Server(VIOS)があります。VIOSを利用することで複数のLPARでPCIデバイスを共有利用することが可能になり、必要最小限のPCIデバイスでネットワークや外部ディスクを共有利用することができます。しかし既存のHAクラスターソフトウェアは物理的なディスク環境に依存しているため、VIOS構成環境での外部ディスクの制御が出来ませんでした。

NEC CLUSTERPROはハードウェア環境に依存せずに稼働できるという特徴を持っており、VIOS構成環境においても使用可能です。

本稿ではIBM Power Systems上にVIOS環境、Linux、RDBMS(DB2)、CLUSTERPROを組み合わせた環境を構築し、クラスターとして正しく動作することを検証した結果を報告します。

2. 環境構成

2.1. ハードウェア構成

本検証では、筐体単位でのメンテナンスを含め、お客様環境を想定して物理的に2筐体を使用し、各筐体上に論理区画(Logical Partition, 以下 LPAR)を構成して筐体間での引き継ぎを行うクラスター構成としています。

各LPARには、OS、アプリケーションが使用するネットワークやディスクI/O用に物理アダプター、仮想アダプターを構成、更には物理/仮想を共存させることも可能ですが、今回は柔軟な構成を可能にする仮想アダプターを構成しています。

クラスターを構成するノード間の共有ディスクにはStorage Area Network(SAN)ディスクを使用し、SANスイッチ経由で接続しています。

2.1.1. 仮想アダプターとVirtual I/O Server

IBM Power Systemsが提供する仮想化機能(PowerVM)の1つに、仮想アダプターがあります。仮想アダプターは、LPAR上でネットワークならびにディスクI/O用のインターフェースを提供し、名称のとおり各LPARには物理アダプターの割り当てを必要としません。一方、外部ネットワークとの通信やSANディスクのような物理ディスクへのアクセスには物理アダプターでの接続が必要です。各LPAR上に構成する仮想アダプターと外部ネットワークやディスクとの接続、つまり仮想と物理の橋渡しを図1のようにVirtual I/O Server(以下VIOS)が担います。

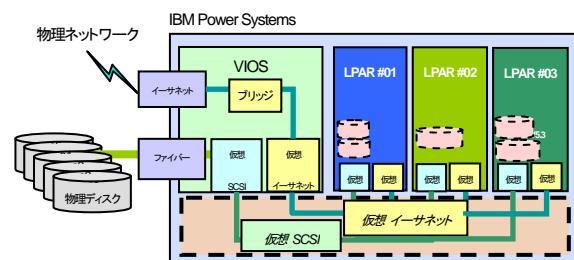


図1 Virtual I/O Server 概要

VIOS 区画にはネットワークやディスクと接続に使用する物理アダプターを構成し、その物理アダプターと各 LPAR 上に構成した仮想アダプター間の橋渡しを VIOS が行います。可用性ならびに保守性を考慮して VIOS に構成する物理アダプターを冗長化することが可能であり、更には VIOS 区画自体を 1 筐体内に冗長構成とすることが可能です。

2.1.2. ディスク構成

本検証におけるディスク構成を図 2 に示します。

2 台の Power Systems 上に各々 LPAR を構成し、各 LPAR には OS 領域用とデータエリア用のディスクをそれぞれ割り当てます。外部 SAN ディスクに RAID5 を構成し、OS 領域 / データエリア共に RAID5 上に作成した LUN を使用します。

OS 領域用には LPAR 毎に独立した形態でディスクを割り当てます。この OS 領域用ディスクは CLUSTERPRO による引き継ぎ対象とはなりません。一方、データエリアは引き継ぎ対象とするため、クラスターを構成する LPAR 間で共有ディスクとして構成します。本検証では、この共有ディスク上にデータベースが使用するデータを配置し、稼動系から待機系への引き継ぎ後にも同一データにアクセス可能な環境としています。データベースの構成に関しては、後述 2.3 章 データベース構成 を参照ください。

次に、仮想化環境における LPAR へのディスクの割り当てに関して記します。各筐体には可用性/保守性を考慮して VIOS を冗長構成としています。クラスターを構成する LPAR は、この冗長化した VIOS が各々提供するパスを経由してディスク(LUN)を multipath ディスクとして構成します。このように 1 筐体内に VIOS を冗長化し、各 VIOS から同一 LUN に対してパスを提供する multipath 構成により、片側の VIOS が停止中であっても LPAR はもう片側の VIOS 経由でディスクにアクセスする事を可能し、可用性を高めています。

(例) 図中の "VIOS1-1" が停止中であっても "VIOS1-2" 経由で RHEL5.4 およびデータエリアにアクセス可能

また、各 VIOS から外部 SAN ディスクへのアクセスに使用する物理アダプター(本検証ではファイバー・チャネル(FC)アダプター、図中の F)を冗長化し、物理アダプターの障害にも耐え得る構成としています。

LPAR の視点からは、同一 LUN に対して 2 VIOS を経由した 4 パスが確保されている構成となり、LPAR 上に 2 物理 FC ポートを構成した環境より高い可用性レベルが提供されることになります。当構成では VIOS に物理的に構成している 4 ポートの接続が全て切れるまで LPAR はディスク障害を検知することなく継続稼動可能な構成となります。

2.1.3. ネットワーク構成

本検証におけるネットワーク構成を図 3 (次ページ) に示します。ディスク構成と同様、2 台の Power Systems 上の各々 LPAR を作成し、各 LPAR にはパブリックとインタコネクトの 2 系統の LAN を構成します。パブリック LAN はサービス用、インタコネクト LAN はクラスターを構成する LPAR 間の死活監視の 1 つであるネットワーク経由のハートビート用に使用します。

仮想化環境における LPAR 上のネットワークの構成に関してですが、各筐体に冗長化した VIOS 経由により複数経路が確保されます。これにより、片側 VIOS が停止中の場合も LPAR 上に構成した仮想アダプターはもう片側の VIOS を経由して外部ネットワークと通信する事が可能です。物理アダプターの耐障害性を考慮して VIOS 上に割り当てるアダプターを冗長化し、EtherChannel もしくは EtherChannel Backup を構成することも可能ですが、本検証においては EtherChannel/EtherChannel Backup 共に構成していません。

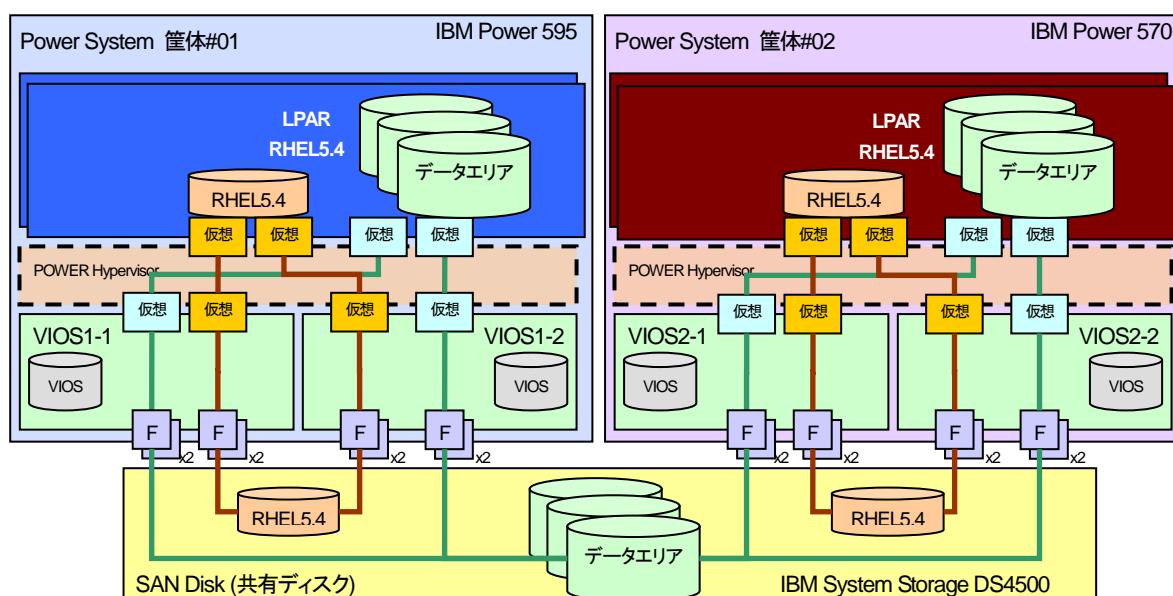


図 2 検証環境 ディスク構成

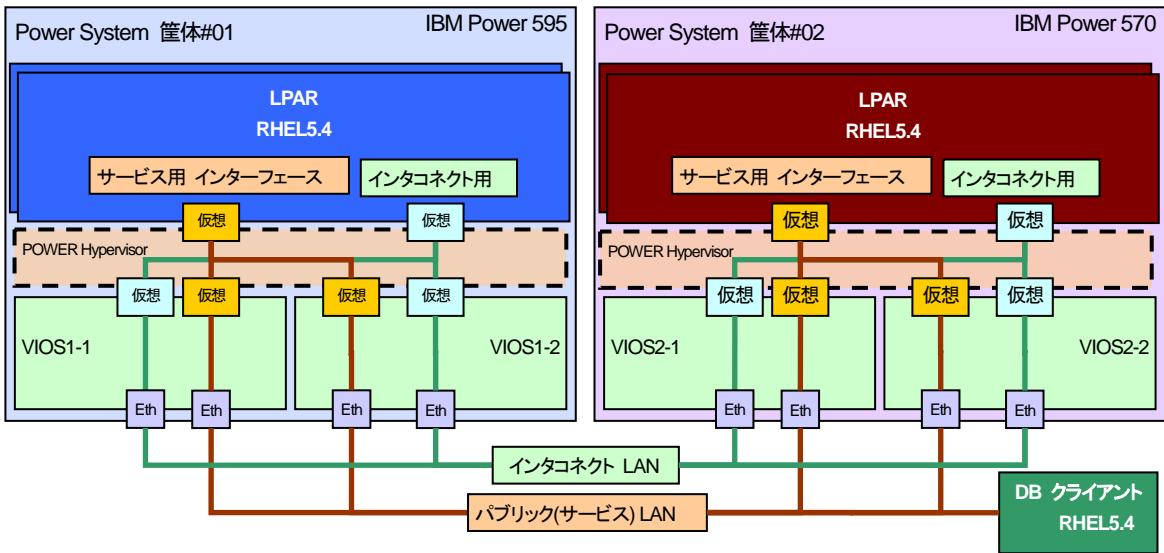


図 3 検証環境 ネットワーク構成

なお、本検証環境である VIOS 元長構成において LPAR からの経路を 2 経路提供する機能は SEA Failover により実装しています。VIOS に Primary と Backup を定義し、Primary 側の VIOS が停止した場合には Backup 側の VIOS 経由での通信に自動的に切り替わります。SEA Failover 機能は VIOS が提供し、冗長 VIOS の死活監視は VIOS が行います。

2.2. OS構成

LPAR ～ Red Hat Enterprise Linux 5.4 (以降、RHEL5.4) for IBM POWER を導入し、共有ディスクでわりあてた LUN へマルチパスアクセスするためのドライバー構成、パーティション作成、RAW デバイス定義を行います。

2.2.1. Linux導入

2 つの VIOS により冗長化されたストレージ接続パスを使用して、外部ディスクストレージの LUN に RHEL5.4 を導入します。通常の RHEL5.4 の導入手順と比べて主に次の 2 点が異なります。

- インストール時には、RHEL 5.4 に同梱されている Device Mapper Multipath(以降、DMMP)を使用します。
- 同一ネットワークセグメント上の PC にて VNC Viewer を使用し、ネットワーク経由で GUI のインストーラーを操作してインストールします。

また、VIOSにおいては次の設定を行います。

- 2 つの VIOS においては、RHEL5.4 をインストールする LUN の hdisk の属性「reserve_policy」を予め「no_reserve」にします。
- RHEL5.4 導入後は、LPAR の SMS にてブートリストに 2 つの vSCSI を指定します。

インストール作業は次の手順で実施します。

1) メディア準備

インストール対象の LPAR に DVD ドライブを接続し、RHEL 5.4 のインストール・メディアをセットします。

2) LPAR 起動

LPAR を起動して、SMS メニューより、DVD ドライブを指定して Normal Mode Boot を実行します。

3) インストーラー起動

コンソールに、インストーラーの起動メッセージが表示されますので、boot: ブロンプトに続けて、次のオプションを入力します。マルチバスドライバー DMMP を使用するために、「mpath」を指定します。また、VNC を使用しますので、vnc 関連のオプションを指定します。IP アドレスやパスワードは実際に使用するものを指定します。

入力例 :

```
linux vnc vncpassword=xxx ip=10.7.10.123
netmask=255.255.0.0 mpath
```

4) VNC 接続

VNC Viewer からの接続を要求するメッセージが表示されますので、同一ネットワーク上の PC の VNC Viewer から、GUI のインストーラー画面に接続します。認証画面が表示されますので、先に vncpassword で指定したパスワードを入力すると、GUI インストーラーである Anaconda の画面が表示されます。

これ以降の手順は通常の RHEL と同じ手順ですが、インストール先デバイスがマルチバスデバイスになっていることを確認します。

- ディスクパーティションレイアウト作成時に、確認画面を表示させ、デバイス名が「/dev/mapper/mpath0」になっていることを確認します。
- デバイス名が「/dev/sda」等になっている場合は、マルチバス環境が正しく構成されていません。

次に、Dynamic LPAR 等の PowerVM 仮想化機能や、信頼性・可用性・保守性(RAS) の機能を RHEL5.4 で利用するために、IBM から提供される Service and productivity tools を追加導入します。

1) rpm パッケージ準備

次のサイトから該当するシステムのリンクを選択し、

「RHEL 5」のタブにある rpm パッケージファイルをダウンロードします。

<http://www14.software.ibm.com/webapp/set2/sas/flopdiag/shome.html>

今回の構成では、「Red Hat」カテゴリーの「on HMC-or IVM-managed servers」を選択します。

2) rpm パッケージ導入

ダウンロードした RPM ファイルを導入対象の LPAR にファイル転送し、rpm コマンドで導入します。

実行例：

```
# rpm -Uvh*.rpm
```

以上で導入作業は終了ですが、マルチパスドライバー DMMP を使用して導入した RHEL5.4 環境は、運用するにあたり、次の点に注意してください。

インストーラにより、下記の3つのファイルに、ストレージ LUN のユニークな ID である wwid 番号が自動的に書き込まれます。システムバックアップ・リストア時などに新しい LUN へ導入する必要がある場合は、この wwid を新しい wwid に書き換える必要があります。

1) /var/lib/multipath/bindings ファイル

マルチパスデバイス mpathX と wwid の関連付けが定義されています。LUN の異なる環境ヘリストアする際等はリストア直後に新しい wwid へ変更してください。

2) /etc/multipath.conf ファイル

DMMP にて管理する対象 LUN の wwid が定義されています。LUN の異なる環境ヘリストアする際等は「blacklist_exceptions」の wwid を新しいものへ変更してください。

3) /boot/initrd-XXX ファイル

Linux 起動時に DMMP がマルチパスデバイスとして設定する LUN の wwid が、起動スクリプトファイルで定義されています。LUN の異なる環境ヘリストアする際等は、上記の2ファイルを書き換えてから、mkinitrd コマンドでこのファイルを再作成してください。

2.2.2. データベース領域マルチパスドライバー構成

Linux でのデータベース領域のマルチパス構成は、2 つの VIOS により冗長化されたストレージ接続パスを使用するため、RHEL5.4 のインストール時と同様に、マルチパスドライバー DMMP を使用して構成します。

構成は以下の手順で実施します。

1) wwid 確認

LUN 毎の wwid を scsi_id コマンドで調査します。

実行例：

```
# /sbin/scsi_id -g -u -s /block/sdc
```

2) DMMP 設定ファイル編集

wwid を DMMP 設定ファイルに記述します。

設定ファイル「/etc/multipath.conf」の「blacklist_exceptions」に wwid を追記

3) デバイスファイル生成

設定ファイルの変更を反映させます。

「multipath」コマンドを実行することで、DMMP に LUN を認識させ、デバイスファイル（例：「/dev/mapper/mpath1」）を生成させます。

データベースソフトウェアや OS 上のアプリケーション（fdisk 等）からは、ここで生成したデバイスファイルを指定して LUN へアクセスします。

尚、VIOS を使用しない構成では、マルチパスドライバーは DMMP ではなく、RDAC を使用する必要があります。RDAC の入手については次のサイトをご参照ください。

<http://www.lsi.com/rdac/>

クラスター環境で RDAC を使用する場合は、メインパス障害の時に次のようなパスの切り替えを繰り返す現象が発生します。

1. メインパス障害の発生ノードにマッピングされた全ての LUN でコントローラーの Ownership が切り替わります。これは、他ノードが使用中の LUN も含みます。
2. 他のノードが Preferred Path 経由での該当 LUN への接続を検知し、Ownership を Preferred Path に戻します。
3. メインパス障害ノードが再度 Ownership を切り替えます。

以後、2. と 3. を繰り返す現象が発生するため、約 1 分毎に一定量のエラーメッセージがシステムログに出力されます。切り替え自体はわずかな時間で行われますが、繰り返す現象を発生させないように、RDAC の設定変更が可能です。

RDAC の設定項目である Disable LUN Rebalance を設定し、優先パスの復旧時に Ownership の切り戻しを発生しないようにする手順です。VIOS を使用しない環境においては、必要に応じて設定してください。

- 1) /etc/mpp.conf をエディターで開き、DisableLUNRebalance の値を 3 に変更します。
- 2) mppUpdate コマンドで initrd イメージを作り直します。
- 3) システムをリブートし、新しい initrd イメージで起動します。

2.2.3. 共有ディスクパーティション構成

CLUSTERPRO を用いた HA クラスター環境では、クラスターサーバー間での死活監視に使用するクラスタパーティションが必要となります。

これらのパーティションは、必要なパーティション数の LUN を共有ディスクに作成し、マッピング、マルチパス構成することが可能ですが、1 つの LUN を複数パーティションにわけて構成することも可能です。今回は、1 つの LUN を複数パーティションにわけて構成する方法を用います。

パーティションの作成は RHEL5.4 の fdisk コマンドを使いますが、マルチパス環境では、マルチパスデバイスに対し

て fdisk コマンドを実行します。

実行例

```
# fdisk /dev/mapper/mpath1
```

今回作成した各パーティションの用途の詳細については、
「4.2.3 ディスク構成」に後述しています。

2.2.4. RAWデバイス定義

前述の手順で作成した、CLUSTERPRO の構成に必要な
クラスターパーティションは、RHEL5.4 にて RAW デバイス
として定義しておく必要があります。

Linux 上での RAW デバイスの定義は、RHEL5.4 に同梱
される rawdevices サービス機能を使用します。
構成は以下の手順で実施します。

1) デバイス定義

rawdevices 設定ファイルに実際のデバイス名と RAW デバイ
ス名の関係を定義します。

設定ファイル「/etc/sysconfig/rawdevices」に、
「/dev/raw/raw1 /dev/mapper/mpath1」のように指定します。
必要な RAW デバイスの数だけ定義します。

2) デバイスファイル生成

「service rawdevices restart」コマンドを実行することで設
定ファイルの変更を kernel に認識させ、デバイスファイル
(例：「/dev/raw/raw1」) を生成させます。このファイルは
OS を再起動すると消えてしまいます。

3) 自動生成設定

「chkconfig rawdevices on」コマンドを実行し、OS 再起動
後も自動的にデバイスファイルを生成するようにします。

ここで実施した rawdevices サービス機能による定義方法
の他に、udev 機能でも実施することができます。詳細につい
てはこちらの Web ページをご参照ください。

<http://kbase.redhat.com/faq/docs/DOC-10164>

CLUSTERPRO では、ここで生成したデバイスファイル
を指定します。

2.3. データベース構成 (DB2)

この章では今回の検証で IBM DB2 (以下 DB2) を
CLUSTERPRO 上に導入した際の情報をまとめます。

2.3.1. 導入ソフトウェア

導入したソフトウェアは以下の通りです。

- DB2 本体
 - DB2 v9.7 Enterprise Server Edition for POWER Linux (64bit)
- DB2 が必要とするコンポーネント
 - Runtime for XL C/C++ Advanced Edition for Linux, V9.0 (C++ランタイム)
 - libaio, pdksh (OS の CD-ROM より導入)

DB2 for Linux を動作させるために必要な追加ソフトウェ

アは、以下の URL にまとめられています。

<http://www.ibm.com/db2/linux/validate>

2.3.2. 導入計画 (物理配置)

今回のDB2 検証環境を図 4に示します。

PVC94サーバ
(プライマリ)
PVC84サーバ
(スタンバイ)

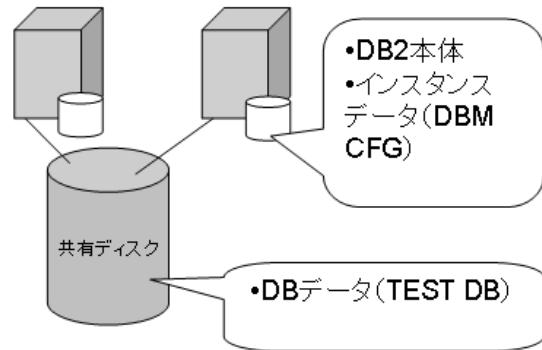


図 4 DB2 検証環境

DB2 を共有ディスク形式のクラスター環境に導入する際
には、まず以下のコンポーネントをローカルディスクと、共
有ディスクのどちらに導入するべきかを決定する必要があり
ます。

DB2 本体とインスタンスデータはローカルディスクにも
共有ディスクにも格納可能です。DB データは必ず共有ディ
スク上に格納します。

DB2 本体とインスタンスデータの両方をローカルディス
クに導入すると、共有ディスクをマウントしない状態で DB2
インスタンスを起動したり、パッチを適用したりすることができます。

これにより稼働系を業務で使用しながら、待機系にパッチ
(Fix Pack) を適用する事が可能になります。一般的にクラ
スターシステムを構築する要件ではできる限り停止時間を短
くするのが望ましいので、今回は表 1 のようにインスタンス
データをローカルディスクに配置しました。

表 1 ディレクトリ構成・配置

コンポーネント	ディスク装置	パス
DB2 本体	ローカル	/opt/ibm/db2/V9.7/
インスタンスデータ	ローカル	/home/ユーザ名/
DB データ	共有	/db2data/sharedb/

この構成では以下の 2 つの注意点があります。

- 1) DB データを保存するディレクトリは完全に空（1つも
ファイルが無い状態）である必要があります。
Linux ではデバイスのトップディレクトリ(今回は

/db2data としてマウントしたディレクトリ)には LOST+FOUND ディレクトリが自動的に作られているケースが多いので、そこには直接データが格納できません。今回はその下に sharedb ディレクトリを作る事で対応しています。

- 2) インスタンスレベルでパラメータを変更した場合は、両系統で更新する必要があります。
今回はインスタンスデータをローカルディスクに別々に保存しているため、インスタンスレベルでパラメータを変更した場合は、もう片方でも同様の操作をしてパラメータを合わせる必要があります。

インスタンスレベルのパラメータ変更は、DBM CFG (データベース・マネージャ・コンフィグ) の変更と、データベース作成や削除がそれに当ります。

2.3.3. 導入計画 (DB2 のユーザID)

導入前にDB2のユーザIDについて決定しておく必要があります。DB2では一般的に3つのユーザを使用します。

- インスタンスユーザ (インスタンスオーナー)
- 分離ユーザ (フェンスユーザ)
- 管理サーバユーザ(Admin ユーザ)

インスタンスユーザは、インスタンス (DB2 のプログラムそのもの) を管理するためのユーザです。分離ユーザはストアドプロシージャやユーザ定義関数を実行する際に使用されるユーザです。

管理サーバユーザは管理サーバーの起動と停止のために必要ですが、DB2の動作に必須ではありませんので、今回は作成していません。そのためインスタンスユーザと分離ユーザを以下のように計画しました。

	ユーザ(ID番号)	パスワード	グループ(ID番号)
インスタンスユーザ	db2inst1 (500) ※ポート番号:50000	password	db2group (500)
分離ユーザ	db2fenc1 (501)	password	db2group (500)

値は任意ですが、ID番号まで決めておく事が大切です。ID番号を両系統で統一しておくことで正しくディスクパーティションの設定が可能になります。

また、インスタンスユーザにはTCP/IPのポート番号が一つ必要になります。1,024より大きい値で、他のデーモン(サービス)が使用していないポート番号を選択してください。今回はデフォルト値の50,000を計画しました。

2.3.4. プライマリへの導入手順

今回の構成では、DB2本体をローカルディスクに導入するため、プライマリとサーバー、計2回導入を行う必要があります。まずプライマリ(PVC94)に導入します。

- 1) 前提ソフトウェアの導入
rpm コマンドで C++ Runtime や libaio, pdksh を導入します。

2) DB2 本体の導入

DB2 のインストーラーは GUI の db2setup とコマンドライン形式の db2_install があります。今回はどんな環境でも使用できる db2_install を使用して導入しました。

DB2 本体のファイル(DB2_ESE_97_Linux_ipSeries.tar.gz)を root ユーザで展開し、中にある db2_install を実行します。今回はインストールディレクトリを /opt/ibm/db2/V9.7 (デフォルト値) とし、導入プロダクトに ESE (Enterprise Server Edition) を選択しました。

3) インスタンスユーザ・分離ユーザの作成

計画に沿ってまずインスタンスユーザが使用する TCP/IP ポートを登録します。 /etc/services ファイルに以下のエントリを追加します。

```
db2c_db2inst1 50000/tcp
(db2c_db2inst1 は任意の名前)
```

次に OS 上にインスタンスユーザと分離ユーザを作成します。

```
# groupadd -g 500 db2group
# useradd -g db2group -u 500 db2inst1
# passwd db2inst1
```

```
# useradd -g db2group -u 501 db2fenc1
# passwd db2fenc1
```

OS上のユーザを使ってDB2のインスタンスを作成します。

```
# chmod 777 /tmp (※/tmp に db2inst1 がアクセス権限を持っていないと警告が出るため)
#/opt/ibm/db2/V9.7/instance/db2icrt -p 50000 -u db2fenc1
db2inst1
```

4) インスタンスの設定

インスタンスに最低限必要な設定を行います。

```
# /opt/ibm/db2/V9.7/instance/db2iauto -off db2inst1
(自動起動の停止)
# su - db2inst1
> db2start
> db2set DB2COMM=TCP/IP
> db2 update dbm cfg using SVCENAME db2c_db2inst1
(TCP/IP 通信が出来るように設定)
> db2stop
> db2start
(インスタンスを再起動しパラメータを反映)
> exit
```

5) 共有ディスクをプライマリ側にマウント

root ユーザで /db2data としてマウントしました。マウントの情報を /etc/fstab に記述すると OS が自動的にマウントしてしまい、CLUSTERPRO が制御できなくなってしまうため /etc/fstab に記述はしないでください。

6) データを保存するディレクトリを作成

root ユーザでディレクトリを作成し、そのオーナーをイン

スタンスユーザ（db2inst1:db2group）に変更します。これでDB2からデータを読み書き出来るようになります。

```
# mkdir /db2data/sharedb/  
# chown db2inst1:db2group /db2data/sharedb/
```

7) テスト用データベースの作成

作成しておいた共有ディスク上のディレクトリ内にデータベースを作成し、接続可能かを確認します。

```
# su - db2inst1  
> db2 CREATE DB TEST1 ON /db2data/sharedb/  
> db2 CONNECT TO TEST1  
> db2 TERMINATE
```

これでプライマリ側のDB2導入が完了しました。インスタンスを停止、/db2dataをアンマウントしておきます。

```
> db2stop  
> exit
```

2.3.5. スタンバイへの導入手順

スタンバイ側の導入手順は、プライマリと同じです。プライマリ側の手順1)～5)を実行してください。6)以降はすでに共有ディスク上にデータがあるため実行しないでください。

プライマリ側でデータベースを作成しましたが、スタンバイ側では作成していないので、スタンバイ側でデータベースが存在する事を CATALOGコマンドで認識させる必要があります。

```
> db2 CATALOG DATABASE TEST1 ON /db2data/sharedb/
```

これでスタンバイ側からもDBに接続できるようになります。接続確認後はインスタンスを停止、/db2dataをアンマウントしておきます。

```
> db2 CONNETC TO TEST1  
> db2 TERMINATE  
> db2stop  
> exit
```

2.3.6. DB2 クライアントの設定

DB2クライアントからは、CATALOGコマンドでリモートサイトのデータベースを認識させます。CLUSTERPROで使用する場合は、一般的にフローティングIPアドレスがサーバー側に用意されるので、CATALOGでもフローティングIPを指定します。

1) クライアント用のインスタンス作成

クライアントマシン上でクライアント用のインスタンスを作成します。

(DB2クライアントを導入後)

```
# chmod 777 /tmp  
# groupadd -g 500 db2group  
# useradd -g db2group -u 500 db2inst1
```

```
# passwd db2inst1  
# /opt/ibm/db2/V9.7/instance/db2icrt -s client db2inst1
```

2) リモートDBをカタログ

CATALOG TCPIPコマンドでリモートDBのIPアドレスとポート番号を指定する。フローティングIPアドレスとして172.16.11.4を使用して設定した。

```
# su - db2inst1  
> db2 CATALOG TCPIP NODE DB2SVR REMOTE 172.16.11.4  
SERVER 50000  
> db2 CATALOG DB TEST1 AT NODE DB2SVR  
> db2 CONNECT TO TEST1 USER db2inst1 USING password  
> db2 TERMINATE
```

2.3.7. DB2をクラスターから制御するコマンド

DB2には、ユーザが通常使用するコマンド以外にクラスターソフトウェアからDB2を制御するためのdb2gcfコマンドが用意されています。

これらはタイムアウトが指定でき、環境変数を設定せずに直接利用できるという特徴がありますので、クラスターソフトウェアの制御スクリプトをカスタマイズする際はこれらのコマンドを利用すべきです。

インスタンスの開始	db2gcf -u -i <インスタンス名> -t <タイムアウト秒数> -L
インスタンスの停止	db2gcf -d -i <インスタンス名> -t <タイムアウト秒数> -L
インスタンスの強制停止	db2gcf -k -i <インスタンス名> -t <タイムアウト秒数> -L
インスタンスの状態表示	db2gcf -s -i <インスタンス名> -t <タイムアウト秒数> -L

db2gcfコマンドの詳細はDB2のマニュアルに記載されています。

(<http://publib.boulder.ibm.com/infocenter/db2luw/v9r7/index.jsp?topic=/com.ibm.db2.luw.admin.cmd.doc/doc/r0010986.html>)

2.3.8. CLUSTERPROからのDB2の監視

CLUSTERPROからのDB2を監視する場合は以下の2つの方法が考えられます。

1) データベース接続の試行

データベースに実際に接続し、表作成などのSQLを定期的に実行することで正常稼働しているかどうかを確認します。CLUSTERPROのDB2モニタリソースは自動的にこの接続試行を行います。

2) インスタンスの状況監視

インスタンス全体の状況を監視するには、インスタンス全体の親プロセスであるdb2sysc(Windows環境ではdb2syscs.exe)の存在をpsコマンド等で監視するか、db2gcfコマンドでステータスを監視する方法があります。上記の監視を実施して正常であれば0、異常であればそれ以外の値を

返却するシェルスクリプトを作成し、CLUSTERPRO のカスタムモニタリソースに登録することで、インスタンスの状況監視を CLUSTERPRO に組み込むことができます。

今回は 1) を設定しました。DB2 モジュールの設定では DB2 導入ディレクトリ (ライブラリへのパス) 指定と、データベースの文字コードを設定する必要があります。

デフォルト設定で導入した場合、ディレクトリは /opt/ibm/db2/V9.7/ になり、ライブラリはその直下にあるため、モジュールに指定するのは /opt/ibm/db2/V9.7/lib32/ もしくは /opt/ibm/db2/V9.7/lib64/ になります。CLUSTERPRO の POWER Linux 用モジュールは 64bit バイナリですので、/opt/ibm/db2/V9.7/lib64/ を指定しています。

文字コードは、DB2 でデータベースを作成(CREATE DATABASE)した際に指定したものと同じものを指定します。指定しないで CREATE DATABASE した場合のデフォルト値は Unicode(UTF-8) です。本検証では UTF-8 を使用しています。

2.4. クラスター構成

本検証では、各 Power Systems の各論理区画(LPAR)上に CLUSTERPRO をインストールし、両方向スタンバイ構成(但しデータベースは片方向構成で実施しました)。

2.4.1. 導入ソフトウェア

本検証では、クラスタリングソフトウェアとして次の製品を使用しました。

- CLUSTERPRO 本体
 - CLUSTERPRO X 2.1 for Linux
 - ❖ アップデート(CPRO-XL050-05)適用
- データベース監視用オプション
 - CLUSTERPRO X Database Agent 2.1 for Linux¹

2.4.2. ネットワーク構成

CLUSTERPRO では、ハートビート経路に冗長性を持たせるため、下記インタコネクト LAN を 1 系統以上含む、計 2 系統以上のネットワーク経路を使用することを推奨しています。

- インタコネクト LAN
クラスターを構成するサーバ間で、互いの死活監視やクラスター情報の交換に使用。他用途での通信には使用しないことを推奨。(ハートビート LAN とも呼称します)
- パブリック LAN
クライアントとの通信に使用する業務 LAN。
インタコネクト LAN 異常時の、バックアップ用として

¹ ppc 版のデータベース監視は、本検証に合わせて開発し、対応完了したものを使用しました。入手方法については末尾の NEC 連絡先にお問い合わせください。(DB2 および Oracle 監視に対応)

も使用。

本検証では、インタコネクト LAN およびパブリック LAN をそれぞれ 1 系統ずつ、計 2 系統を使用しました。

2.4.3. ディスク構成

ストレージ内の同一ボリュームに対し、用途別に以下のパーティションに区切って使用しました。

なお、これらのパーティションは OS によって自動マウントされないよう、CLUSTERPRO のインストール前に設定しておく必要があります。

- クラスタパーテイション
クラスターを構成するサーバ間での、互いの死活監視(ディスクハートビート)に使用。
共有ディスクを使用する場合は、切替パーティションを含む LUN 每に 1 つ以上の作成を推奨します。
本パーティションは CLUSTERPRO のディスクハートビートリソースが RAW で Read/Write のアクセスを行いますので、ファイルシステムを作成する必要はありません。
- 切替パーティション(ext3 形式)
業務データの格納に使用します。
この領域は、稼動系側において CLUSTERPRO のディスクリソースによりアクセス制御および mount され、アクセスが可能になります。
ext3, xfs など主要なファイルシステムに対応。(ただしジャーナリングファイルシステムを推奨)
本検証では、DB2 のデータベース格納用領域として使用しました。
- 切替パーティション(RAW 形式)
RAW 形式でアクセスされる業務データの格納に使用します。
この領域は、稼動系側において CLUSTERPRO の RAW リソースにより OS の RAW デバイスにマップ(bind)され、アクセスが可能になります。
本検証では、RAW パーティションを使用する上位アプリケーションは存在しませんが、検証用ダミー領域として使用しました。

2.4.4. ハートビートとネットワークパーティション解決

CLUSTERPRO では、サーバー間の死活監視のためハートビートを実施しています。

万一ハートビートが行えなくなると、スプリットブレイン状態になり、業務停止や、最悪の場合共有ディスクの二重マウントによるデータ破壊が発生しかねないため、以下の機構を採用することでそのような状況を防止しています。

- 複数の経路使用
本構成では次の経路を使用することで、一部がハードウェア故障等により不通になってもハートビートが継続可能な構成としています。
 - LAN ハートビート(「2.4.2 ネットワーク構成」に記載の 2 系統を使用)
 - ディスクハートビート(「2.4.3 ディスク構成」に記載のクラスタパーテイションを使用)

- Kernel LAN ハートビートの使用
高負荷環境での LAN ハートビート送受信を確実に行うため、各システムとも User 空間のほかに Kernel 空間でもハートビートを実施しています。

サーバ間監視 (LAN, ディスクハートビート)

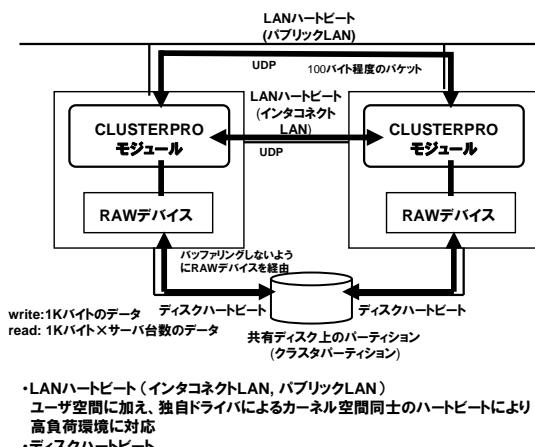


図 5: ハートビート経路

また、万一全てのハートビート経路で異常が発生した場合でも、スプリットブレインによるデータ破壊を防ぐため、次の機構を採用しています。

- Ping ネットワークパーティション解決リソース
各サーバーから通信可能で、常時起動している外部ネットワーク機器を対象として登録しておき、相手からのハートビート途絶時にそのネットワーク機器への通信ができない場合、強制的に自サーバーをシャットダウンする機構。
本機構により、全ハートビート経路断線によりスplitsplit-brainに陥った場合でも、登録した対象ネットワーク機器と通信できない側が自発的にシャットダウンするため、業務二重活性とその延長で発生し得るデータ破壊を防止します。

2.4.5. グループリソース

CLUSTERPRO では、切替対象の単位として各種グループリソースを登録可能です。

本検証では、次のグループリソースを登録しました。

- ディスクリソース
稼動系側で共有ディスク上のデバイスをマウントするための機構を提供します。
また本検証では、このリソースにより管理されるディスク領域を、DB2 のデータベース保管用に使用しました。
- Exec リソース
DB2 など業務サービスを起動/停止するためのスクリプトを動作させる機構を提供します。
なお、本リソースで起動/停止する業務サービスは、OS によって自動起動されないように設定しておく必要があります。

- フローティング IP リソース
仮想的な IP アドレスを提供。これによりクライアントからはどちらのサーバーが稼動系を意識せずに接続することが可能となります。
- RAW リソース
稼動系側で共有ディスク上の RAW デバイスを使用可能とするための機構を提供します。
本検証では、上位アプリケーションから使用されませんでしたが、検証目的で登録しました。

2.4.6. モニタリソース

CLUSTERPRO では、監視対象に応じて各種モニタリソースを登録可能です。

本検証では、次のモニタリソースを登録しました。

- ARP モニタリソース
フローティング IP リソース用の ARP パケットを送出する機構を提供します。
- DB2 モニタリソース
DB2 データベースの監視機構を提供します。DB2 のインスタンスに対し、ダミーデータの読み書きを行う各種 SQL 文(create/drop/insert/update/select)を発行することで、データベースが正しく応答するか、サービス観点で監視を実施します。
- ディスクモニタリソース
ローカルディスクおよび共有ディスクの監視機構を提供します。
- IP モニタリソース
外部ネットワーク機器への導通を監視する機構を提供します。
- RAW モニタリソース
RAW デバイスの監視機構を提供します。
- ユーザ空間モニタリソース
ユーザ空間のストール監視機構を提供します。

3. テスト項目および結果

3.1. 正常状態チェック

以下の項目を実施し、正常状態においていずれの動作も問題ないことを確認しました。

・WebManager 接続

項目	確認内容
WebManager 接続	WebManager(CLUSTERPRO の管理マネージャー)へ接続できることを確認

・サーバー監視

項目	確認内容
サーバー	WebManager よりサーバーの状態が「正常動作中」となっていることを確認

・ハートビート、ネットワークパーティション解決

項目	確認内容
LAN ハートビートリソース	WebManager より全ての LAN ハートビートリソースの状態が「正常動作中」となっていることを確認 (ハートビート専用 LAN、パブリック LAN とも)
ディスクハートビートリソース	WebManager よりディスクハートビートリソースの状態が「正常動作中」となっていることを確認
Ping ネットワークパーティション解決リソース	WebManager より Ping ネットワークパーティション解決リソースの状態が「正常動作中」となっていることを確認

・グループ、グループリソース

項目	確認内容
グループ	WebManager より全てのグループの状態が「正常動作中」となっていることを確認
ディスクリソース	WebManager よりディスクリソースの状態が「正常動作中」となっていること、および対象マウントポイント先にマウントされていることを確認
EXEC リソース	WebManager より EXEC リソースの状態が「正常動作中」となっていること、およびリソースの起動/停止時に、それぞれ開始スクリプト/停止スクリプトが実行されることを確認
フローティング IP リソース	WebManager よりフローティング IP リソースの状態が「正常動作中」となっていること、およびクライアントからフローティング IP アドレスを使用して活性側サーバーと通信可能なことを確認
RAW リソース	WebManager より全ての RAW リソースの状態が「正常動作中」となっていることを確認

・モニタリソース

項目	確認内容
ディスクモニタリソース	WebManager よりディスクモニタリソースの状態が「正常動作中」となっており、監視異常を検出しないことを確認
IP モニタリソース	WebManager より IP モニタリソースの状態が「正常動作中」となっており、監視異常を検出しないことを確認
NIC Link Up/Down モニタリソース	WebManager より NIC Link Up/Down モニタリソースの状態が「正常動作中」となっており、監視異常を検出しないことを確認
PID モニタリソース	WebManager より PID モニタリソースの状態が「正常動作中」となっており、監視異常を検出しないことを確認
RAW モニタリソース	WebManager より RAW モニタリソースの状態が「正常動作中」となっており、監視異常を検出しないことを確認
ユーザ空間モニタリソース	WebManager よりユーザ空間モニタリソースの状態が「正常動作中」となっており、監視異常を検出しないことを確認
ARP モニタリソース	WebManager より ARP モニタリソースの状態が「正常動作中」となっており、監視異常を検出しないことを確認
DB2 モニタリソース	WebManager より DB2 モニタリソースの状態が「正常動作中」となっており、監視異常を検出しないことを確認

3.2. 異常状態チェック

ハードウェアやOS障害を想定して以下のコンポーネントに対し擬似障害を発生させ、いずれの動作も問題ないことを確認しました。

・共有ディスク装置

項目	操作	確認内容
共有ディスク装置 SCSI/FCバス	サーバー側 FC ケーブル を抜線	<ul style="list-style-type: none"> ・ディスクモニタリソースが異常検出すること ・ディスクハートビートリソースが停止すること ・グループが待機系にフェイルオーバすること

・ネットワーク経路

インタコネクト LAN	インタコネクト LAN を抜線	<ul style="list-style-type: none"> ・インタコネクト LAN に対応するハートビートリソースが非活性になること ・グループのフェイルオーバは発生せず元のサーバーで継続動作すること(ハートビートはパブリック LAN をバックアップ用ハートビート経路として使用し継続)
パブリック LAN	パブリック LAN を抜線	<ul style="list-style-type: none"> ・パブリック LAN に対応するハートビートリソースが非活性になること ・IP モニタリソースが異常を検出しグループが待機系へフェイルオーバすること

・VIOS

VIOS	同一筐体内に存在する 2 つの VIOS のうち片方を停止	<ul style="list-style-type: none"> ・もう一方の VIOS により I/O は正常に動作継続すること ・CLUSTERPRO は異常検出することなくグループが稼動系で活性し続けること
------	-------------------------------	--

・グループリソース、モニタリソース

項目	操作	確認内容
ディスクリソース	ディスクをマウントした状態でグループを起動する	<ul style="list-style-type: none"> ・ディスクリソース活性に失敗し、待機系へフェイルオーバすること
EXEC リソース	「exit 1」で終了するスクリプトに設定し、グループを起動する	<ul style="list-style-type: none"> ・EXEC リソース活性に失敗し、待機系へフェイルオーバすること
フローティング IP リソース	ネットワーク内の他サーバーで同一のアドレスを附加した状態でグループを起動する	<ul style="list-style-type: none"> ・フローティング IP リソース活性に失敗し、待機系へフェイルオーバすること
RAW リソース	存在しない RAW デバイス名を指定してグループを起動する	<ul style="list-style-type: none"> ・RAW リソース活性に失敗し、待機系へフェイルオーバすること
PID モニタリソース	EXEC リソースの常駐プロセスを kill コマンドで外部から強制終了	<ul style="list-style-type: none"> ・PID 監視異常によりグループが待機系へフェイルオーバすること

3.3. フェイルオーバグループ遷移チェック

以下の状態遷移を実施し、各サーバーのステータスおよびグループの状態が正しく遷移することを確認しました。

シーケンス	操作	確認内容
1	サーバー1、サーバー2を起動	サーバー1、サーバー2が「起動済」状態になること グループがサーバー1で起動すること
2	グループをサーバー2へ手動でフェイルオーバ	グループがサーバー2へフェイルオーバすること
3	グループをサーバー1へ手動でフェイルオーバ	グループがサーバー1へフェイルオーバすること
4	サーバー1をコマンドからシャットダウン	グループがサーバー2へフェイルオーバすること
5	WebManager からクラスタシャットダウン操作	グループが停止しサーバー2がシャットダウンすること
6	サーバー1、サーバー2を起動	サーバー1、サーバー2が「起動済」状態になること グループがサーバー1で起動すること
7	サーバー1をコマンドからシャットダウン	サーバー1が「停止」状態になること グループがサーバー2へフェイルオーバすること
8	サーバー2をコマンドからシャットダウン	グループが停止状態となること
9	サーバー1、サーバー2を起動	サーバー1、サーバー2が「起動済」状態になること グループがサーバー1で起動すること
10	WebManager からクラスタシャットダウン操作	グループが停止し、サーバー1、サーバー2がシャットダウンすること
11	サーバー2を起動 [相手サーバー起動待ち(5分) 経過後]	サーバー2が「起動済」状態になること グループがサーバー2で起動すること
12	サーバー1を起動	サーバー1が「起動済」状態になること
13	サーバー2をコマンドからシャットダウン	サーバー2が「停止」状態になること グループがサーバー1へフェイルオーバすること
14	サーバー2を起動	サーバー2が「起動済」状態になること

15	WebManager からクラスタシャットダウン	グループが停止し、サーバー1、サーバー2がシャットダウンすること
16	サーバー1を起動 [相手サーバー起動待ち(5分) 経過後]	サーバー1が「起動済」状態になること グループがサーバー1で起動すること
17	サーバー2を起動	サーバー2が「起動済」状態になること
18	サーバー1をコマンドからシャットダウン	サーバー1が「停止」状態になること グループがサーバー2へフェイルオーバすること
19	サーバー1を起動	サーバー1が「起動済」状態になること
20	WebManager からクラスタシャットダウン	グループが停止し、サーバー1、サーバー2がシャットダウンすること
21	サーバー1、サーバー2を起動	サーバー1、サーバー2が「起動済」状態になること グループがサーバー1で起動すること

3.4. データベース動作チェック

DB2 のクラスター セットに対し以下の項目を実施し、いずれの動作も問題ないことを確認しました。(本章では、DB2 Agent を”Database Agent”と記します)

項目	操作	確認内容
起動	WebManager から手動でグループ起動	<ul style="list-style-type: none">データベースが起動し、グループが起動状態となることクライアント端末からデータベースに接続しトランザクションを正常に処理できること
正常系動作	同上	<ul style="list-style-type: none">Database Agent が監視開始状態となり、異常を検出しないこと
停止	WebManager から手動でグループ停止	<ul style="list-style-type: none">データベースが停止し、グループが停止状態となること
DB プロセス異常	DB のプロセスを kill コマンドで外部から強制終了	<ul style="list-style-type: none">Database Agent が異常を検出することグループが待機系にフェイルオーバされ、クライアントからトランザクションを継続して処理できること
サーバー異常	稼動系サーバーで外部コマンドにより OS シャットダウン	<ul style="list-style-type: none">Database Agent が異常を検出することグループが待機系にフェイルオーバされ、クライアントからトランザクションを継続して処理できること

3.5. テスト結果まとめ

上記のとおり、Power Systems 上において DB2 をクラスター化し、クライアントからのトランザクションが問題なく継続可能なことを確認しました。

また、ハードウェア、OS、上位アプリケーション(DB2)それぞれの観点で擬似障害を発生させた結果、いずれも想定どおりフェイルオーバが発生し、クライアントへのトランザクションを継続して提供することを確認しました。

4. おわりに

本稿では、IBM Power Systems 環境において、CLUSTERPRO の導入による DB2 の二重化構成について検証を実施しました。

その結果、正常系および異常系ともに想定どおり動作し、ハードウェア、OS、アプリケーションの各種コンポーネントのいずれかで障害が発生した場合でもフェイルオーバを実施することにより、サービスの可用性が向上することを確認できました。

ました。

また、本構成の導入により、他のクラスタリングソフトとの組み合わせでは実現困難な次のメリットが受けられることを実証しました。

- VIOS 構成で外部ディスク制御機能を備えたクラスター化が可能
- Database Agent の導入により、データベースの動作を単純なプロセス生存監視に留まらず、サービスレベルで監視することが可能

5. 参考URL

- CLUSTERPRO
<http://www.nec.co.jp/clusterpro/>
- IBM Power Systems
<http://www.ibm.com/systems/jp/power/>
- IBM PowerVM
<http://www.ibm.com/systems/jp/power/software/virtualization/>
- IBM Linux for Power Systems
<http://www.ibm.com/systems/jp/power/software/linux/>
- IBM DB2
<http://www.ibm.com/software/jp/data/db2/>

日本アイ・ビー・エム株式会社
パワーシステム事業部

日本電気株式会社
第一IT ソフトウェア事業部
CLUSTERPRO グループ
info@clusterpro.jp.nec.com

[商標登録]

IBM、Power Systems、PowerVM、DB2 は International Business Machines Corporation の米国およびその他の国における登録商標です。

CLUSTERPRO は日本電気株式会社の登録商標です。

Linux は Linus Torvalds 氏の米国およびその他の国における登録商標または商標です。

その他、文中の社名、商品名は、各社の商標または登録商標です。

付録

CLUSTERPRO設定情報

• DB2 起動スクリプト(start.sh)

```
#!/bin/sh
#####
#*          start.sh      *
#####
ulimit -s unlimited

if [ "$CLP_EVENT" = "START" ]
then
    if [ "$CLP_DISK" = "SUCCESS" ]
    then
        echo "NORMAL1"

    # "-t"で指定する値は、インスタンス開始処理を待ち合わせる
    # タイムアウト時間です。開始完了までの最大所要時間を上回る
    # 秒数を指定してください。
        /opt/ibm/db2/V9.7/bin/db2gcf -i db2inst1
-u -t 600
        if [ $? -ne 0 ]; then
            exit 1
        fi

        if [ "$CLP_SERVER" = "HOME" ]
        then
            echo "NORMAL2"
        else
            echo "ON_OTHER1"
        fi
    else
        echo "ERROR_DISK from START"
    fi
elif [ "$CLP_EVENT" = "FAILOVER" ]
then
    if [ "$CLP_DISK" = "SUCCESS" ]
    then
        echo "FAILOVER1"

    # "-t"で指定する値は、インスタンス開始処理を待ち合わせる
    # タイムアウト時間です。開始完了までの最大所要時間を上回る
    # 秒数を指定してください。
        /opt/ibm/db2/V9.7/bin/db2gcf -i db2inst1
-u -t 600
        if [ $? -ne 0 ]; then
            exit 1
        fi

        if [ "$CLP_SERVER" = "HOME" ]
        then
            echo "FAILOVER2"
        else
            echo "ON_OTHER2"
        fi
    else
```

```
        echo "ERROR_DISK from FAILOVER"
    fi
else
    echo "NO_CLP"
fi
echo "EXIT"
exit 0
```

• DB2 停止スクリプト(stop.sh)

```
#!/bin/sh
#####
#*          stop.sh      *
#####
ulimit -s unlimited

if [ "$CLP_EVENT" = "START" ]
then
    if [ "$CLP_DISK" = "SUCCESS" ]
    then
        echo "NORMAL1"

    # "-t"で指定する値は、インスタンス停止処理を待ち合わせる
    # タイムアウト時間です。停止完了までの最大所要時間を上回る
    # 秒数を指定してください。
        /opt/ibm/db2/V9.7/bin/db2gcf -i db2inst1
-d -t 600
        if [ $? -ne 0 ]; then
            /opt/ibm/db2/V9.7/bin/db2gcf      -i
db2inst1 -k -t 600
            exit 1
        fi

        if [ "$CLP_SERVER" = "HOME" ]
        then
            echo "NORMAL2"
        else
            echo "ON_OTHER1"
        fi
    else
        echo "ERROR_DISK from START"
    fi
elif [ "$CLP_EVENT" = "FAILOVER" ]
then
    if [ "$CLP_DISK" = "SUCCESS" ]
    then
        echo "FAILOVER1"

    # "-t"で指定する値は、インスタンス停止処理を待ち合わせる
    # タイムアウト時間です。停止完了までの最大所要時間を上回る
    # 秒数を指定してください。
        /opt/ibm/db2/V9.7/bin/db2gcf -i db2inst1
-d -t 600
        if [ $? -ne 0 ]; then
            /opt/ibm/db2/V9.7/bin/db2gcf      -i
db2inst1 -k -t 600
```

```
        exit 1
    fi

    if [ "$CLP_SERVER" = "HOME" ]
    then
        echo "FAILOVER2"
    else
        echo "ON_OTHER2"
    fi

else
    echo "ERROR_DISK from FAILOVER"
fi
else
    echo "NO_CLP"
fi
echo "EXIT"
exit 0
```
