CLUSTERPRO® システム構築ガイド

CLUSTERPRO® for Linux Ver 2.0

システム設計編(基本/ミラーディスク)

第2版　　2002.07.23

改版履歴

版　数	改版年月日	改版ページ	内　　容
1	2002.5.8		Ver1.0 第2版をベースに新規作成
2	2002.7.23	10 13	図のサブクラスタのCOMポートを削除 付加分散グループについての設定注意事項を削除 共有/ミラーディスクに関するエラーメッセージを削除

はじめに

『CLUSTERPRO®システム構築ガイド』は、クラスタシステムに関して、システムを構築する管理者、及びユーザサポートを行うシステムエンジニア、保守員を対象にしています。

CLUSTERPROドキュメント体系

CLUSTERPROのドキュメントは、CLUSTERPROをご利用になる局面や読者に応じて以下の通り分冊しています。初めてクラスタシステムを設計する場合は、システム構築ガイド【入門編】を最初に読んでください。

■ システム構築ガイド

【入門編】 (必須)設計・~~構築~~・~~運用~~・~~保守~~

クラスタシステムをはじめて設計・構築する方を対象にした入門書です。

【システム設計編(基本/共有ディスク,ミラーディスク)】 (必須)設計・構築・~~運用~~・~~保守~~

クラスタシステムを設計・構築を行う上でほとんどのシステムで必要となる事項をまとめたノウハウ集です。構築前に知っておくべき情報、構築にあたっての注意事項などを説明しています。
システム構成が共有ディスクシステムかミラーディスクシステムかで分冊しています。

【システム設計編(応用)】 (選択)設計・構築・~~運用~~・~~保守~~

設計編(基本)で触れなかったCLUSTERPROのより高度な機能を使用する場合に必要となる事項をまとめたノウハウ集です。

【クラスタ生成ガイド(共有ディスク,ミラーディスク)】 (必須) ~~設計~~・構築・~~運用~~・~~保守~~

CLUSTERPROのインストール後に行う環境設定を実際の作業手順に沿って分かりやすく説明しています。
システム構成が共有ディスクシステムかミラーディスクシステムかで分冊しています。

【運用/保守編】 (必須) ~~設計~~・~~構築~~・運用・保守

クラスタシステムの運用を行う上で必要な知識と、障害発生時の対処方法やエラー一覧をまとめたドキュメントです。

【GUIリファレンス】 (必須) ~~設計~~・~~構築~~・運用・~~保守~~

クラスタシステムの運用を行う上で必要なCLUSTERPROマネージャなどの操作方法をまとめたリファレンスです。

【コマンドリファレンス】 (選択)設計・構築・運用・~~保守~~

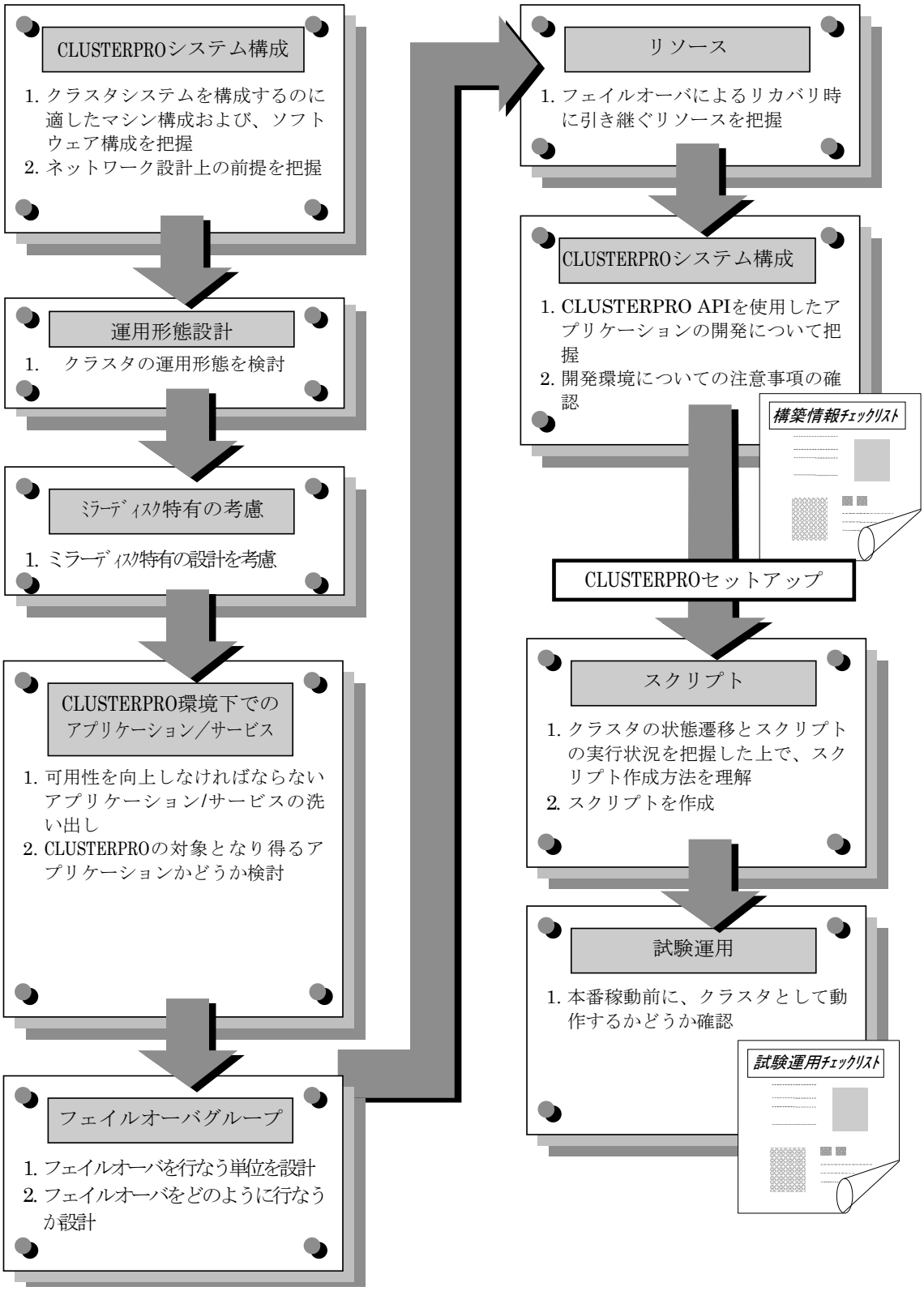
CLUSTERPROのスクリプトに記述できるコマンドやサーバから実行できる運用管理コマンドについてのリファレンスです。

【トレッキングツール編】 (選択) ~~設計~~・構築・~~運用~~・~~保守~~

CLUSTERPROトレッキングツールの操作方法を説明したリファレンスです。
分冊(GUI、システム構成(共有ディスクシステム、ミラーディスクシステム))しています。

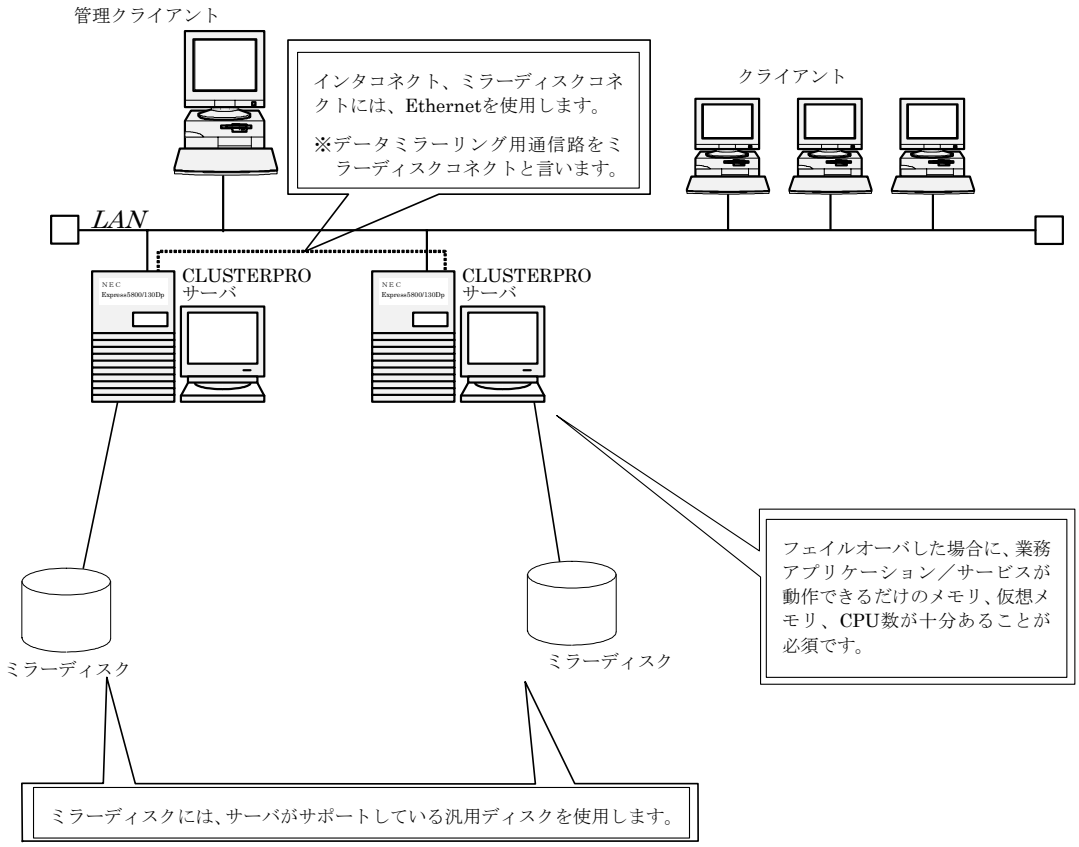
目次

1　CLUSTERPROシステム設計手順



2　CLUSTERPROシステム構成

2.1　マシン構成



2.1.1 クラスタサーバ

* 対象機種内の異なるモデル間での接続が可能です。CLUSTERPROの対象機種およびモデルについては、製品通知を確認してください。
* フェイルオーバした場合に、業務アプリケーションが動作できるだけのメモリ、仮想メモリ、CPU数が充分あることが必須です。
* インタコネクトには、以下の規則があります。
  + １クラスタシステムに対して、最小2、最大16です。
  + プライマリインタコネクトはパブリックLANとの共用できません。

2.1.2 管理クライアント

* CLUSTERPROマネージャをインストールするマシンを管理クライアントと呼びます。
* 管理クライアントのOSは、Windows 95/98/Me, Windows NT 4.0, Windows 2000, Windows XP のいずれかが必要です。

2.1.3 監視用クライアント

* Webマネージャを動作させるマシンを、監視用クライアントと呼びます。
* 監視用クライアントのOSは、Windows 95/98/Me, Windows NT 4.0, Windows 2000, Windows XPです。

2.1.4 ミラーディスク

* ミラーディスクには、各サーバがサポートしている汎用ディスクを使用します。
* 両サーバで同一のタイプのディスクを準備してください。
  片方のサーバがSCSIタイプのディスク、他方のIDEタイプのディスクという運用はサポートしていません。
* 両サーバでミラー用のディスクまたはLUNが同じデバイス名で見える構成にしてください。
* OSが使用している(OSの/etc/fstabなどで制御している)ディスクはミラーの対象にはできません。ミラー用にディスクを増設してください。
* ハードウェアRAIDを使用する場合には、ミラー専用のLUN(RAIDボードベンダによってはパック、システムディスクなどという表現をします)を確保してください。

2.1.5 インタコネクト

* 100BASE-TXのEthernetを使用します。

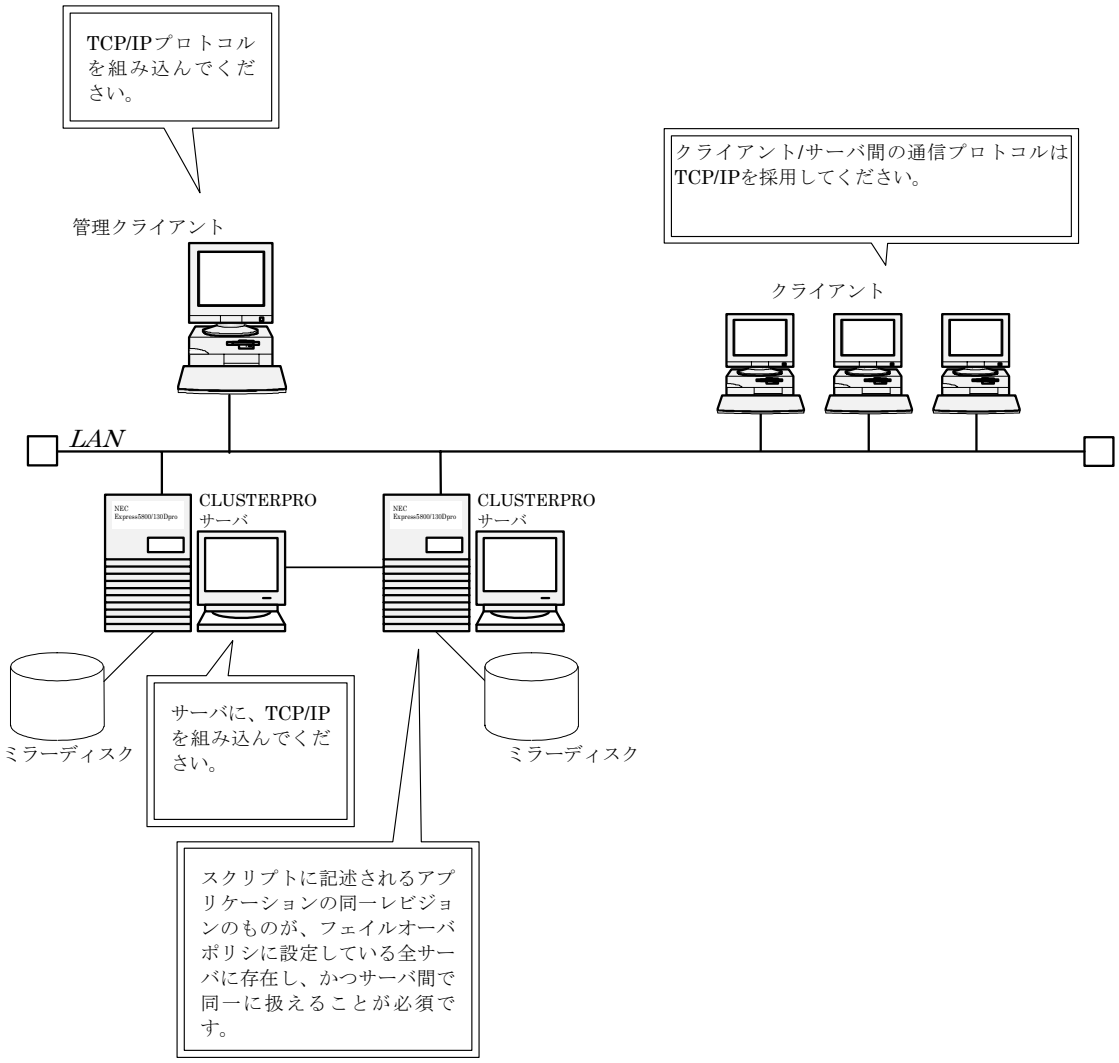
2.1.6 ミラーディスクコネクト

* 100BASE-TX以上のスペックのEthernetを使用します。
* クロスケーブルなどで両サーバを直結してください。
  サーバ間にHUBなどを接続するとディスクのI/O性能が低下します。
* ミラーディスクコネクトは、インタコネクトと兼用になります。Public LANと合わせて、1サーバ当たりの最低必要なLANボードは2枚となります（下表参照）。

Public-LAN用	Ethernet
インタコネクト用	上記とは別のEthernet
ミラーディスクコネクト用	
1サーバで最低必要なLANボード数	2

2.2 ソフトウェア構成

下図は、CLUSTERPROを導入する場合のソフトウェア構成の一例です。



2.2.1 動作環境

* CLUSTERPROサーバのインストールが必要です。
* スーパークラスタ直下には、サブクラスタを合わせて128個配置することができます。
* サブクラスタは、フェイルオーバ型クラスタを構成します。
* システム構築の際に以下の注意点を考慮する必要があります。

  = 拡張性(サポートサーバ数: 2)
  3サーバ以上のクラスタシステムを構築することはできません。

  = 書き込み性能
  ミラーディスクはネットワークを介して書き込みデータを相手サーバに送るため、通常のディスクを使用した場合に比べてデータの書き込時にミラーのためのオーバヘッドが発生します。
  （このため、ディスクに対する更新処理が多い業務には不向きです。）

  = 障害発生後の復旧
  サーバ障害発生後の復旧の際にはミラー再構築が必要な為、共有ディスク装置を使用したクラスタシステムに比べ復旧に要する時間が長くなります。
  又、ネットワークパーティション発生時等には手動でデータ復旧を行う必要があります。

2.2.2 クラスタサーバ

2.2.2.1 ミラーディスクに関する注意事項

* ミラーディスクによる運用の場合、2サーバの構成となります。
* ミラーディスクの同一パーティションに対して、同一マウントポイントにマウントされるように設定してください。
* ミラーディスクには、以下の規則があります。[1]
  + １クラスタシステムに対して、ミラーセットは最大8までです。
  + １ミラーセットについて、クラスタパーティションは必ず1つは必要です。また、最初のパーティションがクラスタパーティションになります。
  + 1クラスタシステムに対して、切替パーティションは最大120個です。
  + 切替ミラーパーティションのファイルシステムはext2またはext3にしてください。
  + 一台のディスクに作成できるパーティションの数はディスクデバイスに依存します。
    但し、各切り替えミラーディスクの第１パーティションはCLUSTERシステム処理用に使われるCLUSTERパーティションとなり一般ユーザからのアクセスは行えません。
    このCLUSTERパーティションはディスクの先頭に基本パーティションとして作成してください。
    また、このパーティションへのファイルシステムの構築は不要です。
  + OSのソフトウェアRAIDを用いたディスクはミラーセットには使用しないでください。

2.2.2.2 ネットワーク環境に関する注意事項

* TCP/IPを組み込む必要があります。
* IPアドレスには、以下の規則があります。
  + １サーバに対して最大16までです(フローティングIPアドレスを除く)。
  + １サーバ内に同一ネットワークアドレスに属するIPアドレスが複数存在してはいけません。
    また、以下のように包含関係にあってもいけません。
    IPアドレス：10.1.1.10, サブネットマスク：255.255.0.0
    IPアドレス：10.1.2.10, サブネットマスク：255.255.255.0

[1] トレッキングツール使用時には制限が発生します。トレッキングツール編を参照してください。

2.2.2.3 クラスタ設定に関する注意事項

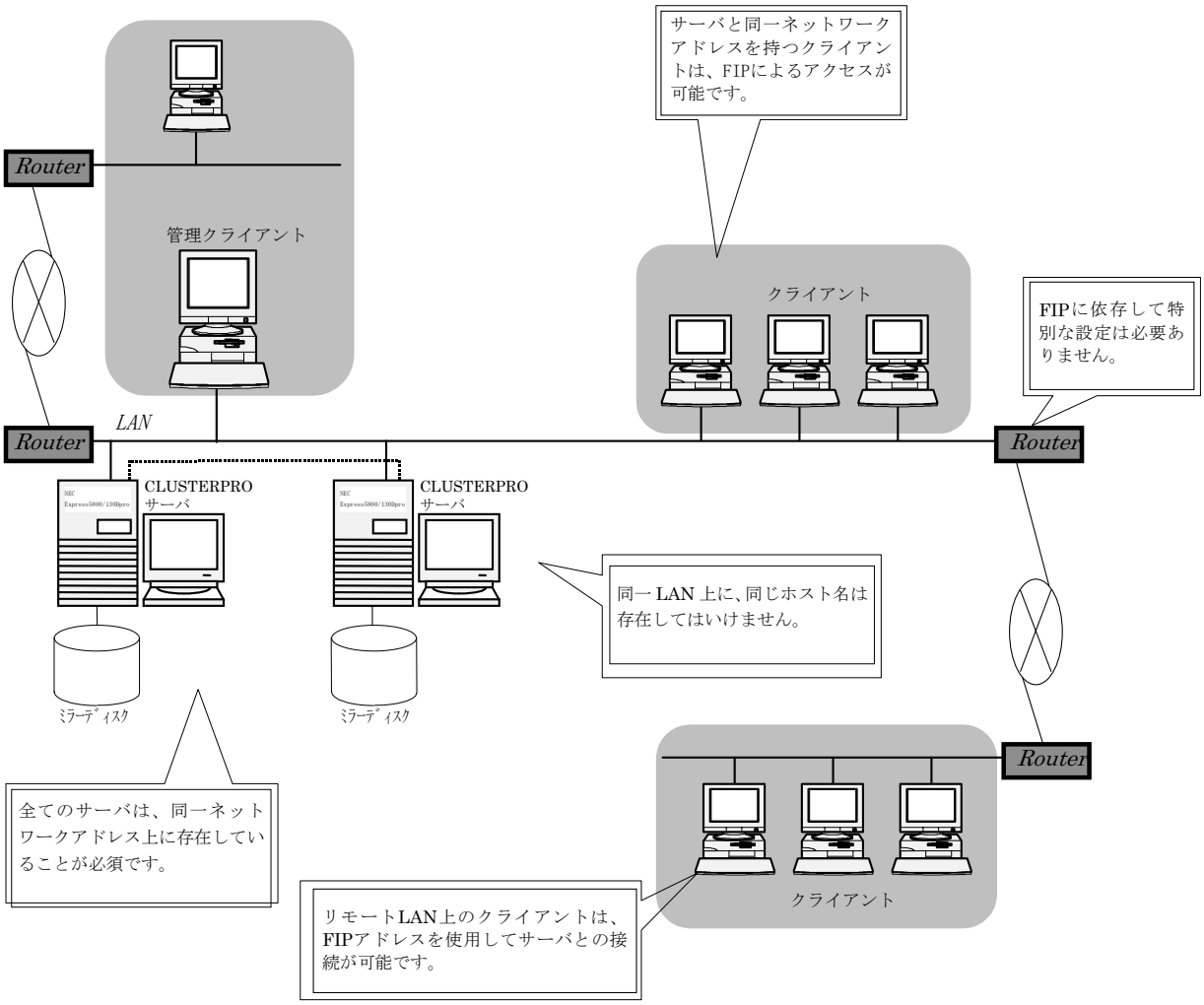
* スーパークラスタ名、サブクラスタ名には、以下の規則があります。
  + 1バイトの英大文字・小文字，数字，ハイフン(-)，アンダーバー(_)のみ使用可能です。
  + 英大文字・小文字の区別はありません。
  + 最大15文字(15バイト)までです。
  + 各クラスタシステムに対して、一意な名前でなければなりません。
* サーバ名には、以下の規則があります。
  + OSで設定可能なコンピュータ名と同じ規則があります。
  + 大文字・小文字は区別します。
  + 最大15文字(15バイト)までです。
* グループには、以下の規則があります。
  + フェイルオーバグループ数は１サブクラスタに対して最大64までです。
  + グループ名には、以下の規則があります。
    = 1バイトの英大文字・小文字, 数字, ハイフン(-), アンダーバー(_)のみ使用可能です。
    = 大文字・小文字の区別はありません。
    = 最大15文字(15バイト)までです。
    = クラスタシステム内で一意な名前でなければなりません。
    = PRNなどのDOS入出力デバイス名は使用できません。
  + スーバークラスタ内で、一意な名前でなければなりません。
* クラスタパスワードついては、以下の規則があります。
  + パスワード長は最大15バイトまでです。
* スクリプトに記述されるアプリケーションの同一レビジョンのものが、フェイルオーバポリシに設定されている全サーバに存在し、かつサーバ間で同一に扱えることが必須です。スクリプトの詳細については、「CLUSTERPRO システム構築ガイド システム設計編(応用)」を、フェイルオーバポリシについては「6.2.5 フェイルオーバポリシ」を参照してください。
* フローティングIPアドレスには、以下の規則があります。
  + １クラスタシステムに対して最大64までです。
  + サーバのパブリックLANと同一ネットワークアドレス内で使用していないホストアドレスを割り当てる必要があります。
* リソース監視については、以下の原則があります。
  + １つのグループ内のリソース監視で監視できるLANの数は、最大16までです。
    （フローティングIPアドレスは、IPアドレス毎の個別の監視設定ができませんので上記の数には含めません。）

2.2.3 管理クライアント

* TCP/IPを組み込む必要があります。
* １つのCLUSTERPROマネージャが管理できるスーパークラスタは最大8です。
* １つのCLUSTERPROマネージャが管理できるサブクラスタは、全スーパークラスタを合計して最大128までです。
* １つのクラスタシステムに接続できるCLUSTERPROマネージャ数は、クラスタシステム内の１サーバ当たり最大32までです。

2.3　ネットワーク設計

2.3.1　ネットワークの概要



フローティングIP(FIP)については「7.2 フローティングIPアドレス」を参照してください。

2.3.2 クラスタサーバ

* クラスタを構成するサーバは、同一LAN上に存在し、同一ネットワークアドレスで構成していることが必須です。
* インタコネクトLAN、パブリックLANは、異なるネットワークアドレスである必要があります。
  インタコネクトLANのIPアドレスは、プライベートIPアドレスでも可能です。

2.3.3 管理クライアント

* CLUSTERPROマネージャのインストールが必要です。
* クラスタサーバと同一LAN上に存在する必要もありません。

2.3.4 ルータ

* フローティングIPのために特別な設定は必要ありません。

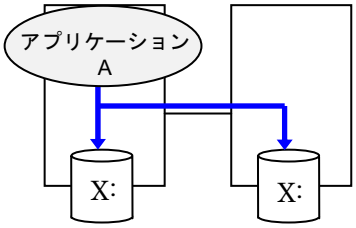
3　運用形態設計

ミラーディスクシステムを設計する場合、通常のクラスタシステムでの設計に加えて考慮する点があります。ここではミラーディスクシステム設計の際に考慮すべき項目に関して説明しています。

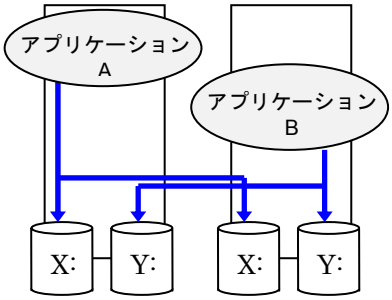
3.1　ミラーディスク運用形態

ミラーディスクを用いたクラスタシステムでは、以下の運用形態でシステムを構築することが可能です

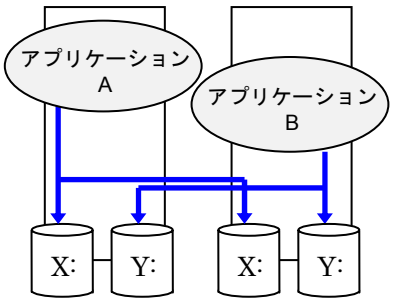
+ 片方向スタンバイ
  クラスタシステム全体で同一の業務アプリケーションが1つしか動作しないシステム形態



+ 同一アプリケーション双方向スタンバイ
  クラスタシステム全体で同一の業務アプリケーションが複数動作するシステム形態

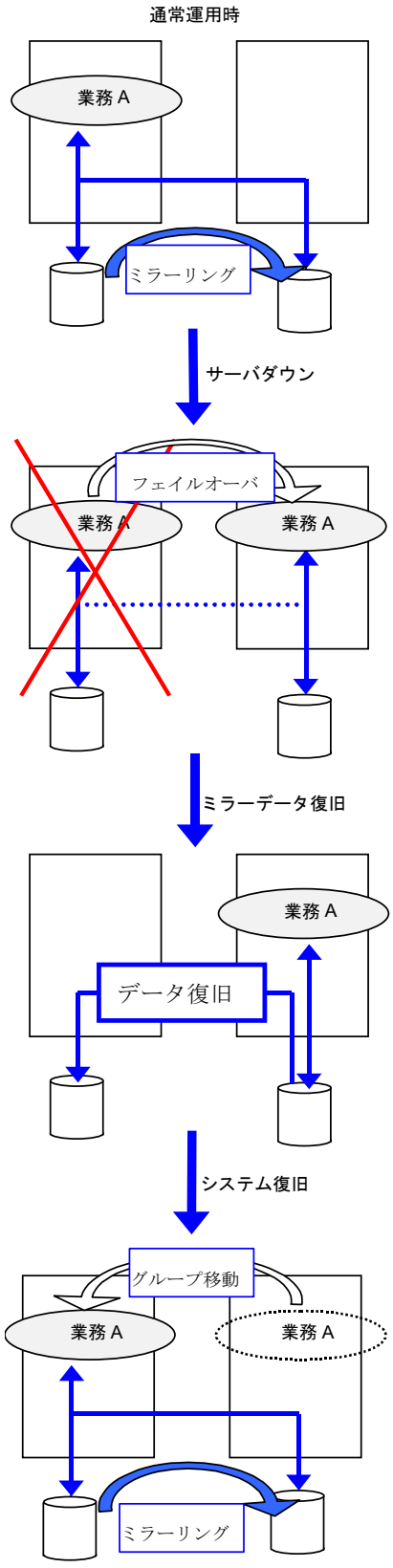


+ 異種アプリケーション双方向スタンバイ
  複数の種類の業務アプリケーションが、それぞれことなるサーバで稼動し、相互に待機するシステム形態



3.1.1 片方向スタンバイ

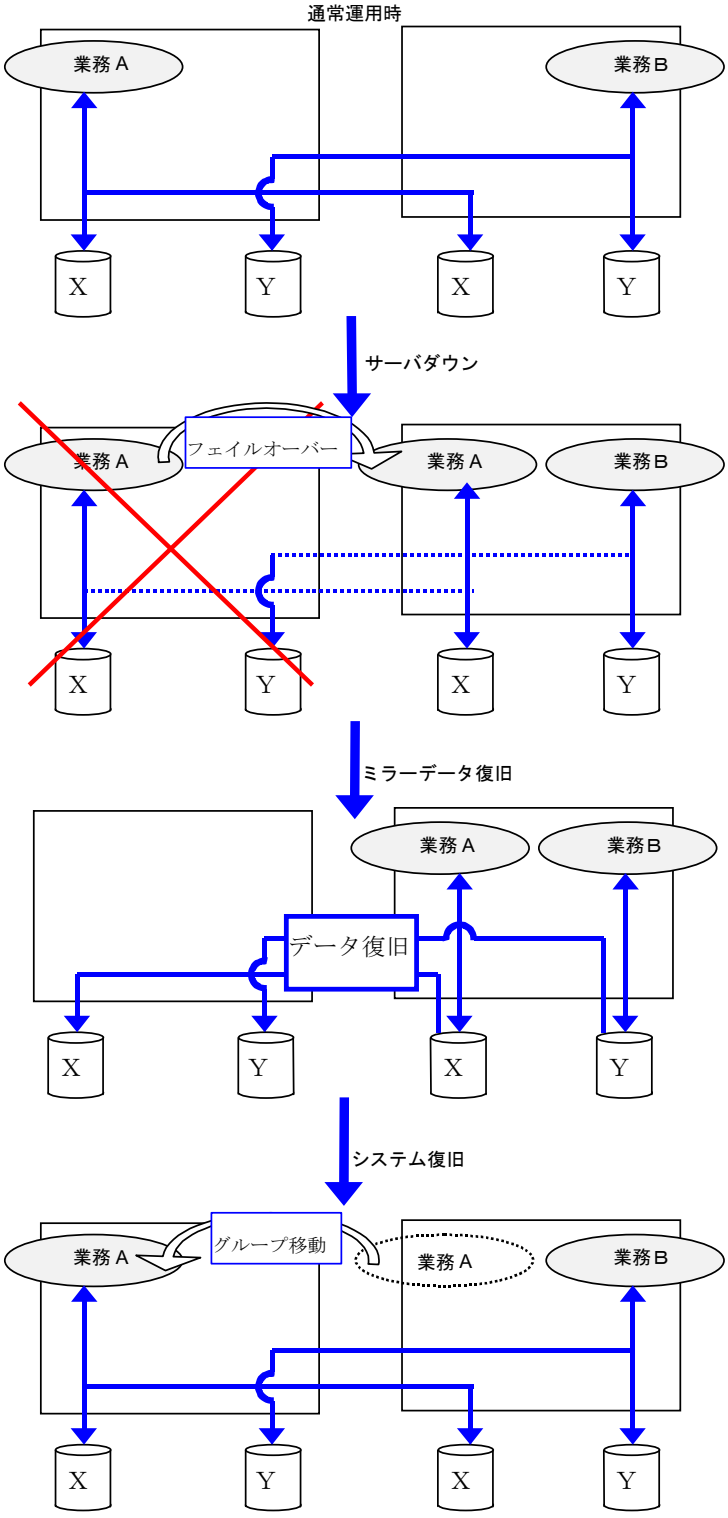
片方向スタンバイとは、ある業務についてフェイルオーバグループを1グループに制限したクラスタシステムです。



3.1.2 双方向スタンバイ

双方向スタンバイとは、ある業務が複数のサーバ上で同時に動作することが可能なクラスタシステムです。

双方向スタンバイには、同じアプリケーションが、複数のサーバ上で動作する、「同一アプリケーション双方向スタンバイ」と、複数の種類のアプリケーションが、複数のサーバ上で動作する「異種アプリケーション双方向スタンバイ」があります。



4　ミラーディスク特有の考慮

4.1　ディスクについて

4.1.1　ディスクの選択

ミラーセットを構成するディスクは同一容量を持つものにしてください。
またサーバ、SCSI アダプタも同一のものにする事を推奨します。
ディスク障害等が発生した場合にそなえ、より高い安全性を持つアレイディスクの利用をおすすめします。

4.1.2　ディスクの追加

以下のディスクはミラーセットとしては使用できません。サーバにこのようなディスクしか接続されていない場合には、別途ミラーリング専用のディスクの追加が必要です。
・ＯＳシステムドライブが存在するディスク
・ページングファイルが存在するディスク
・リムーバブルディスク
また、双方向スタンバイとして使用する場合には、ディスク単位のミラーリングを行うため４台以上のディスク（ミラーセットが２つ）が必要です。

4.1.3　ディスク上のパーティション

ミラーセットを構成するディスクの先頭パーティションは、ミラーディスク管理用（CLUSTERパーティション）として使用されます。
このCLUSTERパーティションには、10MB以上のサイズが必要です。また、このパーティションは基本パーティションとして確保し、フォーマットを行わないでください。
このCLUSTERパーティション以外のパーティションが、ミラー対象となりユーザからの利用が可能です。ミラー再構築時間/業務内容を考慮の上、パーティションサイズを決定してください。

4.1.4　ディスク性能

ネットワークを介したIOを行うためディスクIOにはオーバヘッドが発生します。
ただし通常はファイルシステムを経由してI/Oを行うため、負荷の低い(書き込み頻度の少ない)業務を行っている限り性能低下を意識する必要はありません。
又、アレイディスク使用時にはDISK CACHEをWRITE THRU にすると、性能低下が大きくなるのでWRITE BACK での使用をお勧めします。但し、WRITE BACKで使用する場合は、アレイボード上にバッテリがあるか、UPSが接続されている必要があります。

4.1.4.1　通常(ミラー)運用中のディスクI/O性能

ミラー運用時にはRead対Writeの比率が約2:1のケースで、約20%ほど性能が低下します。
Writeの比率が増えると、更に性能の低下が発生します。

4.1.4.2　ミラー再構築中のディスクI/O性能

ミラー再構築中には、I/OパターンがRead対Write=2:1のケースで、通常運用中よりもさらに約15%ほど性能が低下します。
ミラー再構築中には特に書き込みの多いアプリケーションの処理速度に影響が出ますので、システムを構築する上で十分に注意してください。

4.1.5　アレイディスクのミラーセット

アレイディスクでミラーセットを構成する場合、ミラーリングはアレイ上の構成されたシステムドライブ単位(LUN単位)となります。ＯＳシステムドライブがアレイの異なるシステムドライブに存在しても構いません。また、両サーバのＲＡＩＤが異なっているとミラーセットは構成できません。アレイとシングルディスクの組み合わせも構成できません。

4.2　ネットワークについて

4.2.1 ミラーディスクコネクト[2]の追加

ミラーディスクコネクトは、インタコネクトと同一のLANを用いる為、ミラーディスクコネクト専用のLANは用意する必要はありません。

4.2.2 インタコネクト設定

ミラーディスクを使用したクラスタシステムでは、共有ディスクを使用したクラスタシステムと違い、CLUSTERパーティションを利用したネットワークパーティション解決処理を行うことができません。そこでネットワークパーティションの発生を極力避ける為に、CLUSTERPROマネージャから、全てのパブリックLANをインタコネクトに指定することをお勧めします。
これによりネットワークパーティション発生の際の問題を減少させることができます。

4.3　障害復旧時間について

ミラーディスクを使用したクラスタシステムの場合には、共有ディスクを使用したクラスタシステムに比べ復旧時に要する時間が長くかかります。これは障害サーバやスナップショットバックアップのためにクラスタから切り離されたサーバを、クラスタへ復帰させるとき、復旧時にミラーの再構築を行う為です。
ミラー再構築が完了するまでの間、片サーバのみでの運用となり可用性が低下した状態であるため、システムの設計時点でミラー再構築時間を考慮しておく必要があります。

4.3.1 ミラー構築時間

構築時間に関しては、下表を目安としてください。
ただしこの値は、サーバ性能、ディスク性能及びLAN性能により異なってきます。
また、再構築中に業務を運用している場合には、構築時間が下表よりも長くなる場合があります。

１ＧＢあたりの構築時間

単体ディスク(非アレイディスク)	約　6分30秒 ～
アレイディスク・RAID5(WRITE THRU)	約 11分 ～
アレイディスク・RAID5(WRITE BACK)	約　4分 ～

注：WRITE BACK , WRITE THRU はアレイディスクのDISK CACHEの設定を示す。

[2] ミラーディスクの通信路をミラーディスクコネクトと呼びます。

4.4　その他の考慮

4.4.1 起動スクリプト設定

CLUSTER動作時に、ミラーセットの整合性がとれていない状態で最新のデータを保持してない側のサーバでは、切り替えディスクの起動を成功しないようにしています。
そこでフェイルオーバグループのプライマリサーバ(環境変数 ARMS_EVENT=START、ARMS_SERVER = HOME で起動スクリプトが実行された場合)にて、切り替えミラーディスクの接続に失敗した場合は、フェイルオーバを行うコマンド(ARMFOVER)をスクリプトに記述し、待機サーバでの業務継続を可能にすることをお勧めします。

5 CLUSTERPRO環境下でのアプリケーション/サービス

ここでは、CLUSTERPRO環境下で動作できるアプリケーション/サービスについて、留意すべき事項を述べます。

5.1 業務の洗い出し

CLUSTERPROを導入する場合、まず可用性を向上しなければならないアプリケーション/サービスを、洗い出す必要があります。また、洗い出したアプリケーション/サービスが、CLUSTERPROの環境下で動作するのに適しているかどうかを、見極めなければなりません。

洗い出したアプリケーション/サービスが、CLUSTERPROでのクラスタ対象として適しているかどうかは、次節からの内容を十分検討して判断してください。

5.2 CLUSTERPRO環境下でのアプリケーション/サービス

5.2.1 サーバアプリケーション

対象アプリケーションがどのようなスタンバイ形態で実行するかで注意事項が異なります。

* 片方向スタンバイ[運用-待機] 注意事項: 1 2 3 ~~4~~ 5
  クラスタ内で、あるアプリケーションの稼動サーバが常に一台である運用形態です。

* 双方向スタンバイ[運用-運用] 注意事項: 1 2 3 4 5
  クラスタ内で、あるアプリケーションの稼動サーバが複数台である運用形態です。

* 共存動作 注意事項: ~~1~~ ~~2~~ ~~3~~ ~~4~~ 5
  クラスタシステムによるフェイルオーバの対象とはせず、共存動作する運用形態です。

5.2.2 サーバアプリケーションについての注意事項

(1) 障害発生後のデータ修復

障害発生時にアプリケーションが更新していたファイルは、待機系にてアプリケーションがそのファイルにアクセスするときデータとして完結していない状態にある場合があります。

非クラスタ(単体サーバ)での障害後のリブートでも同様のことが発生するため、本来アプリケーションはこの状態に備えておく必要があります。クラスタシステム上ではこれに加え人間の関与なしに(スクリプトから)復旧が行える必要があります。

CLUSTERPROのフェイルオーバのタイミングでファイルシステムにfsckが必要な場合には、CLUSTERPROがfsckを行います。

(2) アプリケーションの終了

CLUSTERPROが業務グループを停止・移動(オンラインフェイルバック)する場合、その業務グループが使用していたファイルシステムをアンマウントします。このため、アプリケーションへの終了指示にて、切替ミラーディスク上の全てのファイルに対するアクセスを停止する必要があります。

通常は終了スクリプトでアプリケーション終了指示コマンドを実行しますが、終了指示コマンドが(アプリケーションの終了と)非同期で完了してしまう場合注意が必要です。(例えばarmsleepコマンドによって一定時間待ち合わせするなど)

(3)　データ格納位置

CLUSTERPROがサーバ間で引き継ぐことのできるデータは次の通りです。

＝ 切替ミラーディスク上のデータ

アプリケーションはサーバ間で引き継ぎたいデータと引き継ぎたくないデータを分離できる必要があります。

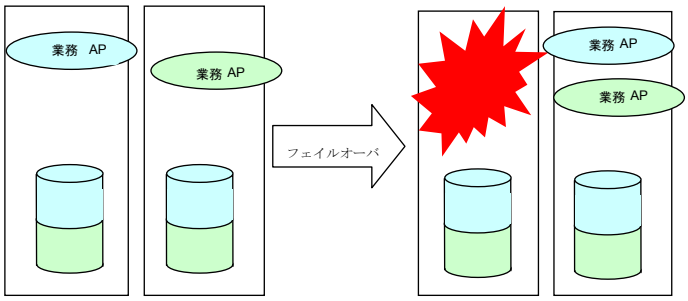
データの種類	(例)	配置場所
引き継ぎたいデータ	(ユーザデータなど)	切替ミラーディスク
引き継ぎたくないデータ	(プログラム，設定情報など)	サーバのローカルディスク

(4)　複数業務グループ

双方向スタンバイの運用形態では、(障害による縮退時)、1つのサーバ上で同一APIによる複数業務グループが稼動することを想定しなくてはなりません。

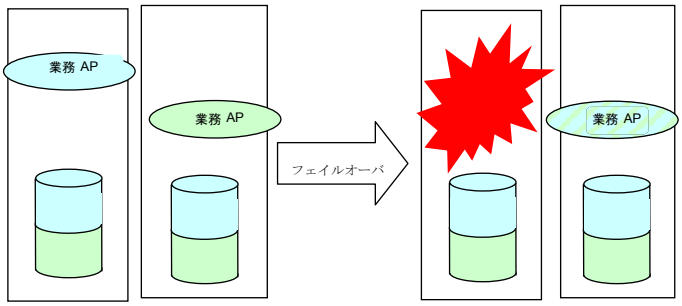
アプリケーションは次のいずれかの方法で引き継がれた資源を引き取り、単一サーバ上で複数業務グループを実行できなければなりません。

□　複数インスタンス起動
新たに別インスタンス(プロセス)を起動する方法です。アプリケーションが複数動作できる必要があります。



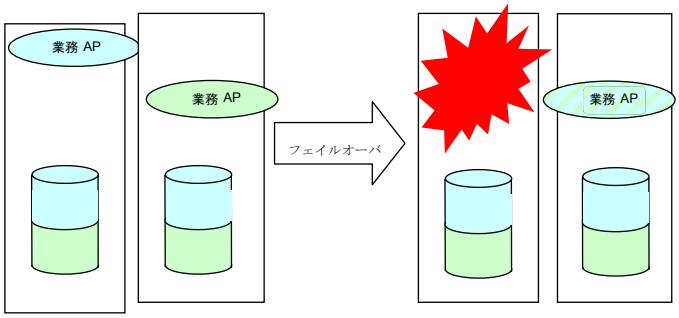
アプリケーション再起動
もともと動いていたアプリケーションを一旦停止し、再起動することで、追加された資源を扱えるようにする方法です。

□



業務 AP を再起動することで、
データを引き継ぐ

□　動的追加
動作中のアプリケーションに対
□　して、自動またはスクリプトか
□　らの指示により資源を追加する
□　方法です。



実行中の業務 AP に動的に
データを追加することで
データを引き継ぐ

(5) アプリケーションとの相互干渉(相性問題)

CLUSTERPROの機能や動作に必要なOS機能との相互干渉によってアプリケーションまたはCLUSTERPROが動作できない場合があります。

* I/Oフィルタリング
  CLUSTERPRO はサーバ間でのミラーディスクのアクセス権利を切り替えるために、I/O フィルタリングを行い非活性状態のディスクへの I/O 要求に対しては"EIO"のエラーを返却します。

* アプリケーションは非活性状態の(つまりアクセス権利のない)切替ミラーディスクにアクセスしてはいけません。
  通常クラスタスクリプトから起動されるアプリケーションは、それが起動された時点でアクセスすべき切替ミラーディスク上のパーティションが既にアクセス可となっていることを想定してかまいません。

* アプリケーションの切替ミラーディスクへのアクセス
  共存動作アプリケーションには、業務グループの停止が通知されません。もし、業務グループの停止のタイミングでそのグループが使用している切替ミラーディスク上のパーティションにアクセスしている場合、アンマウントに失敗してしまいます。

  システム監視サービスを行うようなアプリケーションの中には、定期的に全てのディスクパーティションをアクセスするようなものがあります。この場合、監視対象パーティションを指定できる機能などが必要になります。

* マルチホーム環境およびIPアドレスの移動
  クラスタシステムでは、通常、一つのサーバが複数の IP アドレスを持ち、ある IP アドレス(フローティング IP アドレスなど)はサーバ間で移動します。

  問題点の多くはアプリケーションが gethostbyname で返却される IPアドレスが一つしかないことを前提に作成されている場合に起こります。

5.3 業務形態の決定

5章全体を踏まえた上で、業務形態を決定してください。

* どのアプリケーション/サービスをいつ起動するか
* 起動時やフェイルオーバ時に必要な処理は何か
* 切替ミラーディスクに置くべき情報は何なのか

また、以下を運用の中に必ず組み込んでください。

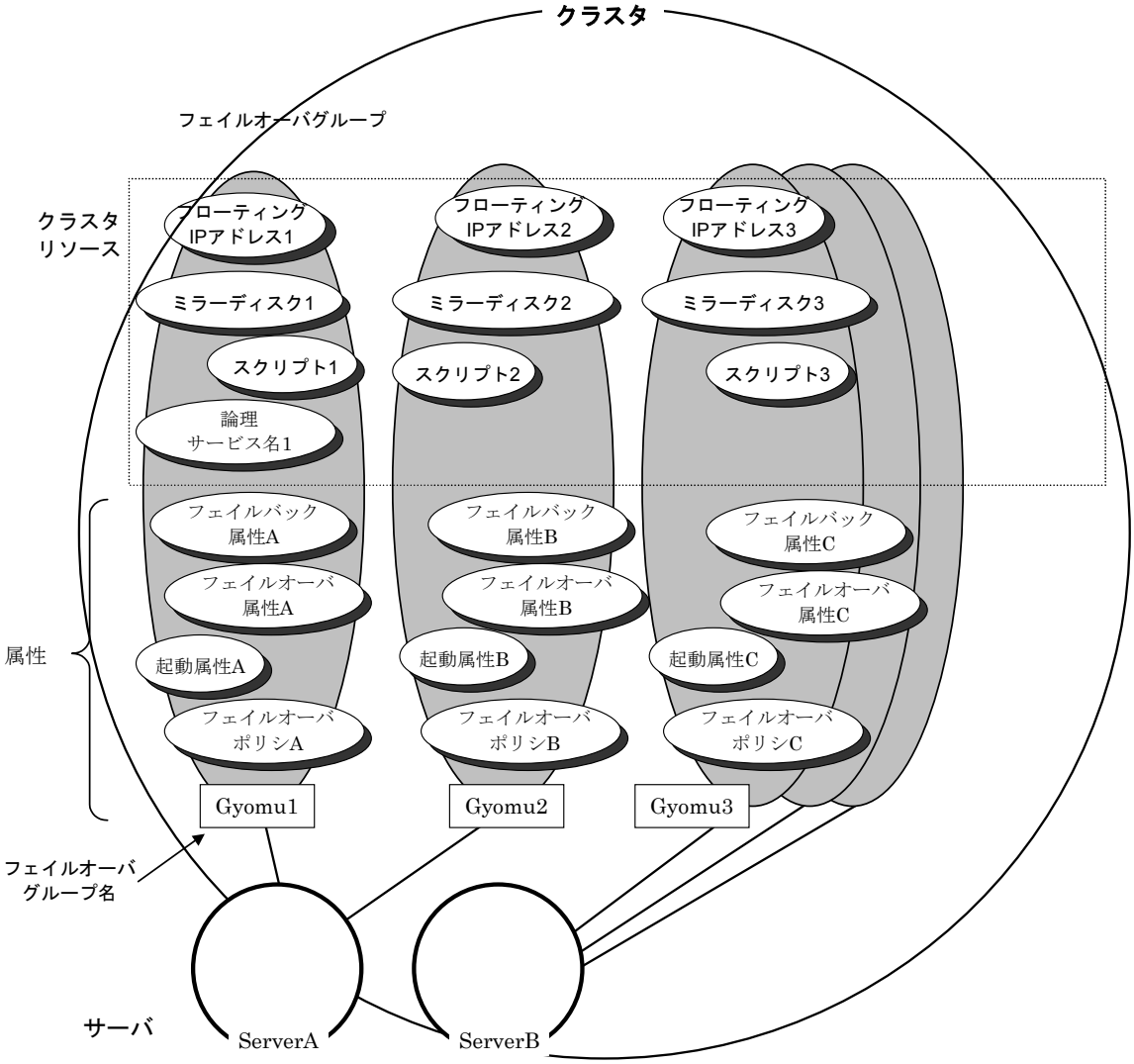
* 切替ミラーディスクの定期的なバックアップ

6　フェイルオーバグループ

フェイルオーバグループとは、クラスタシステム内のある1つの独立した業務を実行するために必要な資源の集まりのことで、フェイルオーバを行なう単位になります。

フェイルオーバグループは、フェイルオーバグループ名、クラスタリソース、属性を持ちます。

1クラスタシステムに対して64フェイルオーバグループまで作成することができます。



各フェイルオーバグループのクラスタリソースは、それぞれひとまとまりのグループとして処理されます。すなわち 、ミラーディスク1とフローティングIPアドレス1を持つGyomu1においてフェイルオーバが発生した場合、ミラーディスク1とフローティングIPアドレス1がフェイルオーバすることになります(ミラーディスク1のみが、フェイルオーバすることはありません)。

また、ミラーディスク1は、他のフェイルオーバグループ(たとえばGyomu2)に含まれることはありません。

6.1 クラスタリソース

フェイルオーバグループは以下のクラスタリソースを所有することができます。
詳細は「7 リソース」を参照してください。

* 切替ミラーディスク
* フローティングIPアドレス
* リソース監視
* スクリプト

6.2 属性

フェイルオーバグループは以下の属性を所有します。

* 起動属性
* フェイルオーバ属性
* フェイルバック属性
* フェイルオーバポリシ

6.2.1 フェイルオーバグループ名

フェイルオーバグループの名前です。
以下の規則があります。

* 1バイトの英大文字/小文字，数字，ハイフン(-)，アンダーバー(_)のみ使用可能
* 大文字/小文字の区別なし
* 最大15文字(15バイト)
* スーパークラスタ内で一意な名前

6.2.2 起動属性

クラスタ起動時にCLUSTERPROによりフェイルオーバグループを自動的に起動するか（自動起動）、もしくはCLUSTERPROマネージャからユーザが操作して起動するか（手動起動）、の属性を指定します。

* 自動起動
  CLUSTERPROにより自動的に起動される。
  クラスタの起動時、フェイルオーバグループは自動的に起動される(活性状態)。
* 手動起動
  CLUSTERPROからは起動されず、ユーザによるCLUSTERPROマネージャからの起動指示により起動される。
  クラスタの起動時、フェイルオーバグループは、起動されない（非活性状態）。その後、CLUSTERPROマネージャから、ユーザが操作して起動される（活性状態）。

6.2.3 フェイルオーバ属性

フェイルオーバ先の決定規則を指定します。
決定規則として、常に一番優先順位の高いサーバにフェイルオーバするか（通常）、常にグループが起動されていないサーバにフェイルオーバするか（排他）、を選択できます。

* 通常
  CLUSTERPROにより自動的にフェイルオーバされる。フェイルオーバ先の決定規則は、常に一番優先順位の高いサーバ。同一サーバで複数のグループが起動されることがある。
  オフラインフェイルバックはあり。
* 排他
  CLUSTERPROにより自動的にフェイルオーバされる。フェイルオーバ先の決定規則は、排他のグループが起動されていないサーバのうち一番優先順位の高いサーバ。このときグループが起動されていないサーバが存在しなければ、フェイルオーバしない。サーバで複数のグループが起動されることはない。オフラインフェイルバックはなし。

6.2.4 フェイルバック属性

フェイルオーバポリシで設定した、最高プライオリティサーバが正常状態に戻ったとき、自動的に最高プライオリティサーバへフェイルバックするかどうかを指定します。

以下のどちらかを選択します。

* 自動フェイルバックする
* 自動フェイルバックしない

既定値は、「自動フェイルバックしない」となります。

6.2.5 フェイルオーバポリシ

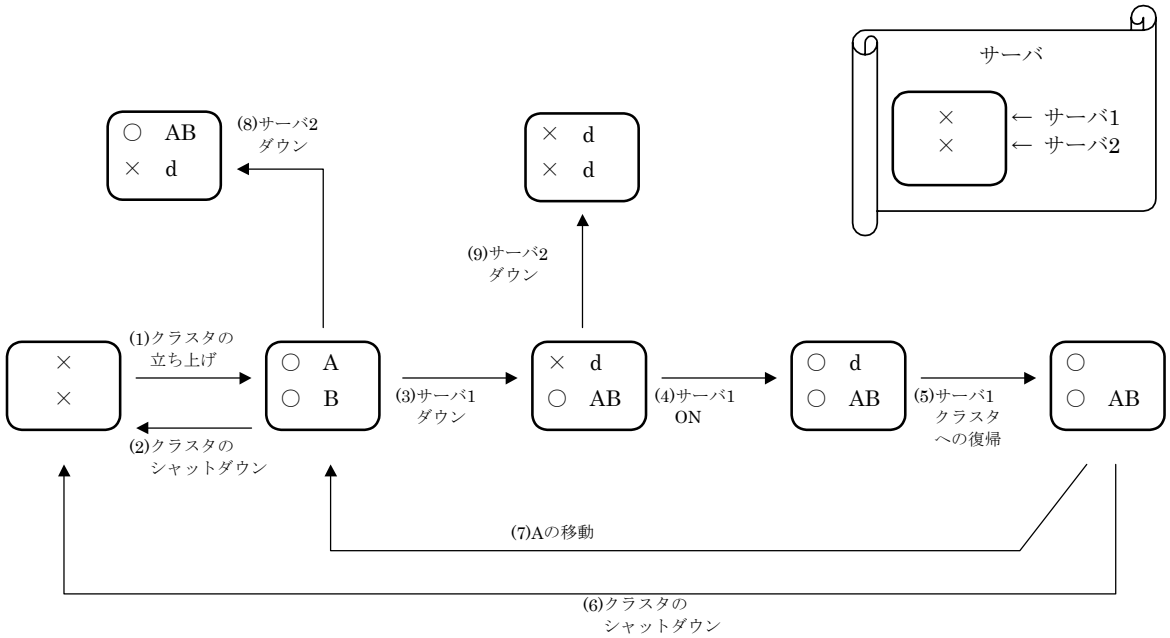
フェイルオーバ可能なサーバリストとその中でのフェイルオーバ優先順位です。フェイルオーバ発生時のフェイルオーバポリシによる動作の違いを説明します。

＜図中記号の説明＞

サーバ状態	説明
○	正常状態（クラスタとして正常に動作している）
×	停止状態（クラスタシャットダウンによる停止状態）
× d	ダウン状態（サーバダウンによる停止状態）
○ d	ダウン後再起動状態（クラスタから切り離されている）

フェイルオーバグループ	フェイルオーバポリシ		
	優先度1サーバ	優先度2サーバ	優先度3サーバ
A	サーバ1	サーバ3	サーバ2
B	サーバ2	サーバ3	サーバ1

フェイルオーバ属性が通常の場合



(1)　クラスタの立ち上げ
(2)　クラスタのシャットダウン
(3)　サーバ1ダウン　：　優先度2のサーバへフェイルオーバする
(4)　サーバ1の電源on
(5)　サーバ1のクラスタ復帰
(6)　クラスタのシャットダウン
(7)　フェイルオーバグループAの移動
(8)　サーバ2ダウン　：優先度2のサーバへフェイルオーバする
(9)　サーバ2ダウン

6.2.6 アプリケーション

クラスタに対応したアプリケーションは、“フェイルオーバ”または、“フェイルオーバグループの移動”が発生した場合に、スクリプトにより相手サーバで再起動されます。よって、同一レビジョンのアプリケーションがフェイルオーバポリシで設定してある全サーバに存在し、かつサーバ間で同一に扱えることが必須です。また、引き継ぐべきデータを共有ディスク上に集められるような性質のものでなくてはなりません。

CLUSTERPRO環境下で動作するアプリケーションは、この他にもいくつかの前提条件をクリアしたものでなければなりません。詳細については、「5. CLUSTERPRO環境下でのアプリケーション/サービス」を参照してください。

6.3 フェイルオーバ要因

フェイルオーバを引き起こす要因としては、以下のものがあります。

* サーバのシャットダウン
* 電源ダウン
* OSのパニック
* OSのストール
* CLUSTERPROサーバの異常
* スクリプトからのCLUSTERPROコマンド(armload)により起動したアプリケーションの障害
  + アプリケーションの障害とは、プロセスの消失を示します
  + armloadには、下記オプションが指定できます
    = 監視対象とする/しない
    = 再起動回数の閾(しきい)値
    = 再起動回数を0クリアするまでの時間
    = アプリケーション単体での再起動もしくはスクリプトからの再起動
    = しきい値を越えた場合の挙動（サーバシャットダウンもしくはフェイルオーバ）
* リソース監視により監視しているリソースおよびパブリックLANで、異常を検出した場合

7　リソース

7.1　ディスクリソース

7.1.1　切替ミラーディスク

切替ミラーディスクとは、クラスタを構成する2台のサーバ間でディスクデータのミラーリングを行うディスクのペアのことであり、CLUSTERPROのリソースとして動作します。

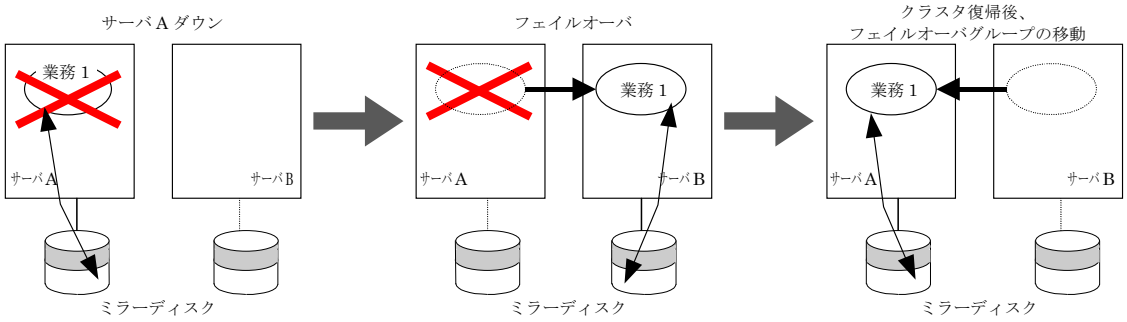
この切替ミラーディスクは共有ディスクを使用した切替ディスクと同様に１台のサーバ(通常はプライマリサーバ)からのみアクセス可能です。

切替は、フェイルオーバグループ毎に、フェイルオーバポリシにしたがって行われます。業務に必要なデータは、切替ミラーディスク上に格納しておくことで、フェイルオーバ時／フェイルオーバグループの移動時等に、自動的に引き継がれます。

切替ミラーディスクのファイルシステムは、必ずext2またはext3にしてください。また、全てのサーバで、同一のパーティションには、同一のマウントポイントを割り付けてください。

切替ミラーディスクの設定手順については、システム構築ガイド「クラスタ生成ガイド」「GUIリファレンス」を参照してください。

切替ミラーディスクは、クラスタを構成するサーバに、それぞれ接続されたミラーディスク上に作成されます。



7.1.2　CLUSTERパーティション

CLUSTERPROサーバが切替ミラーディスク制御のために使用する専用パーティションを、CLUSTERパーティションといいます。

CLUSTERパーティションは、RAWパーティションでなければいけません。 フォーマットは行わないでください。

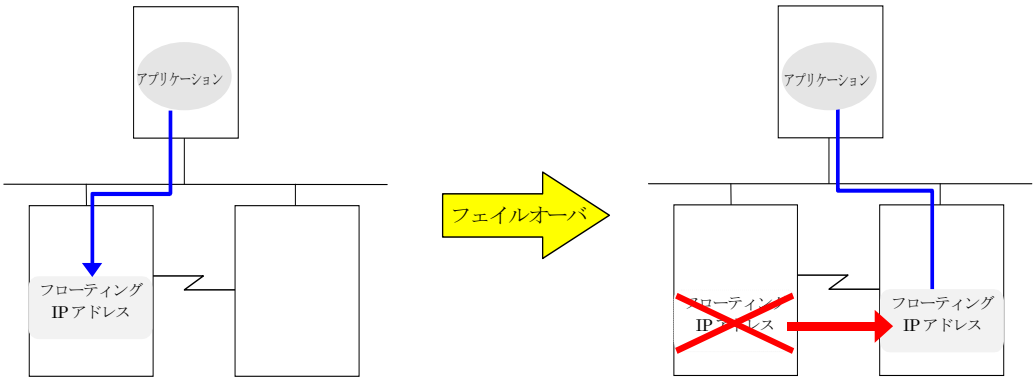
CLUSTERパーティションは､データミラーリング用ディスクの第一パーティションが自動的に割り当てられます。

このためにミラーを行うディスク上ディスクの先頭に、CLUSTERパーティション用の領域（最低10MB）を基本パーティションとして確保してください。

7.2　フローティングIPアドレス

クライアントアプリケーションは、フローティングIPアドレスを使用してクラスタサーバに接続することができます。また、サーバ間でも接続可能です。フローティングIPアドレスを使用することにより、“フェイルオーバ”または、“フェイルオーバグループの移動”が発生しても、クライアントは、接続先サーバの切り替えを意識する必要がありません。

フローティングIPアドレスは、同一LAN上でもリモートLANからでも使用可能です。ARPプロトコルをサポートしているOSであれば使用可能です。



フローティングIPの概要

使用可能IPアドレス	パブリックLANと同一ネットワークアドレス
切替方式	サーバからのARPブロードキャストにより、ARPテーブル上のMACアドレスが切り替わる
クライアントOS	選ばない
ルータ設定	不要
LAN多重化	不可
潜在リソース	フローティングIP設定で設定されたIPアドレス一覧
サーバ間での使用	可能(但しパブリックLANのみ)

7.2.1　アドレスの割り当て

フローティングIPアドレスに割り当てるIPアドレスは、以下の条件を満たす必要があります。

* クラスタサーバが所属するLANと同じネットワークアドレス内で かつ使用していないホストアドレス

この条件内で必要な数(一般的にはフェイルオーバグループ数分)のIPアドレスを確保してください。

このIPアドレスは一般のホストアドレスが変わらないため、インタネットなどのグローバルIPアドレスから割り当てることも可能です。

7.2.2 環境設定

フローティングIPアドレスを使用するには以下の設定が必要です。

+ CLUSTERPROマネージャ でフェイルオーバグループへ
  フローティングIPアドレスの割り当て

クラスタ生成後、CLUSTERPROマネージャの「フェイルオーバグループの追加」→「リソースの設定」→「IPアドレス」→「フローティングIP追加」により選択肢の中から使用するIPアドレスを選択してください。

フローティングIPアドレスの値を変更する場合には、「フェイルオーバグループのプロパティ」→「リソースの設定」→「IPアドレス」により新しいFIPを追加し古いFIPを削除してください。

7.2.3 経路制御

サーバに使用するネットワークIPアドレスの経路制御で フローティングIPアドレスの経路制御も行われますので、フローティングIPアドレスのための特別な経路制御は不要です。

7.2.4 使用条件

以下のマシンからフローティングIPアドレスにアクセスできます。

* クラスタサーバ自身
* 同一クラスタ内の他のサーバ、他のクラスタシステム内のサーバ
* クラスタサーバと同一LAN内 及び リモートLANのクライアント

さらに以下の条件であれば上記以外のマシンからでもフローティングIPアドレスが使用できます。[3]

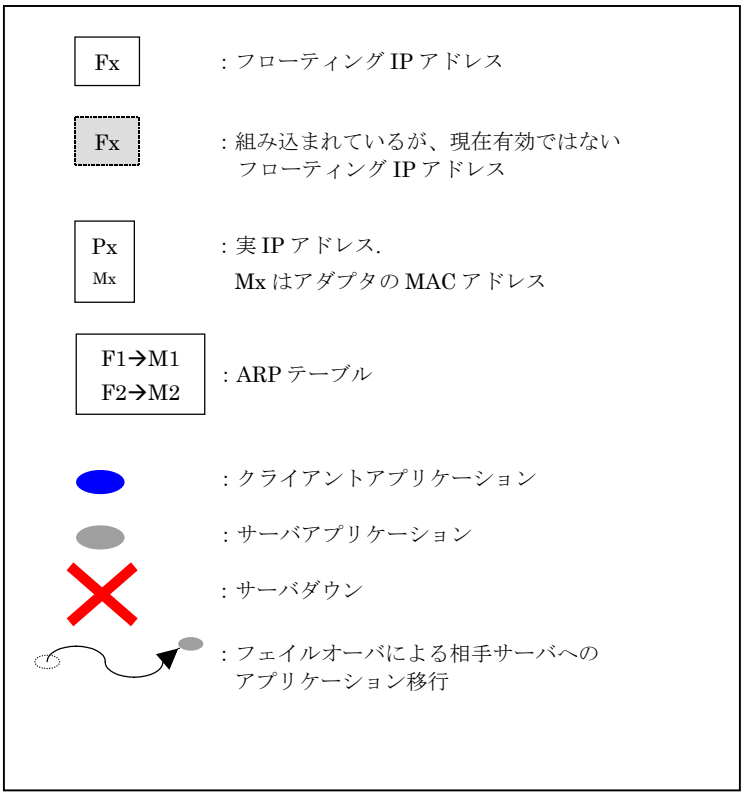
* 通信プロトコルがTCP/IPであること
* ARPプロトコルをサポートしていること

スイッチングHUBにより構成されたLANであっても、フローティングIPアドレスのメカニズムは問題なく動作します。
サーバダウン時には、接続していたTCP/IPコネクションは切断されます。
リモートLAN上のマシンからも、フローティングIPアドレスにアクセスできます。

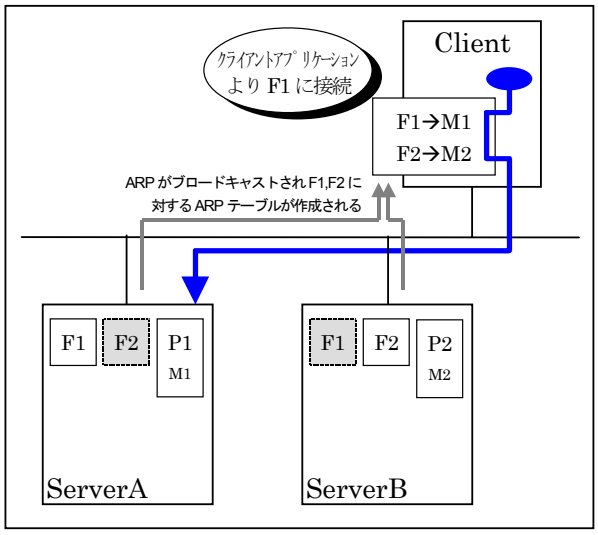
[3] すべてのマシン,アーキテクチャの接続を保証できません。事前に充分に評価をしてください。

7.2.5　フローティングIPアドレスによる接続形態

FIPアドレスによる接続形態を説明します。図中で使用される記号については、以下のように定義します。



7.2.5.1　クライアントからサーバへの接続

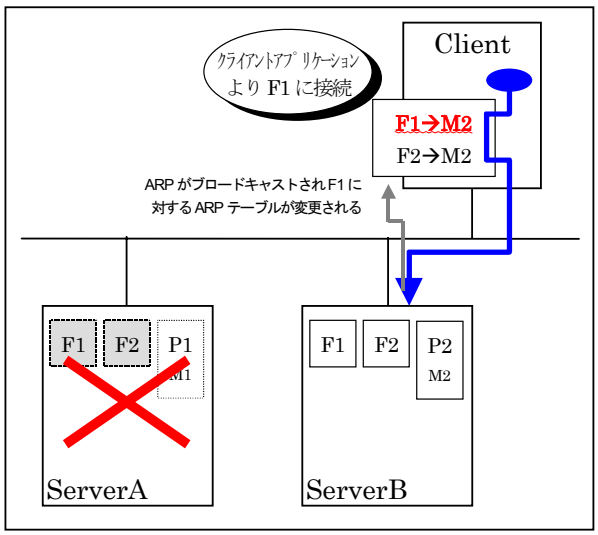


<u>接続形態</u>

*　クライアントからサーバのIPアドレスを指定して接続します。

<u>接続方法</u>

*　接続先にFIPアドレスを指定します。

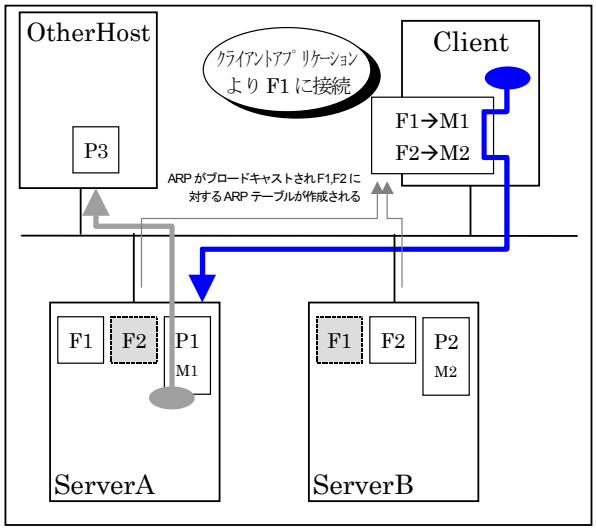


<u>フェイルオーバ時の動作</u>

*　フェイルオーバが発生すると、FIPアドレスに関するクライアントのARPテーブルが変更されます。
　　クライアントは、そのままのFIPアドレスを用いてサーバに再接続することができます。

*　クライアントからサーバへ接続する場合に、FIPアドレスを使用すれば、フェイルオーバの際に接続サーバが変わったことを意識する必要がありません。

7.2.5.2　クライアントからの要求を受けて、他ホストへ接続

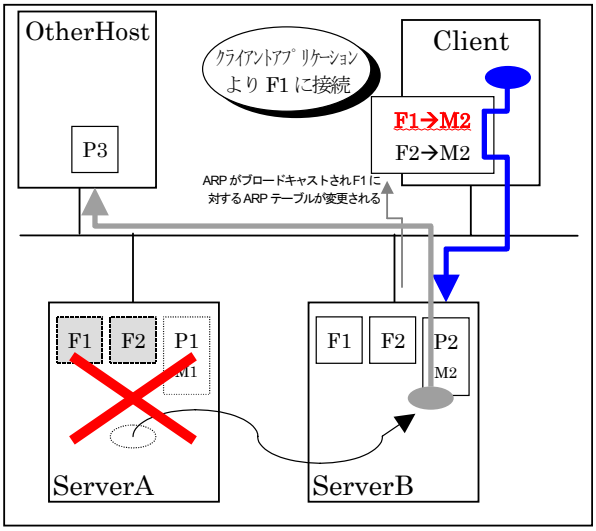


接続形態

* クライアントアプリケーションは、サーバアプリケーションに接続します。サーバアプリケーションはクライアントアプリケーションからの要求を受けて、他ホストに接続し、その結果をクライアントアプリケーションに通知します。

接続方法

* クライアントアプリケーションは、FIPアドレスでサーバアプリケーションに接続します。
* サーバアプリケーションが、クライアントからの要求で他ホストに接続する際は、実IPアドレスが用いられます。
* サーバアプリケーションから接続される他ホストは、どちらのサーバの実IPアドレスからの要求も受け付けるように設定しておきます。
* サーバアプリケーションとクライアントアプリケーションとの接続はFIPアドレスで、サーバアプリケーションと他ホストとの接続は実IPアドレスで行われます。

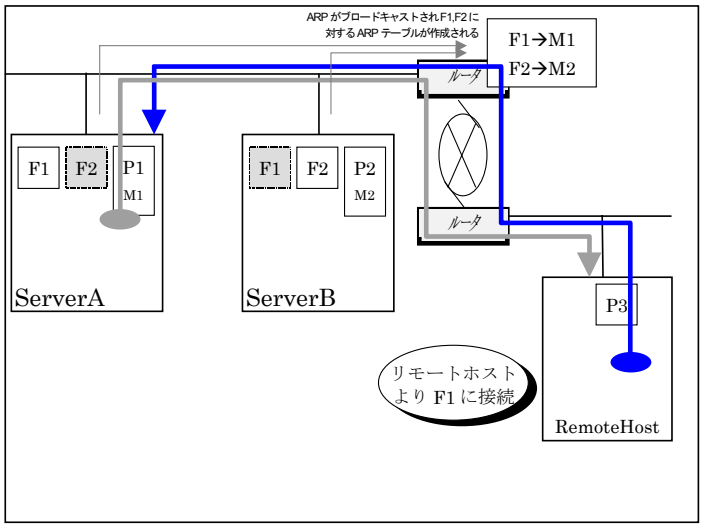


フェイルオーバ時の動作

* フェイルオーバが発生すると、FIPアドレスに関するクライアントのARPテーブルが変更されます。クライアントは、そのままのFIPアドレスを用いてサーバに再接続することができます。フェイルオーバ先のサーバアプリケーションは、クライアントからの要求で他ホストに接続します。

* サーバから他ホストへの接続は、実IPアドレスで接続してください。サーバから他ホストへの接続にFIPアドレスを明示的にbindする必要はありません。

7.2.5.3　リモートネットワーク上の非Windowsホストとの接続

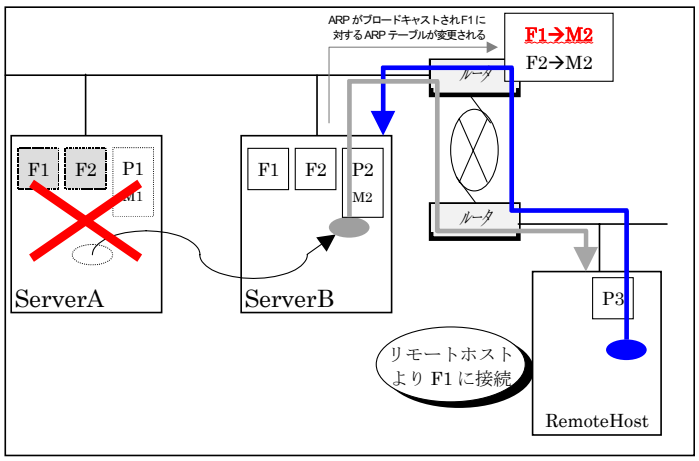


接続形態

* サーバアプリケーションから、リモートネットワーク上のホスト(以下、リモートホスト)に接続します。また、リモートホストからサーバアプリケーションに接続します。

接続方法

* リモートホストは、どちらのサーバの実IPアドレスからの接続要求も受け付けるように設定します。
* サーバアプリケーションからリモートホストへの接続は、実IPアドレスでの接続となります。
* リモートホストからサーバアプリケーションへの接続は、FIPアドレスを指定します。



フェイルオーバ時の動作

* フェイルオーバが発生すると、クラスタサーバ側LANのルータで、FIPアドレスに関するARPテーブルが変更されます。このためリモートホストからは元と同じFIPアドレスを用いて新しいサーバに再接続することができます。また、フェイルオーバ先のサーバからも、リモートホストに再接続できます。

7.3 スクリプト

CLUSTERPROでは、クラスタ対象アプリケーションは、スクリプトによって制御されます。スクリプトはCLUSTERPROによって管理され、起動時、終了時、フェイルオーバ発生時フェイルオーバ，フェイルオーバグループの移動,およびクラスタ復帰の際に実行されます。

shのシェルスクリプトと同じ書式なので、それぞれのアプリケーションの事情にあわせた処理を記述できます。また、CLUSTERPROコマンドをスクリプト内に記述することで、さらに充実した機能を提供しています。

詳しくは、システム構築ガイド「システム設計編(応用)」を参照してください。

7.4 リソース監視

リソース監視は、CLUSTERPROコマンドのarmrspと同等の機能を有しています。

CLUSTERPROのリソースを監視し、異常を検出した場合は、フェイルオーバを発生させるか、またはグループを停止します。

詳しくは、システム構築ガイド「システム設計編(応用)」を参照してください。

8　注意事項

8.1　アクセス許可コマンドに関する注意事項

ミラーディスクアドミニストレータから、切替ミラーディスクに対するアクセス許可コマンドを実行した状態で、HW障害あるいは人為的なシャットダウンが発生した場合は、ミラー不整合となります。

このような状態になった場合には、ミラーの再構築を行ってください。

8.2　ディスクI/Oエラー発生時の注意事項

単体ディスクをミラー対象ディスクとして運用している場合、ディスクI/Oエラーにより切り離されたディスクをミラー再構築できないディスクが一部あります。これは、ディスクのメディアエラーに起因しているため、ディスクの交換が必要となります。又、ディスクの交換無しに、ディスクをSCSIボードのBIOSから物理フォーマット(障害セクタのリアサイン)することにより回避できる場合もありますが、将来のデータ保護の為にディスクの交換をお勧めします。

8.3　ディスクパーティションの変更

一度ミラー対象として指定したディスクは、パーティションの作成/削除といった操作を行うことはできません。もしディスクアドミニストレータにてこの操作を行おうとするとディスクアドミニストレータは操作を完了せずエラーとなります。パーティションの作成/削除/変更を行いたい場合には、ミラー指定を解除しサーバを再起動した後に行ってください。詳細な手順に関しては「CLUSTERPRO システム構築ガイド 運用/保守編」を参照してください。

8.4　ミラーディスクアドミニストレータコマンドの動作制限

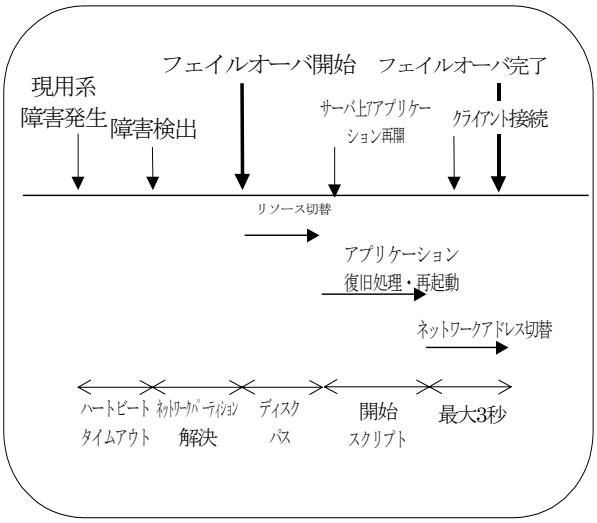
ミラーディスクアドミニストレータコマンドの一部の機能は CLUSTERPRO 動作時には制限が発生します。

サーバ／切り替えミラーディスクの状態と動作不可なミラーディスクアドミニストレータコマンドの機能を下記表に示します。

<table>
<tr><th></th><th>切り替えミラーディスク活性中</th><th>切り替えミラーディスク非活性中</th><th>ダウン後再起動状態</th></tr>
<tr><td>ミラーセット解除</td><td colspan="3">切り替えミラーディスクとして登録されているミラーセットは解除できません。</td></tr>
<tr><td>ミラー構築</td><td colspan="2">切り替えミラーディスクとして動作中のミラーセットはミラーディスクアドミニストレータからミラー構築を実行できません。</td><td rowspan="3">両サーバともが「ダウン後再起動状態」の時のみ可能です。
どちらかのサーバが「正常」状態の場合には行ってはいけません。</td></tr>
<tr><td>強制復帰</td><td colspan="2" rowspan="3">実行できません。</td></tr>
<tr><td>許可</td></tr>
<tr><td>制限</td><td rowspan="2">動作制限はありません</td></tr>
<tr><td>強制許可</td></tr>
</table>

9 付録

9.1 サーバダウン時の切替時間



* ハートビートタイムアウト
プライマリ障害発生後、待機系がその障害を検出するまでの時間で、変更可能です。
出荷時は、3秒×10回の30秒になっています。

* ネットワークパーティション解決
ネットワークパーティション問題を解決するためには、下記1，2のいずれか大きい方の時間が必要です。
１．約1～2回のハートビートタイムアウト時間がかかります。(ハートビートタイムアウトを、既定値である30秒に設定している場合、30秒～60秒が目安です)
２.約1~2回のディスクIO待ち時間がかかります。(既定値である5秒に設定している場合、5秒~10秒が目安です)

* リソース切替（時間は、目安です）
フェイルオーバを行うリソースが複数ある場合は、目安時間×リソース数で計算して下さい。
+ ディスクパス切替(切替ミラーディスク)
最短の場合約1秒で切替をします。ただしファイルシステムの容量によりmountに必要な時間が異なります。また、fsckの実行時間は除きます。
+ フローティングIPアドレス切替
約6秒で切替えます。
+ リソース監視
約1秒で切り替えます。

* 開始スクリプト実行時間
アプリケーションの起動時間、データベースのロールバック時間などが含まれます。ロールバック時間は、チェックポイントインターバルの調整で、ある程度予測可能です。詳しくは、各データベースのドキュメントを参照してください。

* 開始スクリプト実行時間
アプリケーション/サービスの起動時間、データベースのロールバック時間などが含まれます。ロールバック時間は、チェックポイントインターバルの調整で、ある程度予測可能です。詳しくは、各データベースのドキュメントを参照してください。