2A-5

BladeSymphony ファームウェアの開発（3）
Virtage LPAR マイグレーション(コンカレントメンテナンス)の開発

八田ゆかり[†1]　　上野 仁[†2]

(株)日立製作所 IT プラットフォーム事業本部[†1]　同 IT プラットフォーム事業本部[†2]

1 はじめに

OS の稼動やアプリケーションサービスを停止させずに仮想計算機を別のサーバに移動するライブマイグレーション技術が有用とされている。日立サーバ論理分割機構 Virtage は、LPAR 方式でゲスト OS 無停止の論理サーバ移動を実現した。

Virtage による本論理サーバ移動機能は、主要な用途が装置のハードウェア保守であるため、コンカレントメンテナンスと呼んでいる。本機能の方式は高セキュリティが保てる反面、I/O 制御の観点から実現が難しい。本論文ではこの制御方法と運用方法について報告する。

2 コンカレントメンテナンス

コンカレントメンテナンス制御を図 1に示す。移動元 OS のメモリ内容を OS 動作中に移動先物理サーバにコピーし(#6,#7)、OS を一時停止(#8)する間に残メモリイメージをコピーし(#9,#10)移動先論理サーバを起動する(#11,#12)手順が CPU 側の基本制御である。

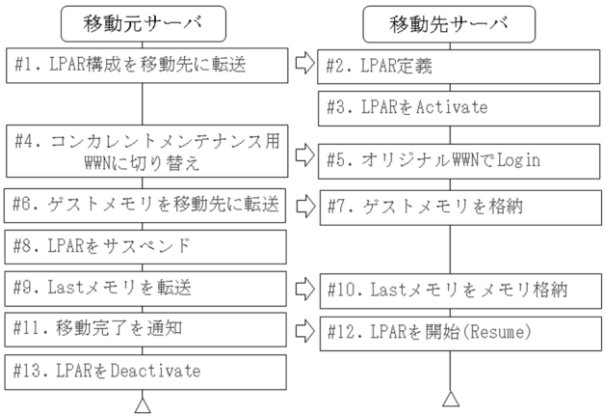

図 1　コンカレントメンテナンス制御

OS の停止(#8)から再開(#12)までの間に、移動元で書換えられたメモリの内容をすべて移動先に反映し、再開時に移動先物理サーバへの動作に切替えるためには、次のような制御が必要となる。

(1) ディスク接続パスの切替え制御

ディスク接続は FC(fibre channel)接続を前提としているため、移動先物理サーバで移動先 LPAR を再開する時(#12)には移動先 HBA で FC-SW の使用ポートに対する Login 処理が完了している必要がある。この処理は数分以上を要する場合があり、これを待っていては OS 無停止を保証することができない。

この時間制約を解決するため、早いタイミングで移動用の FC 接続を並行して起動する方式を開発した。並行起動する FC 接続にはコンカレントメンテナンス用に別の WWN を使用することとした。移動元 LPAR ではコンカレントメンテナンス用の WWN で Login し、オリジナルの WWN は Logout して(#4)、移動先でオリジナルの WWN の Login 状態を確立させる(#5)。

(2) 帯域制御

コンカレントメンテナンスにおいては、移動元の LPAR(ゲスト OS)のメモリの内容を移動先の LPAR のメモリに反映する必要がある(#7)。ところが、転送中に移動元で書換えられる速度が、移動元から移動先にメモリデータを転送する(#6)転送路の速度より速いと、移動元の LPAR のメモリの内容を移動先に反映させる処理が追いつかない。この問題を解決するため、以下の 2 つ方法を用いた。

(i)割り込み報告遅延を利用した帯域制御

I/O 装置側からのメモリ書換えがハードウェアの DMA 転送機能を利用する場合、DMA 転送の帯域を制御する必要がある。物理的な I/O 装置の帯域変更は困難であるため、割り込み制御を利用した実質的な帯域制御を行なう。ここで、

Development of BladeSymphony Firmware (3), Development of Virtage LPAR Migration(Concurrent Maintenance)

[†1] Yukari Hatta, IT Platform Division Group, Hitachi, Ltd.
[†2] Hitoshi Ueno, IT Platform Division Group, Hitachi, Ltd.

実質的な DMA 転送の帯域とは DMA 転送していない時間と DMA 転送している時間を足し合わせた合計時間に行われた DMA 転送量をこの合計時間で割った値である。したがって I/O 装置からの DMA 完了割込みの報告を遅らせれば DMA 転送していない時間を増加でき、実質的な DMA 転送帯域を減少させることが可能となる。

(ii)書き込み保護機能を使用した帯域制御

CPU を介したゲストメモリへの書き込みは、書き込み保護機能を利用して帯域を制御する。書き込みが保護されている領域に対して、データの書き込み要求がされると、書き込み保護例外が発生する。この時、データが実際に書き込まれるためには、書き込み保護例外を解除しなくてはならないが、この書き込み保護例外の解除を遅延させることにより、ゲストメモリに対する書き込み帯域を制御する。

3 仮想マシン方式との比較

LPAR 方式のコンカレントメンテナンスは、OS を停止せずに別の物理サーバに移動するという点において、仮想マシン(VM)方式のライブマイグレーションと共通しているが、両者はストレージセキュリティにおいて大きな相違がある。VM 方式では、移動対象となる物理サーバ間で常時ディスクを共有する必要がある。すなわち、SAN ストレージを使用する場合、ディスクの論理ユニット(LU)のアクセス権限をすべての物理サーバに対して許可しておく必要がある(図 2(右))。したがって多数の物理サーバのうち１台で共有 LU を誤操作すると他のすべてのサーバの動作が影響を受けるという問題がある。

LPAR 方式では、LU のアクセス権限を１台の LPAR だけに許可する。LPAR を移動する場合には LPAR と LU 間の接続権限も同時に別物理サーバに移動するので、VM 方式のようなストレージセキュリティの問題が発生しない。(図 2(左))。

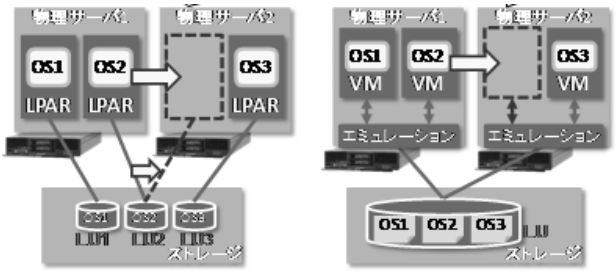

図 2 コンカレントメンテナンス(左)と VM 方式のライブマイグレーション(右)の比較

4 コンカレントメンテナンスの運用

コンカレントメンテナンスでは、ユーザが安心して使用できるよう、コンカレントメンテナンス中に実行状態が確認できる以下のような仕組みを提供している。

(1) 実行ステージの表示

実行ステージを表示することにより、ユーザは LPAR の移動がどの処理まで完了しているかを知ることができる。

(2) メモリ転送性能の表示

移動元 LPAR のメモリの転送速度を表示する。このことにより、ユーザはコンカレントメンテナンス実行時にメモリ転送性能が足りているかを確認することができる。

(3) Dirty メモリ量の表示

コンカレントメンテナンスでは、一度移動元 LPAR(ゲスト OS)に割り当てられた全メモリを転送した後、転送中に書換えられた部分について、メモリ転送を繰り返す。再転送が必要なメモリ容量のことを Dirty メモリ量と呼ぶ。これが一定値以下にならないと LPAR をサスペンドさせた際の Dirty メモリ転送時間が増加し、OS やアプリケーションに影響を与えない短時間内でのメモリやデバイスの状態転送ができない。

メモリ転送中にどれだけメモリの書換え(Dirty メモリ)があるかが分かれば、ユーザはコンカレントメンテナンスの進行状況をより正確に知ることができる。

コンカレントメンテナンス動作に時間を要している場合、上記(1)により進行状況を確認し(2)(3)により原因を判断し対策可能となる。

5 おわりに

LPAR 方式を使用した Virtage でゲスト OS 無停止の論理サーバ移動を実現させた。

また、ユーザが安心して使用できるように、コンカレントメンテナンスの実行状態を確認し、実行できない際にはその原因追及が容易にできる仕組みを開発した。

これにより、ストレージセキュリティを維持しながら、ゲスト OS を停止することなく、安心してサーバの保守ができる機能を実現させた。

参考文献

[1]上野仁：「日立製作所 Virtage によるサーバ仮想化」,仮想化大全 2013, 日経 BP ムック, pp.218-225(2012).