

仮想計算機システム高性能化の一手法

小野一明(富士通株式会社)

1. はじめに

仮想計算機システムは、一台の実計算機から複数の仮想の計算機を作り出し、そのことで複数のOSの動作を可能とするシステムである。この複数のOSが同時に動作することの性質を利用して

- ・システムの移行
- ・複数OSの同時運用

等に使用されました。

しかししながら、従来の仮想計算機システムでは、この性能は必ずしも十分とは言えなかった。これは、このシステム全体を制御する制御プログラム(CPと呼ぶ)のオーバヘッドが相当大きいということが原因であった。

今回、当社が開発した仮想計算機システム(AVM/EF)では、ハードウェアに装備された高速VM機構を利用することにより、オーバヘッドの極めで少ない高性能仮想計算機システムを実現している。そこで、この高性能化のメカニズムと実現された性能について紹介する。

2. 仮想計算機システムの原理と従来の制御方式

仮想計算機システムでは、実計算機システムの資源であるCPU、主記憶、入出力装置を分割あるいはシミュレートという方法を用いて仮想計算機を作り出す。即ち、個々の仮想計算機の要素である、仮想CPU、仮想主記憶、仮想入出力装置は次のように実現される。

- ・仮想CPU — 実CPUを時分割で与える
- ・仮想主記憶 — 実主記憶を分割あるいはページングで与える
- ・仮想入出力装置 — 実入出力装置群を分割あるいは共用使用させる

・CPの制御方式

上記のような仮想計算機を実現するためのCPの基本となる制御方式は従来次のように行われていた。

・仮想CPU制御

実CPUを時分割で仮想計算機に与えるわけであるが、無制限に与えるわけにはいかない。即ち、システム内には複数の仮想計算機が存在しており、その分からシステム全体にかかる制御状態の変更は許さず、実CPUの使用率とおのずと限度がある。

ここで、CPUには動作モードと特権レベルのスーパーバイザモードと一般レベルの問題プログラムモードの2種類が存在する。ここでCPはこの2種類の動作モードの次のようない分けで対処している。

〔スーパーバイザモード — CPが動作

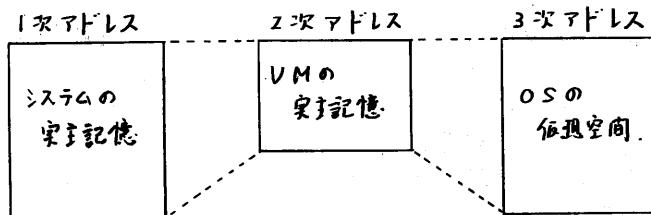
〔問題プログラムモード — VMが動作

このようにVMは問題プログラムモードでの動作しか許さない。この結果、VM上のOSが発行する特権命令(システム全体を制御する命令でスーパーバイザモードでのみ実行可)を直接実行できない。したがって、特権命令はCPによるシミュレートが行われる。

また、VMの動作の結果システムに発生する各種割込みについても、当該VMに關係する割込みだけが発生するよう制御するため、すべての割込みは一端CPUが受け取り個々のVMに割込みをシミュレートすることになる。このため、ハードウェアが割込み等を使用するシステムのハードウェア固定領域(PSA)はCPUの管理域となる。

・アドレッシング制御

仮想計算機の主記憶(仮想主記憶)は実主記憶の一部をえらんで実現されるが、この場合、VM内のOSが使用するアドレスとシステムのアドレス(CPUが最終的に主記憶アドレスのために使用するアドレス)とは一般に一致せず次の關係が存在する。



通常、OSは仮想空間である3次レベルアドレスを使用して動作する。したがって、実CPUがメモリにアクセスするには3次→2次、2次→1次と2度のアドレス変換が必要となる。しかししながら、従来のハードウェアでは1度のアドレス変換しか行えないため、シャドウテーブルと呼ばれる3次レベルアドレスを直接1次レベルアドレスにマッピングするアドレス変換テーブルをCPUが作成しVMを動作させている。この場合、3次レベルと1次レベルの対応關係の変化はシャドウテーブルの再作成につながるため、かなりのCPU介入が必要となる。

・仮想入出力制御

仮想計算機における、実入出力装置を分割あるいは共用せらることにより実現されるが、ここでVMの入出力要求を制御する上で課題となるのは次の二つである。

・チャネルプログラムのアドレッシング

・入出力要求のスケジューリング

まず、チャネルプログラムのアドレッシングについてであるが、前述のように仮想計算機システムでは3レベルのアドレスが存在する。VM上のOSは、3次レベルあるいは2次レベルのアドレスをチャネルプログラムを記述するからチャネルはこのままでは実行できない。したがって、CPは1次レベルアドレスで表現(直)して(チャネルプログラム変換)実行させる。

また、システム内に存在するVMから発信された入出力要求は、実入出力システムで競合することになる。しかししながら、二中間競合は個々のVMには責任がないことであり、CPが介入し個々の入出力要求をスケジュールすることにより競合關係を排除する。

以上述べた事項が仮想計算機システムを実現する上で基本的な制御方法である。

ある。しかしながら、ニホン CP の介入のため基本的な部分でのオーバヘッドの増加によるシステムの性能低下が引き起こさってしまう。

一方、ニホンオーバヘッドを減らすとする努力は、いろいろな形で行われてきた。

- ・ $V = R$ 形態の VM の導入。
- ・ OS と CP とのハンドシェイク
- ・ フームウェア技術の利用
- ・ 専用チャネル形態の導入。

$V = R$ 形態の VM とは、システムの実主記憶のうち低値の連続した領域を VM の主記憶として与え（1 次レベルと 2 次レベルのアドレスが一致）、シャドウテーブルによるアドレッシング制御のオーバヘッド及びチャネルプログラムの変換によるオーバヘッド等を削減しようとするものである。

また、アドレッシング上の問題を OS と CP がハンドシェイクにて解決しようとすると試みがある。ニホン実主記憶のうち任意の連続領域を VM に与え、OS が割り当てられたアドレスを意識することなく、 $V = R$ と同様な効果を狙、たるものである。

フームウェア化についても、VM 上の OS から発信工中の特権命令のシミュレーションをフームウェア技術を用いて高速化しようとするものである。

その他、専用チャネル形態といい、互いチャネル駆下の全入出力装置を特定 VM に専用使用させることにより、仮想入出力制御が不要とした入出力要求のスケジューリング処理を削減し高速化を図っている。

ニホン機能の導入によりある程度のオーバヘッド改善が図られたが、十分とは言えない（依然 10% のオーバヘッドが残る）。また、フームウェア部分のオーバヘッドも無視できなくなる。つまり、制御方式と 12 根本的な改善が又要となる。

今回、当社で開発した仮想計算機システム（AVM/EF）は、ハードウェアに画期的な性能向上を組み、各種機構（高速 VM 機構）を導入することにより高性能なシステムを実現する。以下、高速 VM 機構の概要と CP の制御方法について述べる。

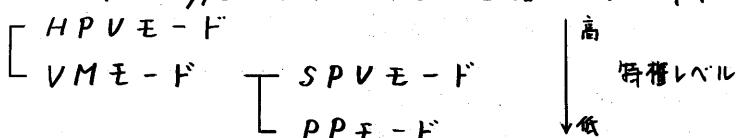
3. 新方式の仮想計算機システム

高速 VM 機構が導入された仮想計算機システムにおいては、動作モード、アドレッシング等に従来のシステムとは違った各種の特徴がある。

- ・ 動作モード — VMモード / HPVモード
- ・ アドレッシング
- ・ 特権命令 / 割込みの実行制御。

3.1. VMモードとHPVモード

従来の計算機システムでは動作モードと 12 問題プログラムモード（PP モード）とスーパーバイザモード（SPV モード）しかなく、だが、新たに VM モードと HPV モード（Hypervisor モード）が追加され次の体系となる。

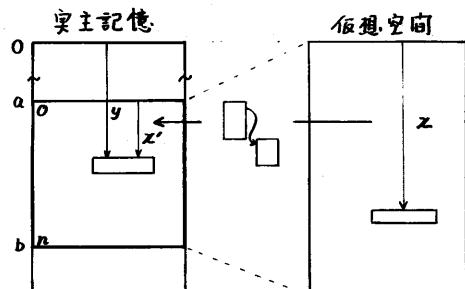


HPVモードとは、システム全体を制御するプログラムであるCPが動作するモードであり、VMモードはVM上のOSが実CPUを与えられた上で動作するモードである。VME-F, HPVモードはPSWのPP/SPVモードを超越したモードで石園の条件と状態は遷移する。たゞ、従来PPとSPVの2種類からなくVMの動作はPPモードに制限されたため実PSWにはVMの指定するPSWをそのままロード不可能である。たゞ(このため従来はCPのみはハードウェアの介入によるPSWの操作が必要だった)。PSWのモードを超越する本モードの導入により実PSWにはVMのPSWをロードしPP/SPVモード任意のモードで動作することが可能となる。たゞ、VME-FはVMがロードした実PSWで全面的にシステムが制御を受けるわけではなく、CPの介入の必要な事象を捕らえるため設けられたPMR(modification register)と呼ばれる3割りレジスタで修飾を受けた実効的なPSWで制御される。通常、CPは入出力割込み、外部割込み、マシンチェック割込みだけは介入の必要があるため常に割込み可となるようPMRを設定している。

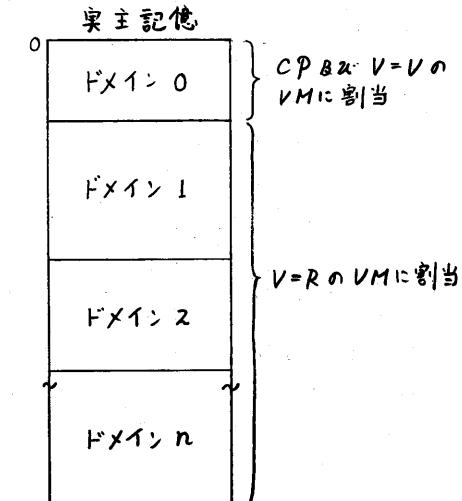
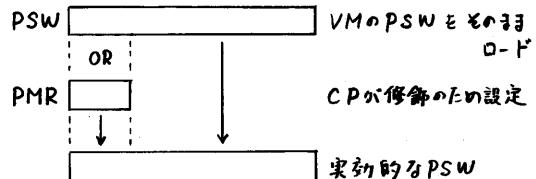
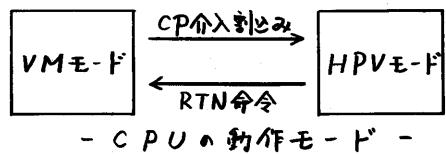
以上のように、新モードの導入及び実効PSWという考え方はVM上のOSから発信される特権命令の直接実行に道を開いていく。

3.2 アドレッシング制御

新しく導入されたアドレッシング制御の特徴を一口で述べると、ハードウェアがVM内でのアドレスをアドレッシング可能となるということである。



VMの主記憶がシステムの主記憶アドレスaから始まる連続領域に割り当てられるときとすると、OSが指定する仮想空間上の任意のアドレスZは、ハードウェアにより自動的に、OSが記述したアドレス変換テーブルを使用してVMの実



アドレスエ'に変換工中、統々当該VMが割り当てられた先頭アドレスを加算され、システム実アドレスが生成工中で主記憶アクセスが行わる。これら自動的なアドレス生成は、CPUとチャネルにおけると行われる。

そこで、まず、当仮想計算機システムではこのアドレッシング制御の特徴を生かすためVMの主記憶はシステムの主記憶上に、図のように連続領域であるドメインによって割り当てられた。

(1) CPUにおけるアドレッシング

VMがCPによりディスパッч(実CPUを与えられたとき)を中心としたアドレッシング制御のため当該ドメインを性格付けた制御情報(ABR/ALR)がハードウェアに対し通知される。

ABR(Address Base Register), ALR(Address Limit Register)は当該ドメインのシステム主記憶上での先頭及び最終アドレスを示す制御レジスタである。

VM内の論理アドレスは制御レジスタ1(CR1)で示されるアドレス変換テーブルを使用して変換工中(当該変換テーブルへのアクセスはABR値が加算工中で行わる)ドメイン実アドレスとなり、プリフィックス変換後ドメイン絶対アドレスとなり、最終的にABR値が加算されたシステム絶対アドレスが生成工中である。VMの実アドレスの場合は、プリフィックス変換後ABR値が加算されたシステム絶対アドレスが生成工中である。

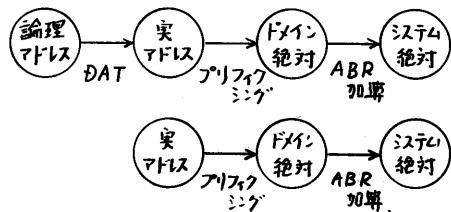
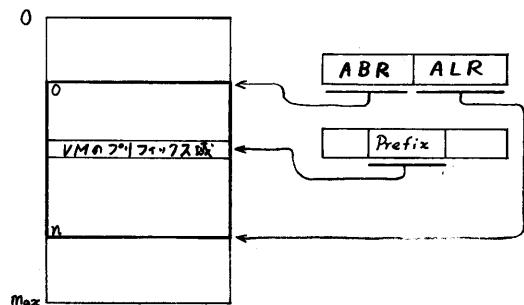
以上の変換過程におけるALRによるドメイン外へのアクセスのチェックが行わることあり、VMの誤動作による他領域への侵害は避けること可能となる、といふ。

ここで述べたアドレス変換がVM動作中常に行わる工事ではない。通常のDATの場合と同様、CPUにはTLBが用意工中おり、ドメイン論理アドレスとシステム絶対アドレスとの対応が既に何工事かはさむを用いて行わるから、高速アドレス生成が可能となる、といふ。

(2) チャネルにおけるアドレッシング

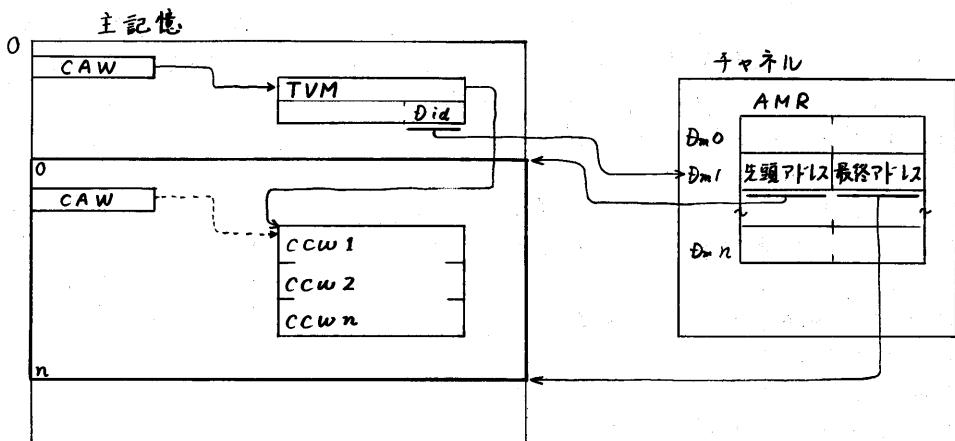
従来の仮想計算機システムでは、チャネルのアドレッシングに合わせるために、VMから発信工中のチャネルプログラムに対して、CPによるアドレス変換や領域チェックが行わる工事だが、高速VM機構の導入により、OSにより記述工中にチャネルプロログラムからチャネルでアクセス可能なアドレスの生成工と領域チェックは、チャネルに下工事自動的に行わるようになる。(これをCHANNEL機能と呼ぶ)

システムには、チャネルの主記憶アクセスのためのアドレス変換レジスタ



(AMR)が用意されており、ここには各ドメインの先頭アドレスと最終アドレスが格納されている。

VMからX出力命令であるSIO(Start I/O)命令が発信されると、CPUは中断を依頼する。OSが記述したチャネルプログラムに、チャネルに対するドメインの開始/終了アドレスを通知するためのチャネルコマンド(TVMコマンド)を附加し、VMに代わってSIO命令を実行する。



チャネルでは、TVMコマンドが現れると以降のチャネルプログラムにおけるD-idを示すアドレス情報を基にCCWアドレス&データアドレスを計算し主記憶アクセスを行う。ここでCPUと同様、ドメイン外へのアクセス侵害チェックが行われる。

3.3 特権命令/割込みの実行制御

当面想定計算機システムでは、CPの介入を絶えずだけなくオーバヘッドを削減しようという観点で、特権命令、割込みも分類し可能な限り直接実行させる。

(1) 特権命令の分類と制御

レベル	特権命令	制御方法
0	IPK, SPT, SPKA, STPT, SPSE	VMで直接実行される。
1	ISK, LRA, IPTE, RRB, SSK, LCTL1,1	ドメインを割り当てられたVM(V=R)は直接実行。 その他はCPに割り出される。
2	LPSW, SSM, STNSM, STOSM, LPSE	VMで直接実行。 ただし命令実行後 PSW変更 割込みが発生する場合がある。
3	入出力命令など上記以外の特権命令	常にCPに割り出され、シミュレートが行われる。

この特権命令の直接実行は、当面想定計算機システムにおけるアドレッサン機構成PSWの制御機構等に負うところが多い。

レベル0に分類される特權命令は、CPU91マあるいはPSWのファイルドを参照更新する命令であるが、これらはVMの自由にまかされた資源でありVMモードでの直接実行が許さない。レベル1に分類される命令は主記憶キー、アドレスシングル部に関連しており、ドメインを割り当たしたVM($V=R$ のVM)にリリースは直接実行が許される。また、レベル2に分類されるものはPSWを変更する命令であるPSWはVMに許さずより直接実行を止め、実行の結果WAITのPSWがロードされた等の場合、新設のPSW変更割込みが発生する可能性がある。最後にレベル3に分類されるものは、割御上CPの介入が必要、あるいは発信頻度が低いなどのもので、CPにより常にインターフェースやシミュレーターである。

$V=R$ のVMの場合、CPは通常レベル0,1,2の特權命令は直接実行を止めようハードウェアを設定しディスパッチする。ただし、CPの介入は、レベル2特權命令でPSW変更割込みが発生した場合とレベル3特權命令のシミュレーションの場合に限定される。また、この直接実行という方式は、従来の方式のようにマイクロプログラムによるシミュレーション型でないため、実計算機と遜色ない高速性が得られる。

(2). 割込みの分類と割御

当高速VM機構では、プログラム割込みの一種としてPSW変更割込みが新設されるとともに、従来の割込みについても割込み先(VMに直接割り込まれるHPVに割り込まれる)を割御可能となる、といふ。

PSW変更割込みについては、レベル2特權命令の説明で述べたが、実PSWにはVMの指定したPSWが格納されよりVMの割御に許さずよりVMのWAITのPSWをロードしようとしたら、あるいは、VMに反映すべき割込みを保留中にVMのPSW操作命令の結果割込み可能となりた場合等、特殊なケースとしてCP介入が必要な場合に発生する割込みである。(プログラム割込みの一種)

当システムにおける割込みは、次のように分類され割御される。

従来の方式に比べプログラム割込みに対するCPの介入の頻度は極度に少なくなる、といふ。さらに、プログラム割込み、SVC割込みはVMの動作と同期した割込み事象であり、CPによるシミュレーションの必要を一部を除き、ハードウェアはVMのPSAを使用して直接割込みを発生させた。

また、入出力割込み、外部割込みは、プログラムの動作とは非同期でありCPの介入が必要なため、割込み先は常にCPであると同時に、PMRを通常に割込み可能として割御される。CPに割り当てられた事象は関係するVMの割込みマスクがチップセット反映処理が行われるが、VMの割込みマスクが閉じてると割込み情報を保留するためのレジスター(IPIR)に設定されたVMからのPSW操作命令の実行により発生するPSW変更割込みを契機に反映処理が行われる。なお、VMが専用チャネル形態で入出力装置を使用する場合には、PM

割込み種別	割込み先
プログラム割込み	特權命令例外 VM/CP
	PSW変更 CP
	上記以外 VM
SVC割込み	VM
入出力割込み	CP
外部割込み	CP
マシンチェック割込み	CP
リストート割込み	CP

Rの入出力割込みマスクドリバはVMの割込みマスクと一致するよう設定され、当該チャネルからの割込み発生時割込み保留という状態が引き起こさないよう制御されるため、ペンドライング处理のオーバヘッドの大幅削減が図られる。

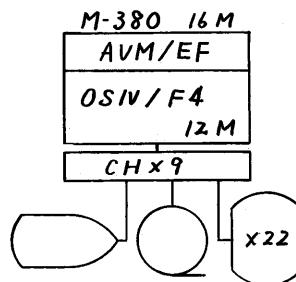
以上が高速VM機構の特徴と新仮想計算機システムへの制御方法であり、従来方式に比べ大幅な性能改善が期待できる。

4. 性能

当仮想計算機システムへのオーバヘッド実測結果を示す。

(1). 測定条件

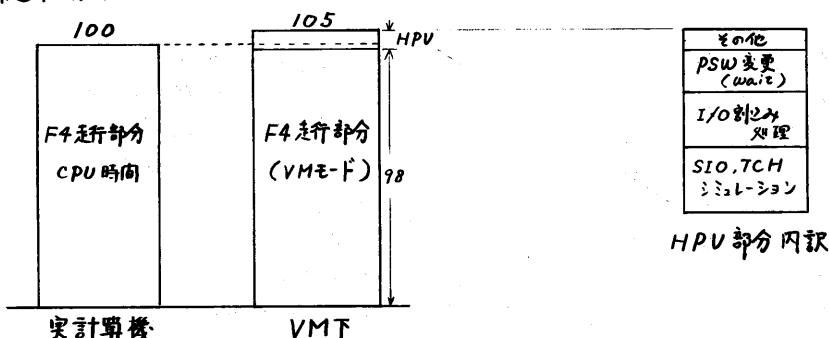
測定マシン	M-380
メモリサイズ	16M (V=R:12M)
チャネル	9 (専用チャネル6)
DASD	22本
使用OS	OSIV/F4
ワーカロード	SBS (BATCH)
イニシエータ数	20



測定は上記のように、高速VM機構付のM-380で当社の代表的OSであるOSIV/F4を使用してBATCH環境を行った。なお、SBSは標準的なBATCH系ワーカロードとて当社で性能測定に使用しないものである。

また、オーバヘッドとは実計算機で使用するCPU時間に対するVM環境でのCPU時間増加の割合であり、このためVM環境と等価な実計算機での測定を行った。測定項目は、実計算機環境でのCPU時間、及びVM環境でのHPV×VMモードのCPU時間等である。

(2). 測定結果



図のよう実計算機のCPU時間に対しVM下では5%のCPU時間増加が見られない。ここで、VMモードのCPU時間に注目すると実計算機の場合に比べ逆に2%の減少が見られる。これはVMモードでの特權命令直接実行の効果を如実に表わしている。次に、HPVモード即ちCPU走行部分を見ると大半が入出力命令及び入出力割込み処理による部分で、I/O高頻度環境を考えるとこの部分をどう扱うかが今後の課題である。