Virtual Machine を活用した大規模ファイルシステム

丸山 伸[1]　北村 俊明[2]　藤井 康雄[1]
[1] 京都大学学術情報メディアセンター
[2] 広島市立大学情報科学部情報工学科

概要:

京都大学学術情報メディアセンターにおいて本年 2 月より稼働を始めたシステムにおいては、利用者向け大規模ファイルシステムを構築するに当たり、Virtual Machine をシステムの中核として活用したシステムとなっている。

大規模ファイルサーバーにおいては NFS プロトコルによってファイルシステムサービスを提供しているが、利用者端末の Virtual Machine 上で稼働するソフトウェアにより NFS プロトコルから SMB プロトコルへと変換を行い、利用者端末の OS は SMB プロトコルによりファイルサーバーを利用している。このようなプロトコル変換はシステムに負荷をかける処理であるため、これまで大規模なシステム上で運用を行うことは難しかった。

今回構築されたシステムにおいてはこのプロトコル変換を 1000 台を超える分散された端末上で行うように工夫したことで、初めて大規模な運用が実用的になった。

Application of Virtual Machine to the Large-Scale File System

Shin Maruyama[1]　Toshiyuki Kitamura[2]　Yasuo Fujii[1]
[1] Academic Center for Computing and Media Studies, Kyoto University
[2] Faculty of Information Science, Hiroshima City University

Abstract

The new system, which started service from this February at Academic Center for Computing and Media Studies, employs Virtual Machines as the central unit of its Large-Scale File System for users of its own.

The Large-Scale File System provides its service using NFS protocol, and convert it to SMB protocol using software which runs on virtual machines on users' terminals, and the OS (Operating System) on users' terminals make accesses to file server with SMB protocol. This kind of protocol conversion requires CPU time, and it was hard to construct this kind of system with large-scale.

We made it possible to do this kind of protocol conversion with an idea of running the protocol converting software on distributed more than 1000 terminals.

1.教育用計算機システムにおけるファイルサービス

京都大学総合情報メディアセンター (現在は学術情報メディアセンター) は 2002 年 2 月に教育用計算機システムの機器更新を行った。利用者用端末（以下クライアント）だけではなく、大容量ファイルサーバーを含むサーバー群、そしてネットワーク機器などが更新された。

教育用計算機システムの特徴としては、利用者数が著しく多く、個々の利用者の利用目的が非常に多岐に渡ることが挙げられる。このようなシステムにおいてファイルサービスを提供する際には非常に大きなディスク容量を用意することが必要となる。今回の構築において大容量ファイルサーバーには 3 万人の利用者に対して一人当たり 100MB のディスクを提供することを想定して、RAID5 構成でかつ 2 重化された状態で物理容量 3TB の SAN を利用したディスクが要求されることとなった。その結果、以下のようなシステムが大容量ファイルサーバーとして導入されることとなった。(Fig 1)

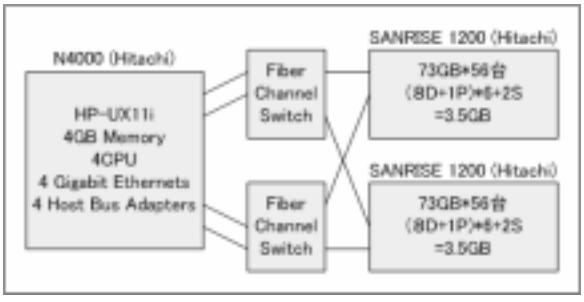

Fig 1: 大容量ファイルサーバーの構成

教育用計算機システムのファイルサービスにおいては、このような大容量ファイルサーバーに対し、同時に 1000 台を越える端末からほぼ同時にアクセスされうる状況であることを考慮して設計を行わなければならない。また、利用者が毎年入れ替わることから、運用を数年に渡り行う中でトータルのアカウント数は数万に上ることが予想される。そこで、このような教育用計算機特有の事情に配慮しつつ、管理を行いやすい大規模ファイルシステムを構築した。

2.端末の概要と Virtual Machine の利用

近年教育用計算機システムに対し Virtual Machine を用いて UNIX 環境を構築する例が増えている。これは限られた予算と端末台数の中で、Windows 環境を利用したいという利用者からの要求にこたえると同時に、コンピュータの原理を教育するためには UNIX 環境を利用することが必要であるという教師側の声の双方に答えるための有効な手法である。

この度の構築においては、クライアントには Microsoft 社製 Windows2000 (以下 Windows とする)を利用することが求められた。そこに VMware 社製の Virtual Machine である VMware3.0[1)]を用いて UNIX 環境（今回は Vine Linux 2.1.5 を用いた）を構築した。

Virtual Machine は起動時にメモリー割り当てを要求する。今回はクライアントの総物理メモリーが 256MB であることを考慮して、Virtual Machine と Windows とにそれぞれ 128MB を割り当てることにした。

なお、クライアントの仕様は以下のようなものとなっている。(Table 1)

CPU:	Intel Pentium3　1GHz
Memory	256MB
Harddisk	20GB IDE
Network	100Mbps Ethernet

Table 1: クライアントの構成

3.ファイル転送プロトコルの問題

ファイルサービスを行うにあたり、サーバーとクライアントとを接続する際にどのようなプロトコルを用いるかを決定する必要がある。今日大規模なファイルシステムを構築する際には NFS プロトコルと SMB プロトコルとが広く用いられている。

	NFS	SMB
接続の管理	Stateless	State full
認証	IP address	ID/Password
Security	△	○
UNIX との親和性	◎	○
Windows との親和性	X (但し SFU を用いれば△)	◎
大規模運用での実績	◎	△

Table 2: ファイル転送プロトコルの特徴

NFS プロトコルは Stateless であることから、サーバーが管理・把握するべき情報が少ない。また UNIX において古くから利用されているものであることから、UNIX との親和性が非常によい。ファイルサーバーはユーザーのパスワードなどの情報を特に管理する必要はなく、クライアントから申告された UID をそのまま利用することになる。セキュリティの観点からは、クライアントの認証は IP アドレスのみによって行われるため、Virtual Machine を用いることが出来る環境において利用者による IP アドレスの偽装を防ぎセキュリティを保つことは難しい。

それに対し、SMB プロトコルは Stateful なプロトコルであり、Windows との親和性がよい。ファイルサーバーは利用者のパスワード情報などを所持ないしは参照する必要がある。クライアントの認証はこのパスワードを用いて行われるため、この点においては NFS プロトコルよりも安全である。

また、管理者にとってのファイルサーバーの管理のしやすさという観点からは NFS プロトコルを利用する方がよい。しかしながらクライアント端末に Windows を選択することになったため、クライアント側では SMB プロトコルを用いるのが適切であると判断された。そのため、サーバーないしはクライアントのいずれかの段階においてプロトコル変換を行う必要性が生じた。

4.Virtual Machine を活用したプロトコル変換

UNIX ファイルシステムないしは NFS プロトコルを、Windows がファイルシステムとして利用できる状態ないしは SMB プロトコルに変換するためのいくつかの方法が検討された。(Fig 2)

方式 1 にある SFU[2] (Microsoft 社製 Service for Unix) は Windows クライアントが NFS プロトコルを利用できるようにするための追加ソフトである。この SFU の利用が検討されたが、Windows 側のユーザー名と UNIX 側の UID との対応付けを行う仕掛けが複雑であり、かつ NIS を利用しないといけないという点とユーザー数が数万となる環境において利用することが難しいという観点から見送られることとなった。

次に Samba[3] という UNIX 上で動作するソフトウェアを用いて、UNIX が利用しているファイルシステムを SMB プロトコルによりアクセスさせることが出来るようにすることが検討された。しかしながらこのプロトコル変換をファイルサーバー上で行うように設計すると 1000 を超えるクライアントが 1 つのサーバーにアクセスすることになる。現在の Samba の

実装について検討をした結果、これを安定して動作させることは非常に難しいと思われたため方式 2 も見送ることにした。

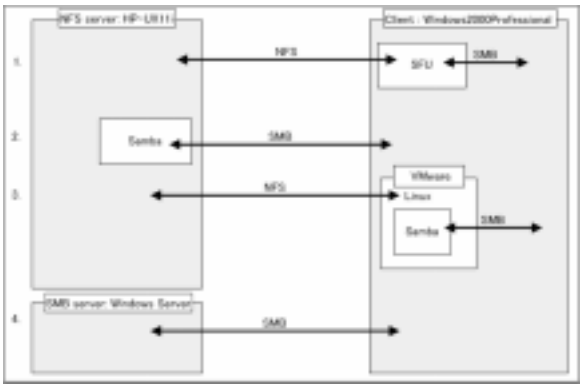

Fig 2: プロトコル変換を行う 4 つの方式

方式 4 は必然的に Windows サーバーを用いることになるが、Virtual Machine 内の UNIX に作成されるアカウントと Windows を利用する際のアカウントの両者の認証を統合することが難しいという観点などから採用を見送った。

そこで今回はクライアント端末で動作している Virtual Machine を活用し、プロトコル変換を利用者端末の上で Samba を用いて行うことにした。

5. システム構築におけるいくつかの工夫

前回のシステムにおいてはさまざまな問題が発生してしまったが、今回の構築においてはそれらに対する対策が行われている。

5.1.利用者端末の文字セット問題

教育用計算機システムにおいては利用者毎にさまざまな文字セットを利用することになる。前回のシステムにおいては、クライアント・サーバーともに Windows であったにもかかわらず、不適切な文字セットを利用されたことが原因でサーバーから参照できないファイルが作成されてしまった。今回もクライアント端末において複数の文字セットが利用されることが想定されたため、この問題を避けるためにクライアントからファイルサーバーに要求を出す際に、ファイル名に対して HEX エンコードを行うように設定した。HEX エンコードとは「“A” を “:41” のように 16 進数で表現する」のと同様の変換をアルファベット以外の文字に対して行うことである。

5.2.Virtual Machine に対する細工

Virtual Machine が利用者自身によって勝手に停止させられたりすることがないように工夫することが必要である。今回の構築においては通常の利用者は Virtual Machine の存在を視覚的に認識することがないように工夫をした。(Fig3) さらには Virtual Machine を停止しようとした際に警告を促すプログラム(Fig 4: VMwrapper)を開発した。

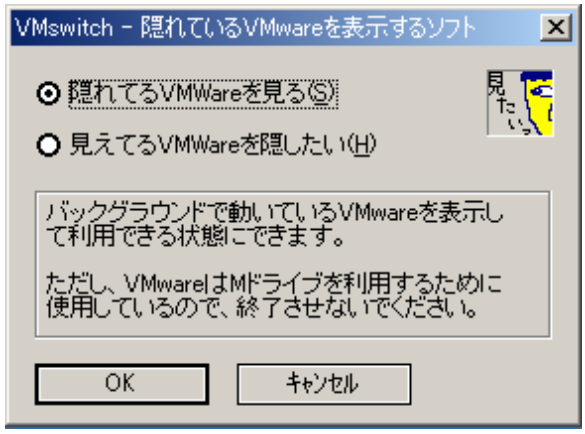

Fig3: VMware を画面に表示するかどうかを選択するためのツール（VMswitch）

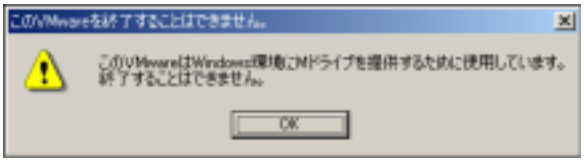

Fig 4: 利用者が VMware を停止しないように警告をする、VMwrapper が出すメッセージ

5.3.Virtual Machine の起動を高速化する工夫

クライアント端末の電源投入時に毎回 Virtual Machine 内の UNIX の起動処理を行うのは非常に時間がかかり効率が悪い。そのため、UNIX が起動した状態で Virtual Machine を Suspend しておくようにした。また、利用者がログオフした際に Suspend イメージを初期状態に復元するようにしている。

5.4.アカウントと UserID のマッピング

前回のシステムにおいて、利用者数に依存した問題がいくつか発生した。一つは「NIS を利用した一部の環境において、利用者数が 3 万を超えたころにデータベースの構築が出来なくなる」というものであった。また、「一部の環境において、UserID が 16bit であると想定されている部分が存在し、UserID が 65536 (= 2^16) である時、それは UserID が 0 であるものと同等に扱われてしまう」というものであった。多くの環境においてこの種の問題は発生しないように対策が行われてきているが、今後も同種の問題が発生しないとも限らない。

そのため、今回の構築においてはアカウント名と UserID との対応を固定しないように細工した。クライアントがファイルサーバーを mount する際には、事前にファイルサーバーで認証を行った上で UserID と mount point の情報を取得するようにした。同時に認証を行った端末に対してのみ mount point を公開することで、クライアントは利用者のホームディレクトリ以外を mount することが出来なくなっている。この工夫により NFS が持つ IP アドレスの偽装に関するセキュリティ問題を回避している。

5.5.ファイルのアクセス権の問題

Windows が標準的に利用するファイルシステムにおいては、作成されたファイルは管理者の目からも保護されるべきであるという観点などから、管理者にすら読み書きや削除が出来ないファイルを作成することが可能である。前回はこのような設定をされたファイルを管理者ですら通常の方法では削除することが出来なくなり混乱を来した。今回ファイルサーバーは UNIX の標準的なファイルシステムを利用しているため、この種の問題は発生しない。

6. パフォーマンス

今回構築された Virtual Machine をプロトコル変換に用いたファイルシステムに対して、パフォーマンス計測を行った。ディスク速度の計測にあたってはフリーソフトで定評のある HDbench[4] を用いた。比較のために、クライアント端末に直接接続されたドライブのデータも計測した。

ローカルドライブと比較して、読み取りは 44%、書き込みは 15%程度のパフォーマンスであることが計測された。読み取り速度は 100Mbps に近づいており(ピーク時で約 94Mbps となっている)、Ethernet の転送速度がほぼ限界に達しているものと推測される。また 100MByte を超えるファイルを読み込む際に著しい速度低下が見られるが、これは VMware に割り当てられたメモリ容量(128Mbyte)に制限されているものと考えられる。

平均 / Read (KB/sec)	サイズ (MB)							
	1	10	20	40	50	100	200	500
M ドライブ(VMware 経由)	10137	11,878	11,675	11,783	11,777	5,511	4,474	4,490
C ドライブ(ローカル)	24975	28,287	24,914	27,619	25,805	24,692	27,334	25,246

Table 3: ファイルサイズ別、Read 速度計測の結果

平均 / Write (KB/sec)	サイズ (MB)							
	1	10	20	40	50	100	200	500
M ドライブ(VMware 経由)	3778	3,452	3,424	4,052	4,066	3,445	3,718	4,070
C ドライブ(ローカル)	11251	26,188	24,008	25,711	22,514	25,183	26,728	25,246

Table 4: ファイルサイズ別、Write 速度計測の結果

平均 / Copy (KB/sec)	サイズ (MB)							
	1	10	20	40	50	100	200	500
M ドライブ(VMware 経由)	1540	2,260	1,651	1,708	1,711	2,083	1,699	1,711
C ドライブ(ローカル)	2590	2,642	2,947	2,652	2,902	2,698	2,454	2,687

Table 5: ファイルサイズ別、Copy 速度計測の結果

しかしながら今回のシステム設計においては、利用者に提供される一人当たりのファイルサーバーの容量は 100MB 程度とされているため、このパフォーマンスは運用上概ね問題ないと言える。

7. 考察

本来は UNIX 教育を目的として導入されている Virtual Machine を用いてプロトコル変換を行う仕組みを構築した。この種のプロトコル変換に Virtual Machine を活用するためには、Virtual Machine が常時起動していることが要求される。そのためには十分な物理メモリーを用意することが必要であることがわかった。メモリーの量の影響は特に大きなファイルを連続して読み込む際に顕著となる。

Virtual Machine を運用目的に用いる際に、利用者がその存在を意識しなくても済むようにした事で、利用者に混乱をきたすことなく運用出来ているものと思われる。パフォーマンスにおいても運用に支障の出るような影響は出ていない。

Virtual Machine を活用してプロトコル変換を行うことで、サーバー側にもクライアント側にも扱いやすい形で大規模ファイルシステムが提供できたと言えよう。

Reference:

1) VMware 3.0 http://www.vmware.com
2) Service for UNIX http://www.microsoft.com/catalog/display.asp?site=759&subid=22&pg=2
3 4 Samba http://www.samba.org/
4) HDbench http://www.vector.co.jp/soft/win95/hardware/se032899.html