5年間使えるシステム作り

丸山 伸・北村 俊明・藤井 康雄・中村 順一

京都大学総合情報メディアセンター

要約

京都大学総合情報メディアセンターでは2002年２月１日運用開始の予定で次期システムを構築している。この構築において最も重視されたことは「5年間運用可能なシステムを作る」ことであった。そのため、「負荷の増大に対処しやすいものであること」「機器が陳腐化した場合においても必要な機能は維持できるような工夫がされていること」「障害が発生した場合においても対処を行いやすいシステムであること」を設計の基本に据えた。

様々な検討の結果、「利用者端末にはWindows2000を採用するがWindows 2000 Serverは利用しない」「Windowsの認証に依存せず、独自の認証方式を実装」「VMwareによる仮想マシンを利用してLinux環境を提供」「学内網に依存することなく独自のネットワークを構築」「スケーラビリティーのあるサーバー構成」といった大きな特徴を持つシステムとなった。

Making the System which can be used for Five Years

MARUYAMA SHIN, KITAMURA TOSHIAKI, FUJII YASUO, NAKAMURA JUN-ICHI

Kyoto University, Center for Information and Multimedia Studies

Abstract:

In the Center for Information and Multimedia Studies, Kyoto University, the next computer system for education is being built by the schedule of starting service on Feb 1st, 2002. Having been most importantly thought in this construction, was that "the system can be utilized for longer than 5 years". Thus, we set following three plans as the foundation of system design, “The system should be easily cope with the increase of system load.", "Even when the machines become old-fasioned, minimal functionality should be served.", and "The system should be easy to be maintained in the case of trouble."

As the result of various examinations, it becomes a system with characteristic features, "Adopting Windows2000 Professional for client machines, but no Windows2000 Server is used", "We don't use Windows based authentication, but use our originally implemented method.", "Serving Linux enviroment in virtual machine for users using VMware.", " Construct our own network system, which does not depend on campus network.", "Server computers are arranged to have scalability."

はじめに

計算機システムには寿命がある。この寿命の要因としては、「利用者の拡大、利用機会の増加に伴う負荷の増大」、「システムの利用が進むにつれて、ディスク消費量やログの量が増大すること」といった様々な理由が挙げられる。

この度京都大学総合情報メディアセンター(以下、メディアセンターと称する)において教育用計算機システムを更新するにあたり、「5年間使えるシステムを構築する」という大きな目標を掲げた。5年という時間は近年の技術革新の速度に対しては長すぎる時間である。5年前と現在とを比べてみるとCPUもOSも2世代の入れ替わりが行われている。この度構築されるシステムを5年間使うことを考えた場合、こういったCPUやOSの世代交代を乗り越えなければならないことは容易に予想される。旧タイプのシステムを使い続けることで様々な障害が発生することも予想されるが、Webのブラウズやメールの送受信、ワープロや表計算ソフトを利用しての資料作成といった基本的な作業に対して支障をきたすような障害は可能な限り避けられるように設計されていなければならない。

本研究においてはこの「5年間使う」という大きな目標を掲げるにあたり、どのような点に注意を払いつつシステム設計を行ったかについて述べる。また、現在構築中のシステムではあるが、どのような実装を行っているかについても要点をまとめる。

設計方針

- 特定のベンダーの非公開情報に依存したシステムを作ることは非常に危険である。そのため今回のシステム構築においては可能な限り汎用的なプロトコルやプログラム言語を用いている。
- 負荷のかからないシステムを作ることは無理である。負荷のかかりそうな部分で特定のベンダーへの依存をなくすことで、万一の障害時の解析を容易にすることが重要である。サーバーにおける様々な処理を、可能な限りクライアントに移行する。
- システム設計において冒険しても良い場所と、安定させなければならない場所とを明確に区別する。冒険をする際には非常時用の代替手段を用意する。

システム設計段階における検討

今回の構築において大きな問題が発生すると想定したのは「メール」「ファイルサーバー」「ネットワーク」「大規模環境における認証」の4点であった。

「メール」

電子メールに対する需要は日々増大している。今回の構築においてメールサーバーについて検討した内容は負荷対策とセキュリティ対策が中心であった。

負荷対策に関しては「授業終了時に携帯電話からメールを読み書きする際に発生するアクセスに耐えられるように」「学内からの事務連絡が全学生宛に一斉に配信されても遅滞なく動作するように」と設計された。

携帯電話からのアクセスを想定すると、メールサーバーにアクセスするクライアント数は、端末数ではなく利用者数に近いものとなる。授業の合間である10分間の休憩時間に6000人が携帯電話を利用してメールチェックをすると仮定した際、メールサーバーに対して平均して秒間10アクセスが行われることになる。アクセスのばらつきを考慮すると、秒間100アクセスとなる可能性も十分に想定される。この規模のアクセスに対して耐えうる、ないしは拡張し得るサーバーを設計しておく必要があった。

事務連絡が引き起こす負荷は最悪のケースでは1万人を超す学生に対して、メールが学内から一斉に配信されるような状況を想定した。1通の同一内容のメールを1万人に送信する場合であったとしても各受取人に対して1つのコネクションを張るようなMTA (Mail Transfer Agent) が中間に挟まれた場合、特定のサーバーから1万コネクションが一斉に要求されることとなる。この状況下においては1万通のメールを受信し終わるまで、これ以外のメールを事実上受信できなくなってしまう。この種の負荷に対処するため同一のIPアドレスからメールサーバーへの同時コネクション数を制限することが出来るようなネットワーク機器を導入する必要性があった。

メールサーバーのセキュリティ対策に関しては「学外からもメールの送受信が問題なく行えるように」設計された。大学の構成員は国内外を問わず学外にいる機会が多く、学外からもメールの送受信を行いたいという要求は強い。また学生の下宿などからの利用に対して、これまではPPPによるダイアルアップを利用して大学に直接接続した場合のみ送受信が可能となる運用を行ってきたが、xDSLなどの回線を利用してのアクセスが増加している現状には相容れないものとなっていた。そのため、不正中継などのセキュリティー問題に配慮しつつ学外からのメールサービスの利用がスムーズに行えるように、ブラウザが用いられる環境であればメールの送受信が出来るWebメールと、サーバーへの接続前に認証を行うためにVPNの導入を行う必要性があった。

「ファイルサーバー」

近年、一般の利用者が扱うファイルサイズが増大している。それとともに市販されているストレージの大容量化と低価格化が進行している。しかしながら大規模で高信頼性を持つファイルサーバーの導入コストは依然として大きい。このような状況下で利用者を十分に満足させる容量のファイルサーバーを導入することは現実的ではなかった。それに代わる形で、利用者が何らかの外部記憶媒体を持ち込み利用できるように配慮することにした。

次にファイルサービスを提供する際に利用するプロトコルとして、NFSプロトコルとSMBプロトコルのどちらを使うかという問題を検討した。大規模運用を行う際にはステートレスであるNFSプロトコルに優位点がある。また、利用者認証をSIDと呼ばれる識別子により行うSMBプロトコルは、クライアント側にも利用者登録を必要とし、Windowsサーバーを利用した認証を必要とするため、万一の障害時に備えてのシステムの透明性を確保することが難しい。また、端末イメージを一斉配信するような環境との連携が難しいことも挙げられる。

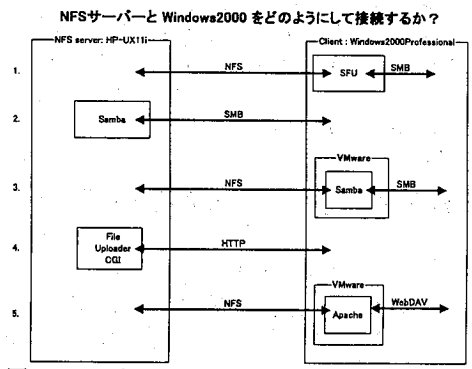

図1. NFSサーバーとWindowsクライアントとをどのようにして接続するか

このようなことを検討しつつ、NFSサーバーとWindowsクライアントとをどのように接続するかについて検討したものが図1である。図中のSFU(Service for Unix)はMicrosoft提供のWindows上で動くNFSクライアントである。Sambaは各種UNIX上で稼動するSMBサーバーソフトウェアである。

「ネットワーク」

教育用計算機システムにおいて発生するトラフィックは講義における指導内容に大きく影響されるため、ある瞬間に集中したトラフィックが発生する可能性が高い。さらには端末とサーバーとを結ぶトラフィックは端末の台数と扱う情報量にそれぞれ比例するため、ここに想定される帯域は非常に大きなものとなる。さらには教育用計算機は学内に分散して配置されるため、建物内のトラフィックよりもむしろ建物間やキャンパス間にまたがるトラフィックを発生させる。

今回のシステムを構築する上で、各種サーバーが配置されネットワークの拠点となるメディアセンター内の計算機室と各学部に分散配置されるサテライト教室との間をどのようなネットワークを利用して接続するかを検討する必要があった。その際に、学内ネットワーク網に依存する形

でメディアセンターのネットワーク網を構築するか、それとも独自のネットワーク網を構築するのかを比較検討することになった。

まず今回のシステムが要求するネットワーク帯域を見積もった。利用者がクライアントを利用し始めた際に必要となるファイル転送量を40MB程度であると見積もった。この容量はファイルサーバー上で個人が利用可能な容量と比較しても妥当な数値である。もし60台の端末で一斉にログイン処理が開始した場合に、この40MBの転送を30秒程度で完了させるためには640Mbpsの帯域が必要となる。この帯域を安定して確保するためには独自のネットワーク網を構築することが妥当であると判断した。

次に各サテライト教室にLayer3スイッチを導入した。これはサテライト教室と学部独自のサーバーとの間で大容量通信が必要となった場合に、学部ネットワークとサテライト教室とを直接接続することを可能とするためである。ここで特別に許可されるトラフィックはファイアウォールを通らないため、セキュリティーを脅かすことがないように末端Layer3スイッチにおいてフィルターをかけることが必要となる。

各端末はそれぞれ末端スイッチに直接収容される形態が好ましい。また、ここで利用される末端スイッチはポートごとにフィルターをかけることが可能である必要がある。これにより末端スイッチのポートにおいてフィルターをすることで端末ごとにネットワーク利用の可否を設定できる。認証サーバーと連携することで、利用者認証を行っていない端末からパケットが流出しないように設定出来る。さらには将来において端末が陳腐化した後においても、利用者認証ベースで接続の可否を制限できる情報コンセントとして活用することも出来る。

「大規模環境における認証」

大規模環境における認証をどのように実装するかについても、十分に検討をしておくべき内容であった。利用者が3万人を超す環境においては、多くの認証システムが破綻することをこれまでに経験してきている。さらに次期システムにおける認証データベースの利用者数は更に増加することが予想されていた。そのため、今回の設計においては利用者数が7万人を超える状態においても稼動するシステムを目標とした。

構築されたシステムの概要

これまでに述べたような検討の結果として、次期システムは以下のような特徴を持つものとなった。

- 学内網とは独立した独自のネットワーク網を構築する
- 認証を受けていない状態ではネットワークを利用できない
- User Interfaceとして Windows環境を提供
- 仮想マシン(VMware)を利用して Linux環境を提供
- 認証は Windowsサーバーに依存をしない形で独自に実装する。
- 携帯電話からの利用にも耐えるメールシステム

ネットワーク

次期システムを設計する上で想定されるトラフィック(640Mbps)は、京都大学の学内ネットワーク網におけるトラフィックの想定を超えるものであったため、次期システムにおいては学内に敷設済みのファイバ網を利用して新たにネットワーク網を構築する方針が妥当であると判断した。このネットワークにおいては冗長性を考慮し、メディアセンター計算機室と各サテライトとを1000Base-LXにより直結する主回線と、学内ネットワーク網にVLANを作成する形で計算機室と各サテライトとを結ぶバックアップ回線とを作成することにした。

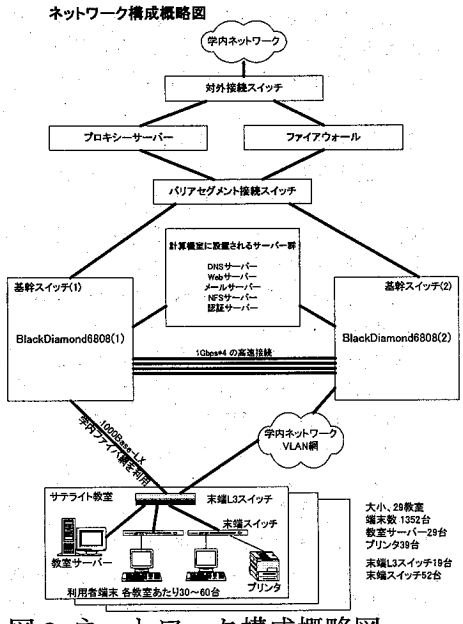

図2: ネットワーク構成概略図

このように独自のネットワーク網を作成した結果として、高速で安定した通信が行えるメリットと同時に、安全な通信を行えることにもなった。

メール

メールサーバーの一番の特徴としては、利用者に対してアナウンスするサーバー名を利用者ごとに異なるものとすることで、サーバー数の増加やメールスプールの移動などを行いやすくしている。標準的に利用されるであろうWebMailを利用する上においては、これらの仕掛けはシステム内に隠蔽されている。このように利用者毎にサーバーを割り振ることが容易になったことで、サーバー数の増減が容易となった。

運用開始当初においては、

- Webmail 2台
- IMAP＋POP 4台
- SMTP＋VirusScan 3台

といった構成で運用している。特定のIPアドレスから著しく多くの接続が同時に行われることを防ぐために、ファイアウォールにおいて同一IP組に対するセッション数の制限を設けた。学外からのメールの送受信は原則としてWebmailを利用して行う。Webmail以外のメーラーの利用を希望する利用者のために、VPNを経由してアクセスする方法も用意した。

メールアカウントはこれまではs60t0305@ip.media.kyoto-u.ac.jpといった機械的に生成されたアカウントを配布していたが、今回はShin@Maruyama.mbox.media.kyoto-u.ac.jpのように、XXX@YYY.mbox.media.kyoto-u.ac.jpのXXXとYYYとを利用者が自由に選択できる形にした。

図3:
利用者に対してアナウンスする内容：

メールを読むためには次の操作をしてください。
Webメールを利用する際には
http://webmail.mbox.media.kyoto-u.ac.jp/
にアクセスして、その後の指示に従う。
(学外からの場合にはSSLを使うことになります)
Webメール以外を利用する場合には次の情報に従って各自で設定を行ってください。

利用コードがa01234567の場合:
SMTPサーバー　a01234567.smtp.mbox.media.kyoto-u.ac.jp
POP3サーバー　a01234567.pop.mbox.media.kyoto-u.ac.jp
IMAPサーバー　a01234567.imap.mbox.media.kyoto-u.ac.jp
アカウント名: a01234567
パスワード: 各自が決めたもの

ダイアルアップ・VPN

学生の下宿の多くが常時接続化されていることから大学のサーバーやファイルなどを学外から利用する環境を提供することは重要である。セキュリティーの問題を考慮しつつこの要求を実現するために、VPN(PPTP)サーバーを設置することにした。ファイルの共有は原則としてWebDAVを利用する。

プリンタ

白黒プリンタ・カラープリンタともに利用者認証を行いつつ、個人ごとに印刷枚数を計測できるシステムを導入することにした。カラー印刷はカラー1ページの出力を白黒10ページに換算して、アカウンティングと利用制限とを行っている。

さらには利用者が利用者端末から印刷要求を出しても即座に印刷されるのではなく、一旦プリンタサーバー内に蓄積されるような設定をした。この設定を行うことで、利用者がプリンタのところに出力された用紙を取りに行く回数を減らすことが出来るため、持ち帰り忘れの用紙が減るのではないかと期待している。

ファイルサーバーの実装

NFSサーバーとNFSクライアントの間で、大きなサイズのコピーをすると、ファイル転送が停止してしまうという相互接続性の問題が発生した。プロトコルの種類などを変えつつ検討した結果、NFSv3/TCPだと問題が発生しないことを確認できたため、今回はこの組み合わせを採用した。

また、WindowsとNFSサーバーとを接続する方法に関しては、図1の各種方針を検討してみたところ、SFUはユーザー数が6万人を超すような規模で運用することが出来ないものであることが判明した。またSambaをUNIX上で大規模運用することにもシステムの安定性の面から不安

があった。結果として、負荷に対しての不安を解消する意味でも3番目の方式が妥当であることがわかった。

VMware内でSambaを動かして、NFS-SMBのプロトコル変換を行う方式を採用したところ、Sambaの行うロック処理がNFSマウントされたボリュームに対して正常に反映されない問題が当初は発生したが、設定変更を行うことで問題を解消することが出来た。パフォーマンス検証の結果、クライアント1台あたり40Mbps以上と、運用に支障のないパフォーマンスが出ている。

WebDAVによる方式も導入した。標準的な設定においてはNFSサーバー上にUTF8コードによるファイル名でファイルが作成されてしまうが、これをEUCコードとなるように修正した。同じファイルシステムをWindowsとLinuxの双方からアクセスする可能性があることを考えると、この変更は妥当であると考えられる。

ファイルサーバーを利用しない形でリムーバブルメディアを接続させるために、USB1.1インタフェイスを提供することにした。(大容量デバイスを接続する可能性が想定されることからUSB2.0を検討したが、時期的に間に合わなかった)大容量ストレージが必要となった利用者自身が半導体メモリーや2.5"ハードディスクのようなデバイスを持ち込めるように配慮している。利用状況を見ると、今後はこの形態で利用者自身がデータを持ち歩くことが広まりそうな勢いである。

利用者環境について

利用者向け端末を提供する上ではWindows環境によるUser Interfaceは必須であると考えられる。しかしながら利用者数が3万人を越す環境においては、Windowsに標準で提供される認証システムは以下のような問題が発生し5年間の運用に耐えられないことが想定された。

・認証サーバーへの負荷の集中
・集中プロファイルの管理コストの増大
・認証情報のサイズの増大
・Windows端末の管理コストの増大

このような問題を技術的に明らかな手段により解決するために、今回の構築においてはWindowsに依存しない認証システムを構築することになった。具体的には、Windowsのログイン認証を行っているGINA.DLLを独自に作成し、認証は任意のプロトコルにより認証サーバーにおいて行うことが可能とした。またログイン後にWindowsを利用する際には内部的には全ユーザーで同一のユーザーIDを用いることで、個人別のプロファイルを管理する必要もなくなった。そして認証を行った際にネットワークを利用することを可能とするように設定した。

Windows端末の設定が変更されることで生じるトラブルも、夜間などに端末のハードディスクイメージをマルチキャスト配信を行うことで解決する。1セグメントあたり30分程度ですべての初期化が完了する。

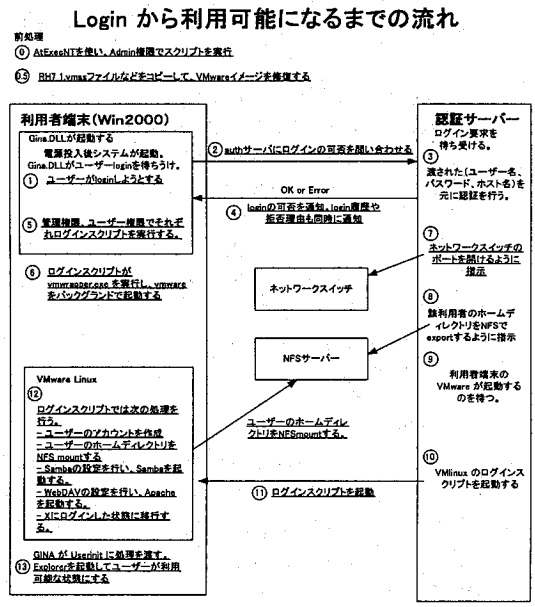

図4: ログインから利用可能になるまでの流れ

まとめ

今回構築されたシステムはこれまでの運用経験に基づき問題の発生しそうな箇所を使うことなく、オープンなシステムとして構築されている。また、負荷の集中しそうな場所はサーバーを追加することできめ細かく負荷の調節が出来る設計となっている。

端末の台数の面においても利用者の人数の面においても大規模な環境の導入事例が、他の多くの管理者の助けとなることを期待する。