マルチメディア通信と分散処理 42－7
（１９８９．５．１９）

分散システム構築におけるネットワーク負荷の評価

齋藤 正史　落合 真一
三菱電機(株) 情報電子研究所

分散システムの構築を行なう際、計算機間の通信速度や計算機の速度の他に接続に使用するネットワークの負荷を考慮する必要がある。そこで、分散システム構築に必要と考えられる基本サービスとして、分散ファイルシステムとディスクレスシステムに注目し、その利用時のネットワーク負荷を測定した。

測定の結果、10Mbpsのネットワークで接続した場合には、ディスクレス計算機で大規模アプリケーション使用時には、2台でネットワークの能力使いきる事が判明した。又、分散ファイルシステムの利用時にはネットワークの負荷が短い期間高くなるが、ネットワークではなく、ファイルシステムを提供するサーバの能力の限界により、性能低下が起こってしまう。

測定に使用した60台程度のシステムでは、分散ファイルシステム、ディスクレスシステムの使用は実用になっていた。しかし、これ以上の台数のシステム構成では、十分な性能が維持できないことがわかった。システム規模の拡大の為には、ファイルサーバ、ファイルキャッシュ手法、メモリ管理手法において改良が必要である。

The Evaluation of the Network Traffic for Distributed System

Masashi Saito　*Shinichi Ochiai*

Information Systems & Electronics Dev. Lab.
Mitsubishi Electric Corp.

In designing a distributed system, we should give consideration to not only communication speed between computers and their computational speed, but also network traffic. We measured the network traffic of an actual distributed system to probe its effect on overall system performance. The system consists of 10 file servers and 50 diskless workstations. File servers are connected via 10 Mbps Ethernet and each file server serves diskless workstations via another 10 Mbps Ethernet. The distributed file facility is provided in such hierarchical network structure of the system.

The reults of measurement indicate that even two diskless workstations running large software causes the over flow of network traffic. It also indicate that the bottle neck of the distributed file system is the sever's perfomance rather than network traffic.

Although the system performance is satisfactory for ordinary usage of the system, it cannot show enough performance for larger system. To expand system scale, we should develop better file servers, file caching techniques and memory management techniques.

1. はじめに

分散システムを構築する為には、使用者にとってそれがローカルの資源であるか遠隔の資源であるかを意識させない程度の性能を持っていること、システム規模を拡大した時にも性能が低下しないことが要求されている。

分散システムを実際に構築する際には、使用者の要求を満足する為に計算機間の通信速度[1][3]、計算機の性能[3]、ネットワークの負荷を考慮する必要がある。

我々は、分散システム構築に必要と考えられる基本サービスとして、分散ファイルシステムとディスクレスシステムに注目し、これらサービス利用時のネットワーク負荷の測定を行なった。本稿では、測定結果を示すとともに、結果より得られた分散システム構築に対する要件を提案する。

実際の測定には、 Sun Microsystems 社の開発した NFS (Network File System) とディスクレスシステムをそれぞれ分散ファイルシステム、ディスクレスシステムの例として取り上げた。

2. 測定条件

測定に使用した分散システム環境を図1に示す。幹線LAN (Ethernet) にファイルサーバを接続し、そこから支線LAN (Ethernet) に4〜6台のディスクレス計算機が接続されており、全体で約60台よりなる分散型システムである。このような接続形態の元で、サーバは支線LAN に接続されているディスクレス計算機の為の私用領域とスワップ領域を提供していると共に、幹線LAN、支線LAN のゲートウェイとして機能している。サーバにおいてNFS (Network File System) を使用し、システム全体でファイルシステムの共有を行なっている。

この環境の元で、NFS使用時のネットワークの負荷、ディスクレスシステム使用時のネットワークの負荷、通常使用状態における負荷について測定を行なった。

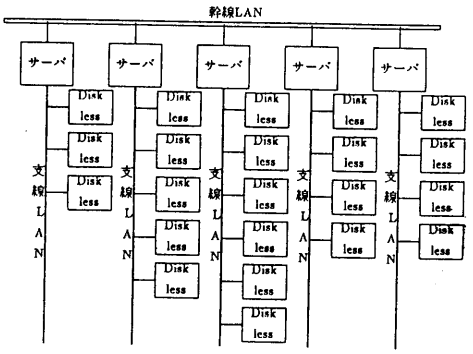


図1. 測定環境

3. NFS使用時のアクセス特性

3.1. 測定項目

NFSを使用した時には、ファイルの先読みやディスクキャッシュがネットワークに及ぼす影響、その時の性能特性を調査する必要がある。そこで、以下の項目について測定を行なった。

1)可変サイズファイルのread/write

逐次read/writeを行なうサイズを変化させた時のネットワークの負荷の測定。又、その時のアクセス時間の測定。

2)複数クライアントからの同時read/write

複数の計算機が同時に異なる2Mbytes のファイルを逐次read/writeを行なう時のネットワークの負荷の測定。又、その時のサーバの負荷(loadaverage)の測定。

3.2. 可変サイズファイルのread/write時の特性

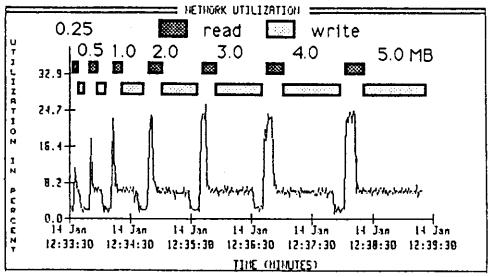
アクセスするファイルのサイズを0.25MBytes, 0.5MBytes, 1MBytesと増加させた時の幹線ネットワークの負荷をグラフ1aに示し、パケットの衝突率をグラフ1bに示す。

ディスクへのwriteの時には、約8%の使用率であり、read の時には25%の使用率を示している。

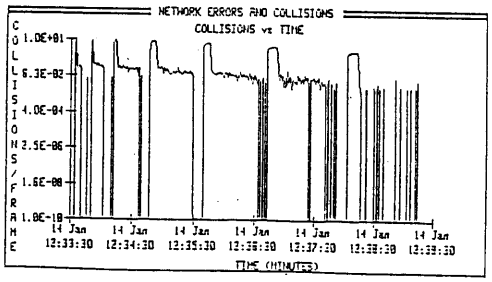
グラフ2に各read/writeに要する時間を示す。グラフ2中の□で示した線が、経過時間(elapse time)であり、○で示した線はOS内での処理にかかる時間(system time)

である。サイズとread/writeに要する時間が比例しているのがわかる。

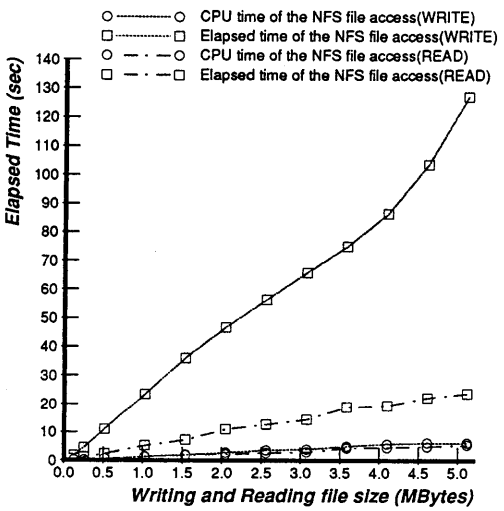
グラフ1a. 可変ファイルサイズのread/write時の負荷



グラフ1b. 可変ファイルサイズのread/write時の衝突



グラフ2. 可変ファイルサイズのread/writeの性能

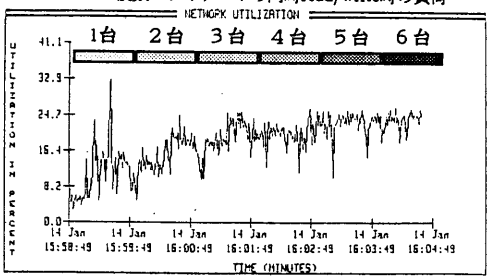


3.3. 複数クライアントからの同時read/write

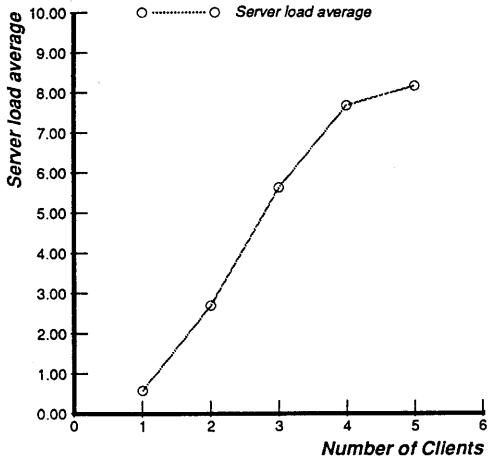
同一サイズのファイル(2MBytes) のread/write を 1〜5台の計算機から行なった時の幹線ネットワークの負荷をグラフ3に示す。

又、計算機の数とサーバの負荷の関係をグラフ4に示す。ここで示している負荷はUnixではuptimeとして知られている値である。

グラフ3. 複数クライアントの同時read/write時の負荷



グラフ4. クライアント数とサーバの負荷



3.4. NFSの特性

以上の測定結果を踏まえて、NFSを使用した場合の特性を考察する。

グラフ2で示した可変サイズファイルのread/write 時の特性により、 NFS のread はサイズが大きくなったとしても、それほど性能低下にはならない。 5MBytes の時には、215KB/sec のread 性能を示している。しかし、writeの時には5MBytes の時に40KB/sec の性能である。

これは、NFSが従来のUnix のディスクアクセス同様のキャッシュ機構を中に持っている事により、readの先読みでのキャッシュヒット率が高い為、高性能を維持している。しかし、writeではファイルの一貫性をとる為と、最新のファイルの属性情報を得る為に、サーバに対してwrite データ全てが送られるので、 readに比べ多くの時間を要する[1]。

ネットワークの負荷を示したグラフ1aに注目してみる。グラフ中、read時のネットワーク負荷は約25%を占めるが、 write時のネットワーク負荷は約8%と低い値となっている。つまり、NFSではreadを高速に実行する為にネットワークに高負荷がかかってしまう。その結果、グラフ1bに示されるように、 read 時にはパケットの衝突による再転送が時々行なわれてしまう。

グラフ4より同時にread/write要求を出す計算機の数の増加にほぼ比例して、ファイルを保持しているサーバの負荷は増加し、クライアント数が4以上でサーバの負荷が飽和している。これは、サーバにキューイングされる要求が増加する事を示し、サーバのへの要求がこれ以上増加しても、サービスしづらくなっている事を示している。

4. ディスクレスシステム使用時のアクセス特性

4.1. 測定項目

ディスクレス計算機ではプロセス実行の為にネットワークをアクセスする。つまり、プロセスサイズとページフォールトの関係やその時のネットワークの負荷特性に注目する必要がある。そこで、以下の項目について測定を行なった。

1)可変サイズプロセスのメモリアクセス

データ空間をランダムに5000回アクセスするプロセスをディスクレス計算機上で動作させ、データ空間を変化させた時のネットワークの負荷の測定。

データ空間量を変化させ、その領域をランダムに5000回アクセスする時の、プロセスの経過時間とOS内処理の時間及びその時のページフォールトの回数の測定。

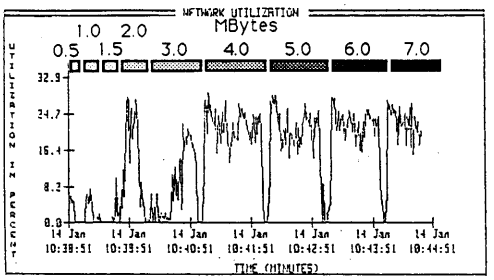
2)複数クライアントからの同時メモリアクセス

2MBytes のデータ空間をランダムに5000回アクセスするプロセスを複数のディスクレス計算機で動作させた時のネットワークの負荷の測定。

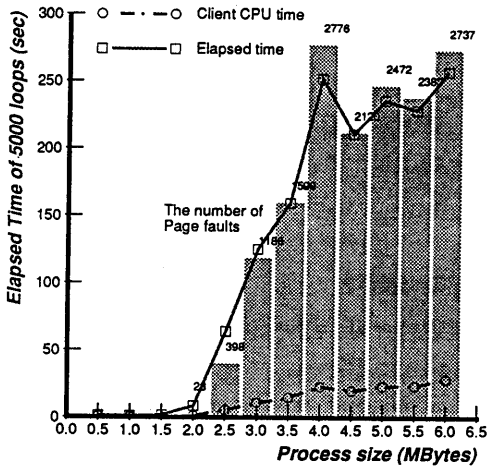
4.2. 可変サイズプロセスのメモリアクセス

データ領域を1MBytesから1分間に 1Mbytes づつ6MBytes まで増加させた時に、そのデータ領域中を1byteづつランダムにアクセスした時の支線ネットワークの負荷をグラフ5に示す。又、グラフ6は各サイズにおける経過時間を□で示し、 OS 内で処理にかかる時間を○を示している。又、同時に棒グラフでページフォールトの回数も示してある。

グラフ5. 可変サイズプロセスのメモリアクセス時の負荷



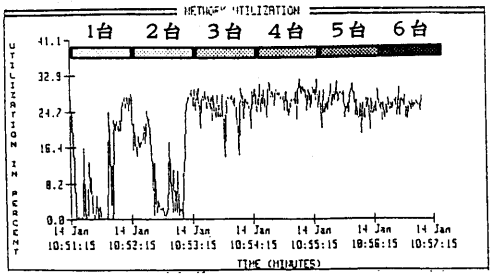
グラフ6. 可変サイズプロセスのメモリアクセス性能と
ページフォールト回数



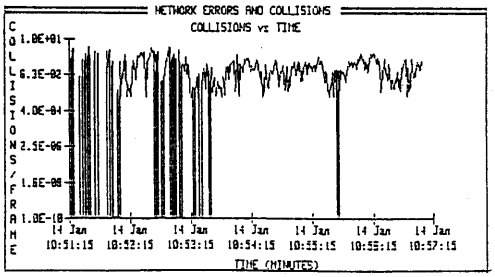
4.3. 複数クライアントからの同時メモリアクセス

データ領域は2Mbytes 固定とし、そのデータ領域中を1byte づつランダムにアクセスするプロセスを1〜6台のディスクレス計算機で実行した時の、支線ネットワークの負荷をグラフ7aに示し、パケットの衝突率をグラフ7bに示す。

グラフ7a. 複数クライアントからの
同時メモリアクセス時の負荷



グラフ7b. 複数クライアントからの
同時メモリアクセス時の衝突



4.4. ディスクレスシステムの特性

4.4.1. ページングの特性

グラフ5によると、データ空間が3Mバイトまではピーク時に20%の使用率を示しているが、定常的に大きくはない。これを越えると使用率が定常的に25%を示し、衝突による再送が非常に増加してしまう。

グラフ6に示されるように3MBytes を境にページフォールトの回数が極端に増加している。従って、上記のネットワーク使用率増加の原因はページフォールトによる物である事がわかる。これは、測定に使用したディスクレスシステムのユーザが使用可能なメモリ領域が約3MBytes であるので、それ以上のプロセスではページフォールトを起こしてしまうことに起因する。グラフ6から、4MBytes 以上のプロセスではアクセスの50% でページフォールトが起こっている事がわかる。

グラフ7aによると、ディスクレス計算機が2台の時からネットワークの使用率が25%という高い値を示している。又、グラフ7bからは、衝突の率も定常的に20%という高い値となっている。

グラフ5、グラフ7aの時にネットワーク使用率が25%以上増加しないのは、おそらくサーバの能力が限界となっているからであると考えられる。

5. 通常使用状態におけるアクセス特性

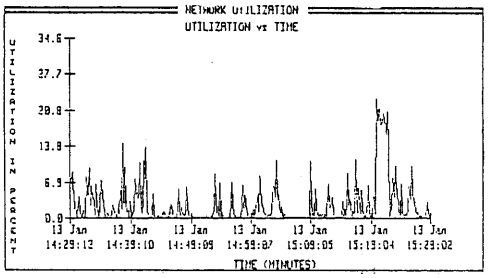
実運用システムでのネットワーク負荷の状況を把握する事も分散システム構築の重要な基礎データとなる。利用内容の詳細は不明であるが、約70% の計算機が使用されている状態のおいて、ソフトウェア開発や文書作成、一部では CAD S/W の使用時のネットワーク負荷の測定を行なった。

定常状態と考えられる1時間の幹線LANの負荷をグラフ8 に、支線LAN の負荷をグラフ9 に示す。基本的にグラフ8 はNFS アクセスの使用のされ方、グラフ9 はページングとNFS によるアクセスである。

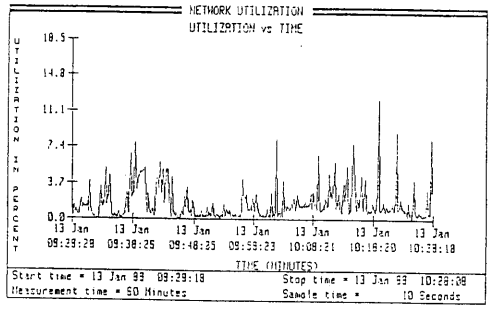
グラフ8により、幹線LAN は定常的には0.75% の使用率であり、ピーク値でも23%という状態であるので、LAN がボトルネックとはなっていない。又、使用率は時々増加しその他の時はほとんど使用されていないという傾向がわかる。これは、NFS でのファイルのreadによる負荷であると考えられ、時々は増加するが、その時間は短く全体に及ぼす影響が少ない事を示している。

グラフ9により、支線LANは定常的には1.72% の使用率であり、幹線LAN の約2.2倍の使用率を示している。しかし、ピーク値は12% であり幹線LANよりも低い負荷となっている。

グラフ8. 通常使用状態の幹線の負荷



グラフ9. 通常使用状態の支線の負荷



6. 測定結果評価

測定結果に対する評価を以下に示す。

1)分散ファイルシステム

- 測定に使用したシステム構成では、NFS は十分に満足できる性能をもっている。
- NFSはread の高速性を達成する為に短い期間ネットワークの負荷は高くなるという特徴がある。
- NFS を利用したシステムでは、ネットワークの負荷ではなく、ファイルサービスを行なうサーバの能力の限界により性能低下が起こってしまっている。

2)ディスクレス計算機

- 測定に使用したシステム構成において、ソフトウェア開発に使用するのであれば6〜8台程度は十分に使用できる。

  これは、ソフトウェア開発においてはエディット作業など小規模なツールをインタラクティブに使用することが多いという特性により、ページフォールトはあまりおこらないからであると考えられる。

・大規模ソフトウェアを動作させた時には、ネットワークの負荷により性能低下が起こってしまう。

・ページフォールトによるネットワークの負荷の増大はネットワークの能力の限界まで達する可能性があるので、LAN を階層構造にする必要がある。

7. 分散システム構築における要件

測定に使用したシステム構成では、NFS、ディスクレス計算機の使用に対して大きな問題とはなっていない。システム規模が拡大した時にはサーバの負荷が増大し、このような構成ではシステムサービスが維持できない可能性がある。そこで、分散システムをより大規模にする時のシステム構成の要件を提案する。

1)ファイルサーバ

測定に使用したサーバでは、同時に4つまでの要求をこなせるがそれ以上は多くの場合待たされるだけである。これは、同時に複数のクライアントにサービスを提供する為にディスクの入出力に伴う待ち時間と、通信に伴うプロトコル処理にかかるCPU時間が主な原因となっている[2]。そのどちらがボトルネックになっているかは今回の測定では不明であるが、ディスクアクセス、プロトコル処理とも高速なファイルサーバを構築する必要がある。

2)ファイルキャッシュ手法

NFS におけるread の高速性は要求をだすクライアント側でのキャッシュ中にread データが先読みにより保持されている事による。readデータの先読みは、ネットワークに対して短い期間高い負荷をかけるが、キャッシュのヒット率が向上する事により高負荷となる回数が減少する可能性があり、全体的なネットワーク負荷はそう高くならないと予測される。

ネットワークの負荷をあげないようにする為には、クライアント側でのキャッシュが重要である。キャッシングされたデータの一貫性を保ち、且つ有効に利用できるようなファイルキャッシュ手法の導入が必要である[3]。

3)メモリ管理手法

ディスクレス計算機がネットワークの能力を使いきるのは、プロセス実行時にページアウトされている空間をページインする事による。ディスクレス計算機のメモリを増加する事で、ある程度は改善されるが増加したメモリ量以上のプロセスの実行時には同様の状況となってしまう。

ディスクアクセスの少ないプロセスに対しては、ディスクキャッシュ用のメモリを仮想記憶用のメモリに動的に変更することにより、プロセス用のメモリを増加させページフォールトをなるべく起こさないというようなメモリ管理手法を導入する必要がある[3]。

ディスクレス計算機を個人で使用する際には、定型的に使用することが多いと予測される。従って、各アプリケーションプロセス毎に最適なメモリ管理方式を導入する事で、高速性が期待できる[4]。

このように、分散システムを構築する為には現状の技術の組合せでは不足している点があり、解決するべき技術課題は多く残されている。

8. おわりに

本稿では、NFS とディスクレスシステム使用時のネットワークの負荷の測定結果を示し、その測定結果の評価を行ない、システム規模の拡大のために解決するべき要件を提案した。

NFS、ディスクレスシステムの使用では測定に使用した60台程度では実用になっている。しかし、これ以上の台数のシステム構成では、十分な性能が維持できないことがわかった。システム規模の拡大の為には、ファイルサーバ、ファイルキャッシュ手法、メモリ管理手法において改良が必要である。

今後は、これらの要件を実現した分散システムの構築を行なっていく予定である。

参考文献

[1] R.Sandberg et.al.:Design and Implementation of the Sun Network Filesystem, USENIX 1985 Summer

[2] S.Menees et.al.:Scale and Performance in Distributed File System, ACM Transaction on Computer Systems Feb. 1988

[3] M.Nelson et.al.:Caching in the Sprite Network File System, ACM Transaction on Computer Systems Feb. 1988

[4] M.Young et.al.:The Duality of Memory and communication of a Multiprocessor Operating System, Technical Report, Carnegie-Mellon University, Feb. 1987