システムソフトウェアと
オペレーティング・システム 76－6
（1997．8．21）

二次サーバを活用した広域分散ファイルシステム Aria

宮澤 元　千葉 滋　益田 隆司

東京大学 大学院 理学系研究科 情報科学専攻
〒113 東京都 文京区 本郷 7-3-1

要旨

本稿ではファイルサーバ(一次サーバ)の処理の一部を肩代りするホスト(二次サーバ)を広域ネットワーク(WAN)上の適切な場所に配置することによって、広域分散ファイルシステムの性能を改善できることを示す。二次サーバを使うと、WANの低速部分の通信量が削減され、さらに一次サーバの負荷が二次サーバに分散されることによってファイルシステムの性能が改善される。我々は、二次サーバを用いて広域分散ファイルシステムの性能を改善できる事を確かめるために、広域分散ファイルシステム Aria を開発した。プロトタイプを用いた実験により、低速部分の通信量が削減されファイルシステムの性能が改善される効果を確かめた。実験に用いた環境では、一次サーバと二次サーバ間の通信速度が、二次サーバとクライアント間の通信速度の約 40% 以下ならば、二次サーバを利用した方が広域分散ファイルシステムの性能を改善できることが明らかになった。

Aria: A wide-area distributed file system that utilizes secondary servers

Hajime Miyazawa, Shigeru Chiba and Takashi Masuda

Department of Information Science, Graduate School of Science,
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113

Abstract

This paper points out that the performance of wide-area distributed file systems can be improved if *secondary servers*, which are file servers that serve collections of files instead of the original file server(the primary server), are appropriately placed on network. This is because the amount of the communication over low speed networks is reduced and because the primary server processes less file requests when secondary servers exist. To show the effectiveness of this idea, we have developed a wide-area distributed file system called *Aria*. Our experiments using this file system revealed that secondary servers improved the performance if the network bandwidth between the primary server and the secondary server was about less than 40% of the bandwidth between the secondary server and the clients.

1 はじめに

本稿では、ファイルサーバ(一次サーバ)の処理の一部を肩代りするホスト(二次サーバ)を広域ネットワーク(WAN)上の適切な場所に配置することによって、広域分散ファイルシステムの性能を改善できる場合があることを示す。ここでいう二次サーバとは、元のファイルサーバ(一次サーバ)の持つファイルの一部を受けとり、以後一次サーバにかわってそれらのファイルのファイルサーバとして働くホストであり、典型的にはクライアントからネットワーク的に近い場所に配置される。ファイルを参照するクライアントの配置には局所性があり、特定のファイルを参照するクライアント群は互いにネットワーク的に近い場所に配置されていることが多いので、二次サーバを使うことによって広域分散ファイルシステムの性能を改善できる。

二次サーバを使って性能が改善できる具体的な理由は二つある。一つ目は、低速なネットワークを介した通信の量を削減できるからである。現実の WAN では任意の2つのホスト間のネットワークの速度が一定ではなく不均一であり、一次サーバと二次サーバの間の通信が遅く、二次サーバとクライアントの間の通信が速くなるような場所に二次サーバを配置することによりファイルシステムの性能を改善できる。二つ目は、一次サーバの処理の一部を二次サーバが肩代りすることによって負荷分散が行なわれるからである。一次サーバは二次サーバに渡したファイルに対してはファイルサーバとして働かなくてよいので、その分性能が改善できる。

二次サーバを利用して広域分散ファイルシステムの性能を改善できる事を確かめるために、我々は二次サーバを利用する広域分散ファイルシステム Aria を設計し、プロトタイプを Unix 上にエミュレータとして実装した。プロトタイプを用いた実験により、低速ネットワーク上の通信量が削減される効果について、一次サーバと二次サーバ間の通信速度と二次サーバとクライアント間の通信速度の違いがどのくらいならば、二次サーバを利用することによって性能を改善できるかを調べた。

2 二次サーバを用いた広域分散ファイルシステム

ファイルサーバ(一次サーバ)の処理の一部を肩代りするホスト(二次サーバ)をネットワーク上の適切な場所に配置することによって分散ファイルシステムの性能を改善することができる場合がある。

2.1 二次サーバとは

二次サーバとは一次サーバの持つファイルの一部を受けとり、それ以後それらのファイルのファイルサーバとして働くホストのことである。ファイルの内容と共に管理権限も一次サーバから二次サーバへ移動するので、あるファイルに対して二次サーバを生成すると、それ以後一次サーバはそのファイルに対する要求を処理しない。一次サーバがクライアントからそのようなファイルに対する要求を受けた場合は、二次サーバに要求し直すようにクライアントに通知する。また、一つのファイルのファイルサーバとして働く二次サーバが複数存在することもない。

一次サーバから二次サーバへのファイル転送はクライアントの要求に従って行われる。あるファイルに対して二次サーバが生成されるときには、ファイル転送は行われず、クライアントからそのファイルへの要求を最初に受けたときに二次サーバは一次サーバにファイル転送を要求する。一度一次サーバから二次サーバに転送されたファイルは二次サーバに保持されるので、同じファイルへの二度目以降の要求に際しては一次サーバから二次サーバへの転送は行われない。

二次サーバの保持しているファイルが更新された場合でも、二次サーバがファイルを保持している限り、この更新は一次サーバには反映されない。更新の一次サーバへの反映は、二次サーバが消去される時に一括して行われる。

二次サーバは、ファイルの使用状況などの情報を考慮して、ユーザないしはシステム管理者の明示的な指定によって生成され、WAN 上の適切な場所(典型的にはクライアントが接続されているのと同じローカルエリアネットワーク(LAN)上)に配置される[1]。また、二次サーバの消去も同様に明示的な指定によって行なわれる。このとき、二次サーバが保持しているファイルの管理権限が一次サーバに戻され、更新されたファイルの内容が一次サーバに反映される。二次サーバが消去されると、一次サーバは二次サーバから管理権限が戻された

[1] ファイルの使用状況に関する統計などを利用して、二次サーバを自動的に配置するのは今後の課題である。現時点では人間が二次サーバの配置を決定するのが最も簡単で確実であると思われる。

ファイルのファイルサーバとして働く。

2.2 二次サーバによる性能の改善

二次サーバを使って広域分散ファイルシステムの性能を改善できる理由を述べる。

2.2.1 低速ネットワーク上の通信量の削減

WAN を単純化して考えると、高速ネットワークで接続された複数のクライアントが LAN を構成し、さらに LAN 同士が低速ネットワークを介して相互に接続されている。LAN 同士の間のネットワークに ATM のような高速ネットワークが使われる場合でも、通信の混雑のため実効速度は低いと考えられる。

広域分散ファイルシステムでは、一般にファイルサーバとクライアントは別々の LAN に属しており、これらが通信するには低速ネットワークを経由しなければならないため時間がかかる。そこで、二次サーバをクライアントの接続されている LAN 上に配置すれば、低速ネットワーク上の通信量を削減して、ファイルシステムの性能を向上することができる。

二次サーバを使うと、以下のような場合に低速ネットワーク上の通信量を削減することができる。

- クライアントはファイル更新時にキャッシュを同じ LAN 上の二次サーバに書き戻す。二次サーバがない場合に低速ネットワークを介して一次サーバに書き戻さなければならないのに比べ、低速ネットワーク上の通信量が削減される。

- 同一 LAN 内で複数のクライアントが同じファイルにアクセスする場合に、低速ネットワークを介したファイル転送の回数を減らすことができる。この場合、低速ネットワーク上のファイル転送が起きるのは一台目のクライアントがアクセスする際に一次サーバから二次サーバへ転送する時の一回だけであり、二台目以降のクライアントは、LAN を経由して二次サーバが保持しているファイルにアクセスするだけで済む。したがって、各クライアントが個別に一次サーバにアクセスする場合に比べて全体としての性能が向上する。

2.2.2 二次サーバによる負荷分散

一次サーバがファイルを二次サーバに渡すと、それ以後一次サーバはそのファイルへの要求を処理しない。その分一次サーバの負荷が軽減され、全体としての性能が改善される。クライアントの合計台数が同じならば、ファイルアクセスのレイテンシが小さくなる。一方、レイテンシが同じならば、より多くのクライアントのファイル要求を処理する事ができ、スケーラビリティが向上する[2]。

3 Aria ファイルシステム

我々は二次サーバを利用した広域分散ファイルシステムとして Aria ファイルシステムを開発している。本節では、Aria ファイルシステムの構成とプロトタイプを用いた実験について説明する。

3.1 Aria ファイルシステムの構成

Aria ファイルシステムの構成はカーネギーメロン大学 (CMU) で開発された Andrew ファイルシステム (AFS)[1, 2, 3] に二次サーバ機能を追加した形になっている。主要な構成要素は、ファイルサーバ (一次サーバ)、二次サーバおよびクライアントである。

- 二次サーバ

  二次サーバの生成と消去はユーザの明示的な指定により行なわれる。二次サーバを生成する時には、ユーザはコマンドを使って二次サーバとするべきホスト名と一次サーバから二次サーバに渡すファイル (ディレクトリ) とを指定する。消去する時にも同様に、二次サーバのホスト名と、二次サーバ生成時に一次サーバから渡されたファイル (ディレクトリ) を指定する。二次サーバの有無はクライアントが自動判別するので、ユーザは二次サーバの生成・消去の指定を行なう以外は、二次サーバを使っていることを意識しないでよい。

  二次サーバが生成された後、一次サーバから二次サーバへのファイル転送は、クライアントからのファイル要求に従って行なわれる。クライアントが、二次サーバ生成時に指定したファイル (ディレ

[2]ただし、この負荷分散によって性能が改善される効果はまだ実験で確かめられていない。

クトリ) を二次サーバに要求すると、二次サーバは一次サーバからそのファイル (ディレクトリ) を転送し、ローカルディスク上に保持する。

ローカルディスク上に保持したファイルに関して、二次サーバは AFS のファイルサーバと同様に動作する。二次サーバは、保持しているファイルをクライアントから要求されると、一次サーバにアクセスせず、保持しているファイルをそのままクライアントに転送する。また、二次サーバはクライアントから更新されたファイルが書き戻された時に、そのファイルをキャッシュしている別のクライアントが存在すれば、そのクライアントのファイルを無効化する。

二次サーバは、消去される時に、ローカルディスク上に保持しているファイルのうち、更新されたファイルを一次サーバに書き戻す。

- 一次サーバ

  Aria の一次サーバは、二次サーバが生成されていない場合には AFS のファイルサーバと同様に動作する。一次サーバは、クライアントからのファイル要求を受ければそのファイルを転送し、クライアントから更新されたファイルが書き戻されれば必要に応じて別のクライアントの持つキャッシュを無効化する。

  二次サーバが生成された後は、一次サーバは二次サーバからの要求に従ってファイルを二次サーバに転送する。二次サーバに転送したファイルに対する要求を受けとった場合、一次サーバはファイル (ディレクトリ) を要求したクライアントに対して二次サーバにその要求を出し直すように通知する。これにより、クライアントは一次サーバに対する要求を自動的に二次サーバに振り向ける。例えば、既にクライアント A がマウントしていたファイルに対して二次サーバが生成され、そのファイルが二次サーバに転送されてしまったとする。その後、クライアント A からのファイル要求を受けると、一次サーバはクライアント A に対して二次サーバにファイルを要求し直すように通知する。

- クライアント

  クライアントは一次サーバのディレクトリを指定してマウントする。マウントしたディレクトリ上のファイルがアクセスされると、システムは一次サーバにファイル要求を出す。

  マウント要求やファイル要求を出す時に、もし二次サーバが作られていれば、一次サーバから二次サーバに要求を出し直すよう通知されるので、それ以降は直接二次サーバに要求を出す。

  要求したファイルが (一次または二次) サーバから送られた後は、AFS と同様にそのファイルをローカルディスクにキャッシュして処理を行なう。

3.2 実験

現在、Aria ファイルシステムは、そのプロトタイプが FreeBSD 上のエミュレータとして実装されている。このプロトタイプを使って実験を行ない、一次サーバと二次サーバの間の通信速度と二次サーバとクライアントの間の通信速度の違いがどの程度なら、二次サーバを使った方がより性能が向上するかを調べた。

実験は Ethernet(10Mbps) の孤立ネットワーク上で行ない、一次サーバ、二次サーバ、クライアントのそれぞれに表 1に示す仕様を持つホストを使用した。

実験には、CMU で開発された Andrew ベンチマークを使用した。Andrew ベンチマークとは、あらかじめ用意されたプログラムソースのディレクトリツリーに対し、(1) 同じディレクトリ構造を作る、(2) ディレクトリをコピーする、(3) 再帰的に stat を実行する、(4) 再帰的に grep を行なう、(5) コンパイルする、の各フェーズを連続して実行し、所要時間を計るものである [2]。

二次サーバとクライアントの間のネットワークの速度は 940KB/s で固定し、一次サーバと二次サーバ (二次サーバを使わない場合はクライアント) の間のネットワークの速度をソフトウェアのエミュレーションによって変化させた。940, 200, 100, 67, 40, 29, 20KB/s の 7 通りのネットワークの速度に対して Andrew ベンチマークを実行した結果を図 1に示す。二次サーバを使う場合と使わない場合のグラフの交点から、一次サーバと二次サーバの間のネットワークの速度が約 360KB/s 以下ならば二次サーバを使って性能が改善できることがわかる。これは二次サーバとクライアントの間のネットワークの速度 940KB/s の約 40% にあたる。

なお、複数のクライアントで同一ファイルを共有する場合に二次サーバを使って広域分散ファイルシステムの

	CPU	メモリ	ハードディスク
一次サーバ	PentiumPRO 200MHz	64MB	SCSI 1GB
二次サーバ	Pentium 166MHz	40MB	SCSI 1GB
クライアント	Pentium 120MHz	16MB	E-IDE 800MB

表 1: 各ホストの仕様

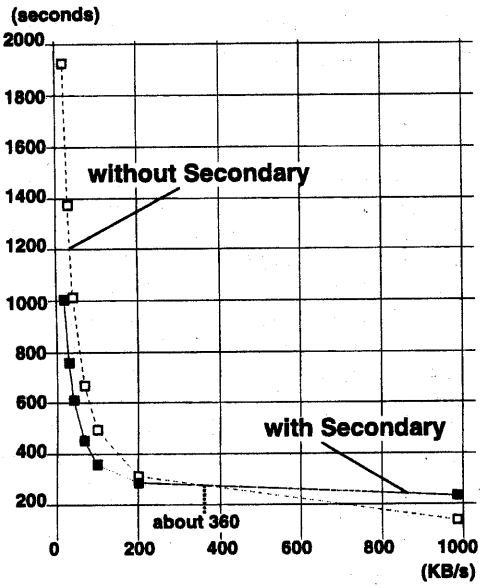

図 1: ネットワーク速度と Andrew ベンチマークの関係

性能を改善できる効果についても実験により確かめたが、本稿では紙面の都合上割愛する。

4 関連研究

この章では、広域分散ファイルシステムに関連する研究について紹介する。

- AFS[1, 2, 3]
  CMU で開発された代表的な広域分散ファイルシステムである。WAN で使用されるため、ファイルサーバ一台当たりのクライアントの台数が多く、スケーラビリティを向上するためにサーバの負荷を軽くするのが主な設計目標となっている。このため、ファイルの全体を一度に転送する、クライアントのローカルディスクにファイルをキャッシュする、callback を使ってクライアントのキャッシュを無効化する、セッションセマンティクスを採用するなどの技術を使っている。

- Coda[4, 5, 6]
  AFS の後継システムであり、WAN でしばしば起こると考えられるネットワークの故障に対応するためにサーバの複製を行っている。さらに Coda では、耐故障性を高めるため、複製されたサーバのいずれにもアクセスできない場合にも、クライアントの持っているキャッシュだけを使って作業を継続できる disconnected operation 機能も追加されている。

  サーバを複製する手法は、全く同じファイルを複数のサーバが供給するという点で、二次サーバ方式とは異なる。二次サーバ方式では、一次サーバはファイルの管理権限まで含めて二次サーバにファイルを渡すので、あるファイルのサーバは常にひとつである。また、サーバを複製する場合、各サーバの持つファイルの一貫性保持の問題が生じる。通信速度の遅いネットワーク上では、厳密な一貫性を保証するのは困難であり、サーバを複製するコストもかなり高いと考えられる。

- Multi-level caching ファイルシステム [7]
  ミシガン大学において AFS への拡張の形で開発されているシステムである。ファイルサーバとクライアントとの間に、何段階かのキャッシュ・サーバ (中間サーバ) を置くことにより、一次サーバの負荷を下げてスケーラビリティを向上している。中間サーバからファイルサーバへの書き戻しを遅延させることにより、遅いネットワークでの性能を改善したり、移動体計算機への対応を進めたりといった研究も行われている [8]。

  二次サーバ方式では、二次サーバが持っているファイルは、二次サーバを消去するまで一次サーバに書

き戻しを行わないのに対し、キャッシュ・サーバはファイルサーバからみればあくまでクライアントの一つであり、ファイルが更新されればファイルサーバにキャッシュの書き戻しを行う点が異なる。遅いネットワーク上で使う場合、このキャッシュの書き戻しのオーバーヘッドがかなり大きくなってしまうと考えられる。

- Web Proxy[9]
  もともとファイアウォールの中から World-Wide Web(WWW) へのアクセスを提供するための技術である。ファイアウォール内部のクライアントは proxy を介して外部へアクセスする。外部からの WWW データを proxy にキャッシュすることにより、外部の遅いネットワークを介した通信を削減して性能を向上できる。キャッシュは、ファイアウォールの有無に関わらず有効なので、ファイアウォールを使用しない環境でも proxy を使うケースは多い。

  Web proxy は、WWW の世界における二次サーバのようなものだと考えることができる。ただし、Web proxy ではクライアントによるファイルの更新が考慮されていない点が、二次サーバを用いるファイルシステムとは本質的に異なる。

5 まとめと今後の課題

二次サーバを活用することによって広域分散ファイルシステムの性能を改善できる場合があることを示した。二次サーバによって性能が改善できる理由は二つあり、一つは WAN の速度の不均一性に注目して低速ネットワーク上の通信量を削減できること、もう一つは一次サーバの処理の一部を二次サーバが肩代りすることによって負荷が分散されることである。二次サーバを使って広域分散ファイルシステムの性能を改善できることを示すために二次サーバを使うファイルシステムとして Aria ファイルシステムを設計し，そのプロトタイプを FreeBSD 上のエミュレータとして実装した。さらにプロトタイプを用いた実験を行い、WAN のネットワーク速度の不均一性に注目した場合に二次サーバ方式が有効と考えられるネットワークの速度の条件について調べた。実験に用いた環境では，一次サーバと二次サーバの間のネットワークの速度が，二次サーバとクライアントの間のネットワークの速度の約 40% 以下ならば，二次サーバを使う方式が有効であることが明らかになった。

二次サーバによって一次サーバの負荷が分散されることによる性能改善効果についてはまだ実験で確かめておらず、今後の課題である。

謝辞

本研究では、Mahadev Satyanarayanan 氏が CMU で開発した Andrew ベンチマークを使用しています。

参考文献

[1] Satyanarayanan, M., Howard, J. H., Nichols, D. A., Sidebotham, R. N., Spector, A. Z. and West, M. J.: The ITC Distributed File System: Principles and Design, in *Proceedings of the 10th ACM Symposium on Operating Systems Principles*, pp. 35–50 (1985).

[2] Howard, J. H., Kazar, M. L., Menees, S. G., Nichols, D. A., Satyanarayanan, M., Sidebotham, R. N. and West, M. J.: Scale and Performance in a Distributed File System, *ACM Transactions on Computer Systems*, Vol. 6, No. 1, pp. 51–81 (1988).

[3] Spasojevic, M. and Satyanarayanan, M.: An Empirical Study of a Wide-Area Distributed File System, *ACM Transactions on Computer Systems*, Vol. 14, No. 2, pp. 200–222 (1996).

[4] Satyanarayanan, M., Kistler, J. J., Kumar, P., Okasaki, M. E., Siegel, E. H. and Steere, D. C.: Coda: A Highly Available File System for a Distributed Workstation Environment, *IEEE Transactions on Computers*, Vol. 39, No. 4, pp. 447–459 (1990).

[5] Kistler, J. J. and Satyanarayanan, M.: Disconnected Operation in the Coda File System, in *Proceedings of 13th ACM Symposium on Operating Systems Principles*, pp. 213–225 (1991).

[6] Mummert, L. B., Ebling, M. R. and Satyanarayanan, M.: Exploiting Weak Connectivity for Mobile File Access, in *Proceedings of the 15th ACM Symposium of Operating Systems Principles*, pp. 143–155 (1995).

[7] Muntz, D. and Honeyman, P.: Multilevel Caching in Distributed File Systems, in *Proceedings of USENIX 1992 Winter Conference* (1992).

[8] Muntz, D. A., Honeyman, P. and Antonelli, C. J.: Evaluating Delayed-Write Policy in a Multilevel File System, in *Proceedings of 2nd Intl. Conf. on Dist. Platforms*, pp. 415–429 (1996).

[9] Luotonen, A. and Altis, K.: World-Wide Web Proxies, in *Proceedings of the First International WWW Conference* (1994), `http://www1.cern.ch/PapersWWW94/luotonen.ps`.