Mogami: 高遅延環境において広帯域を達成する分散ファイルシステム

堀内美希[†]　田浦健次朗[†]

1. はじめに

大量の医学文書のインデクシングを行ったり，大量の天文学画像データを用いて超新星の発見を試みたりなどの，データ集約的アプリケーションでは，複数の拠点間にまたがって大規模なデータ解析を行うことがある．その際，広域分散ファイルシステムによって，ローカルファイルを扱う感覚で透過的に共通のストレージを扱えると都合が良い．このように，データ解析のワークフローの基盤として使用される広域分散ファイルシステムだが，高遅延環境下でデータ転送を行うと実効帯域幅が物理的に可能とされる帯域幅と比べて大きく劣ることがある．アプリケーションから分散ファイルシステムに対して必要な部分のデータを要求し，遠隔のノードから実際のデータを転送するが，この要求データサイズが小さすぎると，要求と転送を何度も繰り返すこととなり，特に高遅延環境では全体として物理帯域を使い切ることができない．今後，このようなデータ解析において扱うデータのサイズがますます増大することを考慮すると，広域分散ファイルシステムの高遅延広帯域環境への最適化は必須である．そこで，存在する物理帯域を有効利用し，高遅延環境においても高いデータ転送性能が出る広域分散ファイルシステム Mogami を提案し，実装，評価する．

2. 関連研究

Gfarm[1] は，複数拠点をまたいで利用できる広域分散ファイルシステムである．単一のメタデータサーバと複数のストレージサーバを束ねる構成をしており，Mogami に近い構成である．Gfarm はデータ転送を行う際に，現在の実装では 1MB 程の単位でファイルデータの要求とデータ転送を繰り返す．そのため，高遅延広帯域環境において大きなファイルを読み出す際には高遅延の影響を受けて広帯域を活かしきれないことがある．Gfarm を用いてデータ転送実験を行ってみたところ，物理的に可能とされる帯域幅が 10Gbps，iperf を用いたデータ転送時の帯域幅が 2.5Gbps，遅延が約 27msec の環境で，1GB のファイルを cat コマンドで読み出した結果，実効帯域幅は 130Mbps 程度となった．一度に転送するデータサイズを単純に大きくすれば，シーケンシャルな読み込み性能は改善されるが，ランダムアクセスの際には無駄なデータ転送が起こることとなってしまう．

他にも，Lustre[2]，Ceph[3]，GPFS[4] 等の分散ファイルシステムが存在するが，これらは広域環境で使用されることを想定して設計されていない．仮に広域環境で使用することが可能であったとしても，高遅延環境を想定した最適化は行っておらず，物理的に可能とされる帯域を満足する通信を行うことはできないと予測される．

3. 設計・実装

広域分散ファイルシステム Mogami を設計・実装するにあたり，高遅延環境においても広帯域を達成可能とするためのアプローチとして，以下のような手法をとる．

- ファイルをシーケンシャルに読み込んでいる時にのみ，先のデータも使用されることを予測し，空いている帯域を効率的に用いてあらかじめデータ転送を行っておく（プリフェッチ）
- 高遅延環境に TCP 輻輳制御パラメータを最適化する

本稿では，前者のアプローチに注目し実装と評価を行った．

また Mogami は，単一のメタデータサーバ，任意の数のストレージサーバ・クライアントを持つ構成とする（**図 1**）．Mogami でファイルの読み込みをする際の手順を図 1 中に記す．

クライアントのファイルシステムは，FUSE[5] を用いる．これにより POSIX 互換の API を用いて，ファイルの読み書き等の操作をすることが可能となる．

4. 性能評価

高遅延環境下での使用に最適化した広域分散ファイ

† 東京大学大学院 情報理工学系研究科

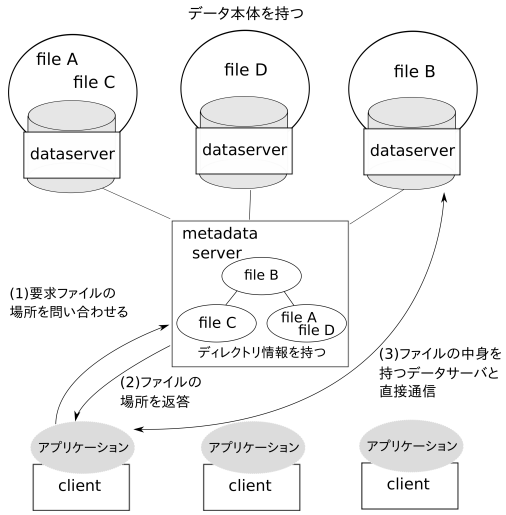

図 1 Mogami のシステムコンポーネント
Fig. 1 System component of Mogami

ルシステム Mogami の性能評価を行う．Mogami は高遅延環境下でのデータ読み込み実効帯域幅の向上を目標としている．そこで実際に高遅延広帯域な実験環境において，ファイルのデータ読み込みを行い，実効帯域幅を測定することによって評価を行う．実験環境は InTrigger プラットフォーム[6] の東京-福岡（物理帯域=10Gbps，RTT=約 27msec）の拠点であり，東京に存在するファイルを福岡で読み込んだ場合の，データ転送帯域幅を iperf，Gfarm，sshfs[7]，Mogami 間で比較する．ファイルの読み込みは，一回につき 64KB ずつ，ファイルの先頭から末尾までシーケンシャルに行う．ここで bs とはブロックサイズであり，Mogami でクライアントがデータサーバに一度に要求・転送するデータのサイズである．

性能評価として，読み込んだファイルの大きさと実効帯域幅の関係を，**図 2** に示す．Mogami でプリフェッチを行ってデータを先読みした時の実効帯域幅は，1GB のファイルを読み込んだ時に関して述べると，Mogami でプリフェッチを行わない時や他ファイルシステムの 7 ～ 10 倍程になっている．結果として，プリフェッチを用いると，iperf には及ばないものの，ファイル読み込みの帯域幅をかなり大きくすることができることを確認できた．

また，今回はランダムにファイルを読み込んだ場合について言及していないが，Mogami の場合，シーケンシャルにファイルが読み込まれている時のみプリフェッチを行うため，ランダム読み込み時に無駄なデータ転送を行うことはない．

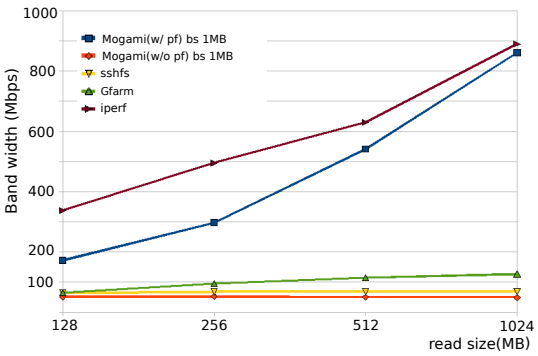

図 2 ファイルの読み出し性能比較
Fig. 2 Comparison of file reading performance

5. おわりに

高遅延環境においても，広帯域を達成する広域分散ファイルシステム Mogami を提案し，実装，評価を行った．結果として，Mogami で用いたアプローチにより，高遅延広帯域環境においての広域分散ファイルシステムのデータ転送速度を大幅に向上することができた．今後の課題としては，未実装となっている，TCP の輻輳制御パラメータの高遅延環境への最適化をはじめ，複数クライアントの通信競合の回避等が挙げられる．加えて，単に広域分散ファイルシステムとしての性能向上を目標とするだけではなく，上位レイヤで実行することになるワークフローとの親和性が高く，手軽にかつ高性能に広域分散環境での大規模なデータ解析を可能とする，広域分散ファイルシステムの構築を目指す．

参 考 文 献

1) Osamu Tatebe, K. H. and Soda, N.: Gfarm Grid File System, *New Generation Computing*, Vol. 28, (2010).
2) Lustre file system :http://www.lustre.org/.
3) Weil, S. A., Brandt, S. A., Miller, E. L., Long, D. D. E. and Maltzahn, C.: Ceph: A scalable, high-performance distributed file system, in OSDI '06, pp. 307–320, Berkeley, CA, USA (2006), USENIX Association.
4) Schmuck, F. and Haskin, R.: GPFS: A shared-disk file system for large computing clusters, in *Proceeding of the Conference on File and Storage Technologies*, pp. 231–244 (2002).
5) FUSE :http://fuse.sourceforge.net/.
6) InTrigger Platform :http://www.intrigger.jp/.
7) sshfs :http://fuse.sourceforge.net/sshfs.html.