情報処理学会論文誌　コンピューティングシステム　Vol.5　No.5　128–137 (Oct. 2012)

広域分散ファイルシステムのための適応的な先読み手法

堀内 美希[1,a)]　田浦 健次朗[1,b)]

受付日 2012年3月28日, 採録日 2012年6月9日

概要：分散環境でデータ集約的計算を行う際に重要な役割を果たす分散ファイルシステムであるが，高遅延環境でのファイルアクセスでは遅延の影響を受け，大幅にスループットが下がってしまうことがある．それを回避するために，アプリケーションに変更を加えず適応的に実行可能なアクセスデータの先読み手法を提案する．提案手法により，評価に用いた高遅延広帯域環境下でのファイルアクセスでは，シーケンシャルアクセスで約 700～800%，ストライドアクセスで約 300～400%の読み込みスループット向上を確認することができた．

キーワード：分散ファイルシステム，先読み，自動調節システム，高遅延環境

An Adaptive Prefetching Method for Distributed File Systems

Miki Horiuchi[1,a)]　Kenjiro Taura[1,b)]

Received: March 28, 2012, Accepted: June 9, 2012

Abstract: Distributed file systems play an important role for data intensive computation but current systems often fail to achieve good throughputs in high latency environments. To achieve a good access throughput, we propose a data prefetching method that can be adatively applicable without any modification to applications. In the evaluation environment with high latency and wide bandwidth, the proposed method achieved the "read" performance improvement of around 700 ∼ 800% in sequential access and of around 300 ∼ 400% in stride access.

Keywords: distributed file systems, prefetching, automated tuning system, high-latency environment

1. はじめに

近年，マルチコアアーキテクチャの普及，解析対象データサイズの増大，分散ノードを扱うための環境の整備等により，分散環境で並列計算を行う機会が増えている．Montage [1] 等の，分散環境で大規模なデータを扱って複雑なデータ解析を行うワークフローアプリケーションも登場しており，このような分散環境での並列計算を行う際には，共有ファイルをローカルファイルを扱うのと同じ通常のファイルシステム API を用い透過的にアクセスできるとユーザからみて扱いやすい．そのため分散環境でのワークフローアプリケーション実行の際には，よく分散ファイルシステムが基盤システムとして用いられる．

しかし，複数拠点間にまたがったデータ解析では，高遅延環境で分散ファイルシステムを基盤としたファイルアクセスを行い，遅延の影響を受けて高スループットを達成できない場合がある．これを回避するためにはデータの先読みを行うことが考えられ，本稿では広域分散ファイルシステムのためのデータ先読み手法を提案する．本研究の貢献は以下である．

- 分散ファイルシステムにおいてファイルのどの部分の先読み要求を行えばよいか判断するため，ファイルアクセスの規則性を検出し，その規則性に基づいて先読みを行うための手法を提案する．
- データ先読みを行うときは，先読みが外れたときのペナルティを考えて，ネットワーク性能を限界まで引き出せる最小の大きさで先読みを行うとよいが，その大きさは環境に応じて変化する．この大きさの自動調整

[1] 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan
[a)] mikity@eidos.ic.i.u-tokyo.ac.jp
[b)] tau@eidos.ic.i.u-tokyo.ac.jp

を，既存の TCP 上でシンプルに行う手法を提案する．これにより，信頼性があり世の中で広く用いられているプロトコルである TCP を用いつつ，先読み要求量を自動調整することができる．

以上の点に関して実装・評価を行い，提案手法により高遅延環境下での分散ファイルシステムを用いたアクセスを高速化できていることを確認した．

以降の章では，まず 2 章で本研究に関係して行われている関連研究に関して述べる．次に 3 章でファイルアクセスパターンの検出手法に関して説明し，4 章で先読み要求量の自動調整方法に関して述べる．その後 5 章で，提案した手法評価を行い，6 章でまとめと今後の課題に関して述べる．

2. 関連研究

本研究では分散ファイルシステムの高遅延広帯域環境におけるファイルアクセスを，適応的なデータ先読みによって高速化する手法を提案する．関連する研究ではこれまでに，遠隔ファイルアクセスを Non-blocking RPC を用いて高速化する手法が提案されている [2]．ただし，先読みを行う量を自動調整する手法は提案されておらず，ユーザが手動で調整を加えなければならない．

広域分散ファイルシステムには Gfarm [3]，HDFS [4] 等が広く用いられており，これらの分散ファイルシステムではレプリケーションを行い，レプリカの中で一番ネットワーク遅延が小さい場所に存在するレプリカにアクセスし，なるべく高遅延環境をまたいだファイルアクセスが起こらないようにしている．GPFS [5] では MPI-IO によるデータ先読みを行いファイルアクセススループットを向上させる研究が行われている [6]．また，PVFS [7] を用いて MPI プログラムのコードを解析し，ファイルアクセスに関する部分を別スレッドの処理へ切り出し，さらに事前に実行しておくことで先読みを実現する研究も行われている [8]．いずれの場合も，先読みは並列アプリケーションを変更することで行われ，ユーザが既存のアプリケーションの書き換えを行う必要がある．本研究では，既存のアプリケーションを書き換えることなく，TCP の標準的な用い方で得られる情報をもとに，プリフェッチサイズを自動決定する手法を提案する．

また，高遅延環境でのファイルデータ転送手法として，GridFTP [9], [10] の研究が行われている．しかしこれは FTP の機能をサポートすることを目的としており，ファイル全体の転送しか行わない．本研究ではファイルシステム API をサポートし，それに応じたデータ転送を行う．

3. アクセスパターン検出手法

広域分散ファイルシステムに限らず，ファイルシステムのデータ先読みのためにはデータアクセスパターンを検知し，今後必要とされるであろう部分を予測することが必要である．本章では提案するファイルアクセスパターンの検知手法に関して述べる．ファイルシステムにおける，先読みのためのアクセスパターン検知手法は様々なものが提案，実装されている．本稿では，一般的に広く採用されているような，ファイルアクセスの規則性に基づいた手法を適用する．

ファイルのアクセスパターンにより次のアクセスを予測するために，以前行われたファイルアクセスの規則性に注目する．世の中でファイルアクセスを行うアプリケーションには，ファイルをシーケンシャルにアクセスするアプリケーションが比較的多い．そのほかにも，一定の大きさのファイル読み込み，読み飛ばしを繰り返すアクセスでは，その規則性を検知して先読みを行うことができる．そこで本章では，ファイルアクセスの規則性に応じてアクセスパターン分類を行い，先でアクセスされそうな場所を予測する手法を提案する．

以下でははじめにアクセスパターンの分類と，ファイルシステムのレイヤで保持しておく情報（パラメータ）に関して述べる．次に，分類したファイルのアクセスパターンをファイルアクセス中に検出する手法を提案する．その後，そのアルゴリズムによるアクセスパターン検出を用いた先読み要求の流れに関して述べる．

3.1 アクセスパターンの分類と保持しておくべきパラメータ

本稿では規則性のあるアクセスパターンとして，シーケンシャルアクセスとストライドアクセスを取り上げ，その規則性に基づいて先読みを行う．ストライドアクセスとは，ファイルを飛ばし飛ばし読み込みを行うアクセスのことを指すが，ここでは簡単のため，一定の大きさ N byte のデータを読み込み，一定の大きさ M byte のデータを飛ばす動作を繰り返すアクセス（図 1）のこととする．これらの 2 つのアクセスパターンに合致しないタイプのアクセスを，規則性がなく予測が難しいアクセスであるランダムアクセスと分類することにする．

次に，これらのアクセスパターンに分類する際に，利用するためのパラメータに関して述べる．分散ファイルシステムのレイヤでアプリケーションがファイル読み込みを行う際に得られる情報は，各 read の offset，size のみである．提案手法では，これらと以下に述べるパラメータ群を利用

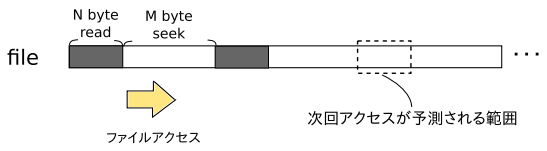

図 1　ストライドアクセス
Fig. 1　Stride access.

Algorithm 1 シーケンシャルアクセスモードのときのアクセスパターン検出アルゴリズム

```
if offset == last_read then
  mode = シーケンシャルアクセスモード
else if offset > last_read then
  mode = ストライドアクセスモード
  stride_read_size = last_read - last_seq_read_start
  stride_seek_size = offset - last_read
else
  mode = ランダムアクセスモード
end if
```

することによってアクセスパターンを検知・分類する．

mode　現在アプリケーションがどのようなファイルアクセスモードと判断されているか．シーケンシャルアクセスモード，ストライドアクセスモード，ランダムアクセスモードの 3 つの種類がある．

stride_read_size　ストライドアクセスのときに，シーケンシャルに読み込むと推定されるデータ量．

stride_seek_size　ストライドアクセスのときに，シーケンシャルな読み込みのあと，シークを行うデータ量．

last_seq_read_start　最後にシーケンシャルアクセスが始まったファイル内 offset．各 read で前回の read の続きを読んだ場合は更新されず，そうでない場合にその read の offset が入る．

last_read　1 つ前に呼び出された read で読んだデータ直後の offset．

これらの情報を保持し，利用することによりシーケンシャルアクセスモード，ストライドアクセスモード，ランダムアクセスモードの 3 つのアクセスモードを遷移ための判断が行える．具体的な遷移アルゴリズムに関しては次節で説明する．

3.2　各モード中のアクセスパターン検出アルゴリズムと先読み

次に，前節で述べたパラメータと，アプリケーションからの各 read の offset と size をもとにアクセスパターンのモード（パラメータの mode）を変更しつつ，そのモードに合わせた先読みを行うアルゴリズムに関して説明する．ファイルを open した際には，ファイルをシーケンシャルにアクセスするアプリケーションが比較的多いことからファイルアクセスモードはシーケンシャルモードで始める．また，last_seq_read_start，last_read の値は 0 に設定しておく．

3.2.1　シーケンシャルアクセスモード

現在の先読みモードがシーケンシャルアクセスモードである場合には，**Algorithm 1** に示したアルゴリズムに従い，今回の読み込み offset と last_read のみに着目してアクセスモードを推定する．このモードから他のモードに変わらない間は，現在アクセスしてる部分からシーケンシャルにそのままアクセスが続くことを予測し，シーケンシャルな先読み要求を行う．

Algorithm 2 ストライドアクセスモードのときのアクセスパターン検出アルゴリズム

```
if offset == last_read then
  # 読み込まれる N バイトのデータが分割されている
  # 場合も考慮した分岐
  if (offset + size) が次にアクセスが予測される部分を越える then
    mode = シーケンシャルアクセスモード
  else
    mode = （引き続き）ストライドアクセスモード
  end if
else if last_read, last_seq_read_start, offset の間隔がストライドアクセスパラメータにマッチ then
  mode = （引き続き）ストライドアクセスモード
else
  mode = ランダムアクセスモード
end if
```

3.2.2　ストライドアクセスモード

シーケンシャルアクセスモードに続いて，ストライドアクセスモードのときのアクセスパターン検知と先読みアルゴリズムを **Algorithm 2** に示す．先にも述べたとおり本提案手法でストライドアクセスとして扱うのは，一定の大きさ N byte のデータを読み込み，一定の大きさ M byte のデータを飛ばすアクセスのみである．現在の提案手法では，このアクセスパターンから少しでもずれてしまうとストライドアクセスとは判断しない．

ストライドアクセスモードのときは，このままのサイズの read と seek を繰り返した場合にアクセスすることになる部分を計算し，先読み要求を行う．このとき，パラメータ stride_read_size，stride_seek_size，last_seq_read_start を用いる．

3.2.3　ランダムアクセスモード

ランダムアクセスモードのときは各読み込みの offset が last_read と一致したとき，つまり前回の read の続きにアクセスした場合のみシーケンシャルアクセスモードに切り替える．またこのモードでは将来読み込まれる部分を予測することが困難であるため，先読み要求は行わない．

3.3　アクセスパターン検出と先読み要求の流れ

分散ファイルシステムでは多くの場合，ファイルはある大きさのブロック単位（これをブロックサイズと呼ぶ）で管理され，データ転送もこのブロックの単位で行われる．このブロックはアプリケーションから意識されることはない．ここでは提案手法におけるファイル読み込みの流れに

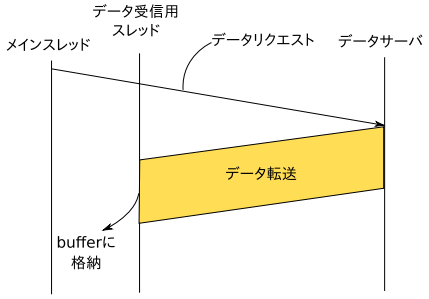

図 2　データ受信用スレッドとデータリクエスト
Fig. 2　Thread receiving data and data request.

関して述べる．まず，ファイルを開いたときに目的のファイルを持っているサーバ（データサーバとする）とコネクションを持つが，それと同時にデータの受信用スレッドを作成する．以降ファイルを読み込むときにはこのデータ受信用スレッドがデータを受け取り，バッファに格納する（**図 2**）．メインスレッドではアプリケーションからの要求を受け取り，アプリケーションに要求されたデータを返すまでの手順を分散ファイルシステムの read システムコールとして実装する．実装した read システムコールの大まかな流れを以下に示す．

（Step1）　read 要求の offset，size と保持しているパラメータを用いて，アクセスパターンを推定する（3.2 節）．

（Step2）　要求されたデータがバッファに存在しない場合のみデータサーバにリクエストを出し，受信用スレッドがデータを受信してバッファに格納するまで待つ．

（Step3）　Step1 で判断したアクセスパターンとパラメータにより次にアクセスされそうなブロックの集合を決定する．このときのブロック数の決定方法は 4 章で述べる．

（Step4）　Step3 で計算したブロックのうち，データサーバに要求をまだ出しておらず，バッファにも格納されていないブロックの転送要求をデータサーバに出す．

（Step5）　Step2 で得た要求データを用いて，read システムコールを返す．

4.　適応的なデータ先読み量自動調整手法

本章では，3 章で延べたアクセスパターンの分類・検出手法を用い，どの程度の量の先読み要求をデータサーバに出すかの調整方法に関して述べる．

4.1　理想的なデータ先読み要求量

まず本提案手法で理想とする先読みデータ要求量に関して説明する．ここでは研究の第一歩としてシンプルな状況を仮定することにし，データサーバ上でファイルデータがキャッシュに存在していることを前提として考える．この場合，遠隔ファイルのアクセススループットはディスクの読み込み帯域に律速されず，ネットワーク帯域に律速されることになる．分散ファイルシステムでネットワークの性能を極限まで引き出したデータ転送を行うためには，データサーバがあるデータ先読み要求を受け取ってから，次のデータ先読み要求を受け取るまでの間にデータをネットワークに転送し続ける必要がある．これをちょうど実現するためには（RTT × ネットワーク帯域幅），帯域遅延積分のデータ先読み要求を送信するようにすればよい．帯域遅延積より小さい量を先読み要求量にした場合には**図 3** の左側に示した図のようなデータ転送となり，ネットワークの性能を活かしきれない．逆に帯域遅延積以上のデータ先読み要求を出してしまうと，図 3 の右側に示した図のようなデータ転送となり，先読みが外れたときに前の先読み要求によるデータの転送後，実際にアプリケーションが要求したデータの転送が行われるため，先読みを行うことによるリスクが大きくなってしまう．したがって，データ先読み要求量を帯域遅延積とすることによって，ネットワーク帯域を十分に活かしつつ，先読みが外れたときのリスクは最小限とすることができる．この場合，先読みが外れたときの，次データ送信にかかる遅延時間は最悪で RTT となる（先読み要求を出した直後にアプリケーションが先読みから外れた場所をアクセスした場合）．

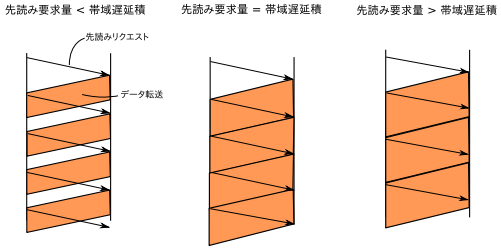

図 3　先読み要求量と帯域帯域遅延積の関係
Fig. 3　Relationship of prefetch request size and bandwidth-delay product.

4.2　先読み量自動調整手法

理想的なデータ先読み要求量を帯域遅延積としたうえで，分散ファイルシステムでの帯域遅延積の推定方法に関して述べる．簡単な方法としてユーザに帯域遅延積の値を設定として与えてもらうということが考えられるが，この方法ではユーザは，（ファイルシステムのクライアントの数 × ファイルサーバの数）通りの帯域遅延積を把握し，設定する必要がある．しかし，その設定は困難なうえにユーザに負担を強いることとなるため，可能ならば自動で推定し調整を行いたい．また，ネットワークの状態は刻一刻と変化するため，帯域遅延積もネットワークの状態に応じて変化

する．手動調整ではその変化に対応することはできず，その意味でも自動調整を動的に行うことは必要である．

本節では，標準的な TCP の用い方で得られる情報をもとに，シンプルに自動決定する手法を提案する．先読み要求量の初期値は分散ファイルシステムの実装ブロック 1 つ分の大きさに設定する．

帯域遅延積を推定するため，まず遅延（RTT）の測定をファイルを開いた際に行う．アプリケーションからあるファイルの open 要求を受け取ったのちに，データサーバに open 要求を送ってからその返答を受け取るまでの時間 t_1 を測定しておき，データサーバ側では open 要求を受け取ってから処理を行い返答を返すまでの時間 t_2 を測定する．そして t_2 を返答と共に送信し，ファイルシステム側で $t_1 - t_2$ を計算することで RTT を得る（**図 4**）．

次に，帯域推定の手法に関して述べる．データを受信し，帯域推定から，データ先読み要求量を帯域遅延積に設定するまでのアルゴリズムを **Algorithm 3** に示す．帯域推定は，データの受信用に作成したスレッドでデータを受信する際の時間計測により行う．ただし，データ受信時にすでにマシンの受信バッファにデータが到着しており，バッファからデータを読み込むだけになった場合は，データ受信にかかった時間から正確な帯域を計算することはできない．しかし，そのような場合は図 3 の左側に示したようなデータ転送の待ち時間は生じてないと判断できるため，ネットワーク帯域を十分に活かした通信が行われており，先読み要求量を増やす必要はない．この場合を検知するために，データ受信の前にそのソケットが読み込み可能であるかどうかをチェックし，すでに読み込み可能であった場合には，帯域推定やデータ先読み要求量の変更を行わないこととする．

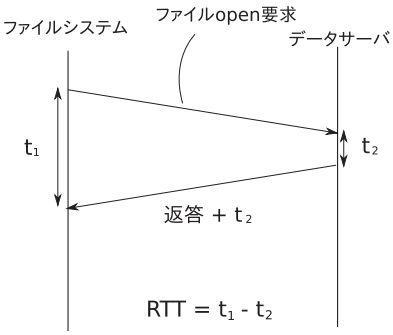

図 4　RTT の算出手順
Fig. 4　Procedure of RTT calculation.

Algorithm 3 データ受信用スレッドのアルゴリズム

```
while ファイルが開かれている do
  if socket からデータ読み出し可能 then
    帯域推定を行わないモードに変更
  else
    帯域推定を行うモードに変更
  end if
  データのヘッダを受信する
  start_time = now()
  ブロックデータを受信する
  end_time = now()
  if 帯域推定を行うモード then
    推定帯域 = 受信したデータサイズ / (end_time - start_time)
    先読み要求量を推定帯域 × 推定 RTT に設定
  end if
end while
```

5. 評価

本章では，ここまで述べた提案手法を 2 つの観点から評価する．まず，3 章で述べたアクセスパターン検知・分類手法を用いたファイルの先読みの効果を評価する．その後，4 章で述べた適応的に先読みデータ量を自動調節する手法の効果を評価する．本評価では分散ファイルシステムを FUSE [11] ベースで独自実装し，評価対象として用いる．この分散ファイルシステムのブロックサイズは 1 MB として実装を行っている．また，評価では 1 MB ごとのデータ読み込みにかかる時間を計測するが，分散ファイルシステムを用いたファイルアクセススループットは大きくばらつくため，直近の 5 MB のデータを（1 MB ずつ）読むのにかかった時間の平均（移動平均）を算出して評価結果に用いている．

5.1 評価環境

評価では，InTrigger プラットフォーム [12] 中の huscs（北海道大学），tsukuba（筑波大学），kyoto（京都大学）クラスタのマシンを用いる．それぞれのマシンの基本性能を**表 1** に，ネットワーク環境を**表 2** に示す．実装した分散ファイルシステム上にファイルを作成し（物理的には huscs 内に位置），tsukuba，kyoto からそれを読み込む．huscs → tsukuba，huscs → kyoto の流れでデータ転送が行われることになるが，この環境は表 2 に示したとおり高遅延の環境である．ファイルはデータサーバ上のキャッシュに存在する状態で行う．高遅延環境での通信をする際，Linux 標準の TCP 送受信バッファサイズで通信を行うと，バッファあふれによりパケットロスが起こり，スループットが十分に大きくならないため，各マシンの TCP 送信受信バッファを十分に大きくとって行う．その際には `sysctl` コマンドにより，`net.core.wmem_default`,`net.core.rmem_default` を 1 MB，`net.core.wmem_max`，`net.core.rmem_max` を 1 GB，`net.ipv4.tcp_mem`，`net.ipv4.tcp_rmem`，`net.ipv4.tcp_wmem` を “4 kB 1 MB 1 GB” に設定する．使用する TCP 輻輳制御アルゴリズムは Linux 標準の TCP-CUBIC である．

表 1　実験マシン基本性能
Table 1　Basic spec of machines for experiments.

ノード	CPU	メモリ	OS	NIC
huscs	Intel Xeon E5530 8cores (w/ HT 16cores)	24 GB	Linux 2.6.26（64 bit）	Broadcom NetXtreme II BCM57711
tsukuba	Intel Xeon E5620 8cores (w/ HT 16cores)			Intel(R) 10 Gigabit
kyoto	Intel Core2 6400 2cores	4 GB		Intel(R) 1 Gigabit

表 2　実験ネットワーク環境
Table 2　Network environment for experiments.

	物理帯域 [Gbps]	RTT [msec]
huscs - tsukuba 間	10	約 27
huscs - kyoto 間	1	約 38

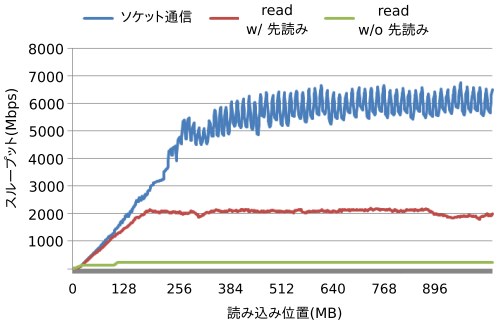

図 5　シーケンシャルアクセスの読み込みスループット（huscs→tsukuba）
Fig. 5　Read throughput in sequential access (huscs → tsukuba).

5.2　アクセスパターン検知による先読み効果の評価

5.2.1　シーケンシャル・ストライドアクセスの先読み効果の評価

提案アクセスパターン検知手法を用いて行う先読みの効果を測定するために，シーケンシャルアクセス，ストライドアクセスそれぞれに関して，ファイルの読み込みスループットを測定した．読み込むファイルは 1 GB の大きさのファイルである．シーケンシャルアクセスではファイルを 1 MB ずつ順番に読み込み，ストライドアクセスでは 1 MB 読み込み，3 MB 読み飛ばすという処理をファイルの末尾まで繰り返し行い，1 MB 読み込むごとのスループットを測定した．

評価は，実装した分散ファイルシステムの先読みを無効化したもの，先読みを有効にして手動設定により先読み要求量を決定したもの（huscs → tsukuba 環境では 25 MB，huscs → kyoto 環境では 10 MB に設定），単にソケット通信でデータを送り続ける簡易プログラムの比較で行う．ソケット通信のみのものは理論上可能な最大スループットを与えるためのものであり，先読みを無効化したものと，先読みを有効にして手動設定により先読み要求量を決定したものの比較により，先読み効果の評価を行う．この評価のシーケンシャルアクセスの結果を**図 5**，**図 6** に，ストライドアクセスの結果を**図 7**，**図 8** に示す．

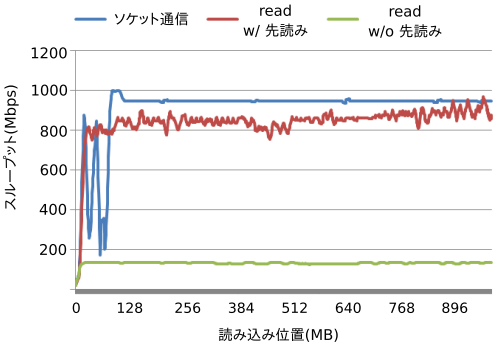

図 6　シーケンシャルアクセスの読み込みスループット（huscs→kyoto）
Fig. 6　Read throughput in sequential access (huscs → kyoto).

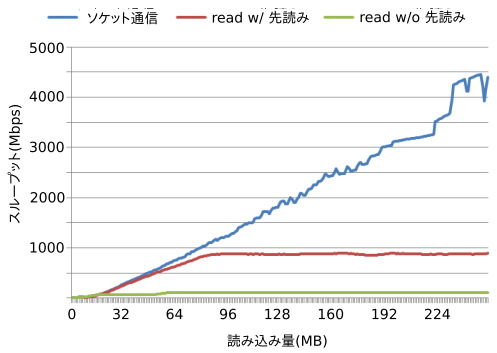

図 7　ストライドアクセスの読み込みスループット（huscs→tsukuba）
Fig. 7　Read throughput in stride access (huscs → tsukuba).

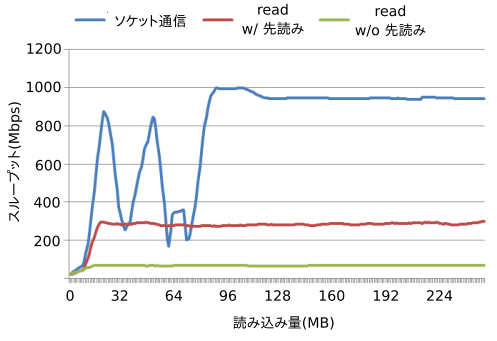

図 8　ストライドアクセスの読み込みスループット（huscs→kyoto）
Fig. 8　Read throughput in stride access (huscs → kyoto).

まずシーケンシャルアクセスに関して，先読みを行うことによってどちらの環境でも読み込み性能が大幅に改善されていることが分かる．しかし，図 6 ではソケット通信

の結果で示された理想スループットに近いアクセススループットが得られているが，図 5 ではこのスループットの 1/3 程度のスループットしか得られていない．

また，ストライドアクセスに関しては，先読みにより大幅にアクセス性能が向上することを確認できたが，シーケンシャルアクセスのときの読み込みスループットには達していない．うまくアクセス予測ができていればシーケンシャルアクセスとほぼ同じスループットが得られるはずである．これに関して調査を行った結果，これは FUSE の先読み機能が引き起こしていることが分かった．FUSE にはデフォルト設定で先読みの機能がついており，アプリケーションが要求したところの少し先までデータ要求が起こる．そのため，アプリケーションが 1 MB の大きさのデータ要求を行ったとき，その少し先までのデータ読み込みが行われる．評価用分散ファイルシステムは 1 MB の大きさのブロックでデータを転送・管理するため，結果的にアプリケーションが要求した 1 MB とその次の 1 MB の転送が行われることとなり，スループットが約半分になっていた．FUSE の先読み機能はオプションで無効化することが可能だが，FUSE の先読み機能には，FUSE モジュールによる内部通信，メモリコピー等の先読みを行えるという利点がある．これを無効にしてしまうと，いくらデータを先読みしてもスループットが大きく低下するため，簡単に改善することは難しい．

5.2.2　可変なアクセスパターン時のアクセスパターン検知手法の評価

次に，3 章で述べたアクセスパターン検出手法による先読みの精度を，可変パターンのアクセスを行い評価する．ここでは，前項と同様に 1 GB のファイルを読み込む．はじめに 128 MB シーケンシャルに読み込み，次の 128 MB を読み飛ばす．そしてその先の 128 MB で，1 MB の大きさ read，3 MB の大きさ seek を続けるストライドアクセスで読み込み，次の 128 MB を読み飛ばす．ここで 128 MB ずつ読み飛ばしているのは次のアクセスパターンに遷移する際に以前先読みを行ったところをアクセスしないためである．ここまでのアクセス全体を 2 回繰り返し，ファイルの先頭から末尾までアクセスした際の，1 MB ごとの read スループットを測定する．ここでは先読みサイズの調整を，提案手法による自動調整で行う．この実験を huscs → tsukuba 間で行った結果を**図 9** に示し，この実験中の提案手法による帯域推定と先読み要求量の推移を**図 10** に示す．

図 9 の赤丸で示した部分のみ，先読みされていないデータへのアクセスが起こっており，その部分のスループットは大きく低下している．しかし，それ以外の部分はうまくアクセスパターンの検出ができ，先読みが行えていることを示している．

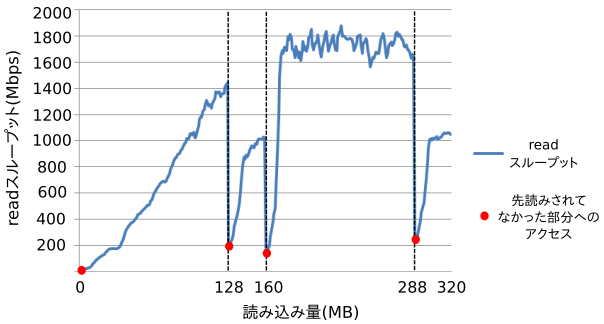

図 9　アクセスパターン可変時の読み込みスループット
Fig. 9　Read throughput in changing access pattern.

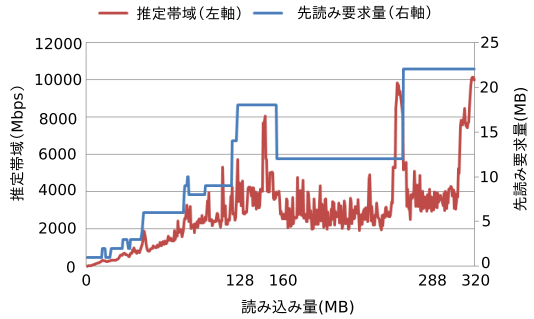

図 10　アクセスパターン可変時の推定帯域と先読み要求量
Fig. 10　Estimatded bandwidth and prefetch request size in changing access pattern.

5.3　適応的な先読みデータ要求量調整手法の評価

次に，本稿で提案する適応的な先読みデータ要求量を自動調整する手法に関して評価を行う．提案手法を用いて先読みデータ量を自動調整した際に，手動で適切とする値にした場合と変わらない読み込み性能を達成できるか，また先読みが外れたときのリスクは想定した範囲内かという 2 つの観点から評価を行う．

5.3.1　先読み要求量自動調整による効果

提案手法による先読み要求量の自動調整手法を評価する．評価は 5.2 節と同じく 1 GB のファイルを，シーケンシャルに読み込んだときの読み込みスループットを測定する．評価対象として，先読み要求量を手動で調整した場合と，自動調整を行った場合を比較する．また自動調整を行ったときのデータ受信用スレッドによる推定帯域と，そこから計算した先読み要求量も測定を行う．huscs → tsukuba で評価を行った結果を**図 11**，**図 12** に，huscs → kyoto で行った結果を**図 13**，**図 14** に示す．

どちらの環境でも提案手法による自動先読み量調整手法は，管理者が手動でパラメータを設定した場合と比べて，大きくアクセススループットを落とすようなことはなかった．ただし，huscs → tsukuba では先読み要求量が増加するのが比較的遅く，ファイルの先頭付近のアクセススループットが少し低い．つまり，高遅延環境で小さなファイルを読み込む場合は広帯域なネットワーク性能を活かした通信を行うところまで至らないことが分かる．したがって，

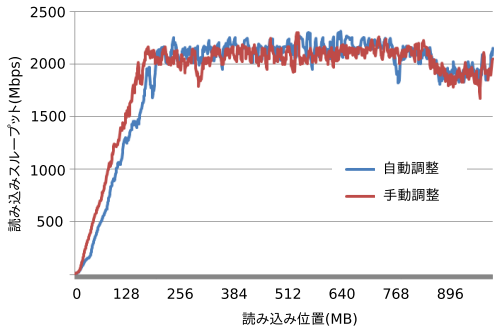

図 11　先読み要求量自動調整手法の評価（huscs→tsukuba）

Fig. 11　Evaluation of prefetch request size auto-tuning method (huscs → tsukuba).

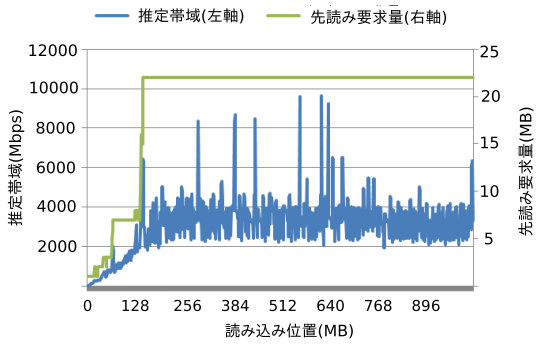

図 12　推定帯域と先読み要求量（huscs→tsukuba）

Fig. 12　Estimated bandwidth and prefetch request size (huscs → tsukuba).

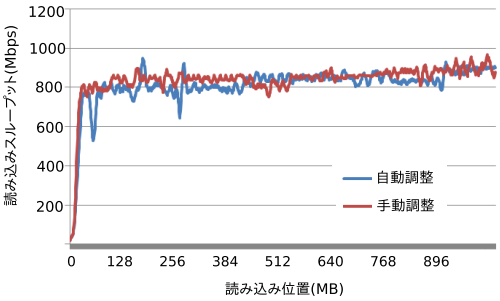

図 13　先読み要求量自動調整手法の評価（huscs→kyoto）

Fig. 13　Evaluation of prefetch request size auto-tuning method (huscs → kyoto).

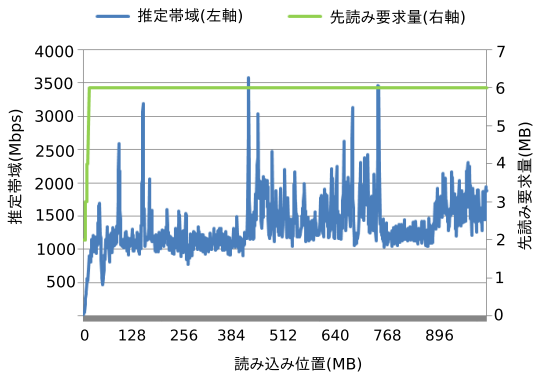

図 14　推定帯域と先読み要求量（huscs→kyoto）

Fig. 14　Estimated bandwidth and prefetch request size (huscs → kyoto).

ファイル先頭を読み込むときの帯域推定アルゴリズムに関してはもう少し改良の余地があると思える．

5.3.2　先読みモード切替え時の性能

次に先読みを行う際に生じるリスクに関して，提案手法を用いて先読み要求量を自動調整する場合と，先読みを無効化した場合を比較して評価する．先読みが外れたのちのデータ転送の遅延を測定するため，1 GB の大きさのファイルを用意し，はじめに 512 MB をシーケンシャルにアクセスし，その後ファイル末尾 1 MB にアクセスを行うときに読み込みにかかる時間を計測する（**図 15**）．実験を huscs → tsukuba で行った結果を**図 16**，huscs → kyoto で行った結果を**図 17** に示している．

まずどちらの環境においても提案手法を用いて先読みを行った場合のほうが，連続領域をアクセスしている途中で他の場所をアクセスしたときのアクセスにかかる時間が短くなっている．提案手法では最悪 RTT 分の遅延を想定していたが，実際は先読みを行っているときのほうがネットワーク帯域が出ている状態であるため，先読みを外して無駄なデータ転送が生じても，この実験環境では短い時間でデータ転送が行えている．

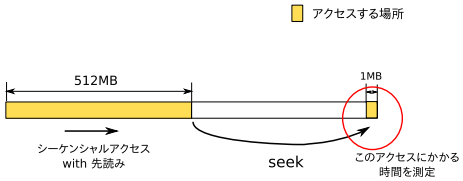

図 15　先読みモード切替えの性能評価

Fig. 15　Performance evaluation of switching prefetching mode.

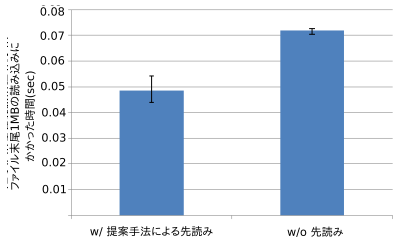

図 16　先読みモード切替え時のリスク評価（huscs→tsukuba）

Fig. 16　Evaluation of risk in switching prefetching mode (huscs → tsukuba).

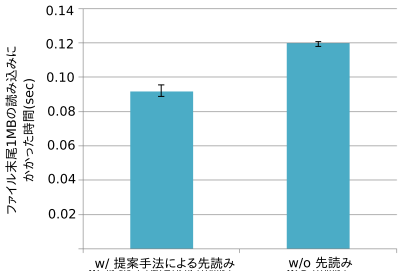

図 17　先読みモード切替え時のリスク評価（huscs→kyoto）

Fig. 17　Evaluation of risk in switching prefetching mode (huscs → kyoto).

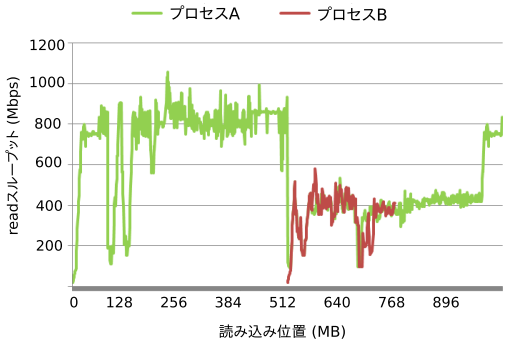

図 18　複数プロセスでの読み込みスループット
Fig. 18　Read throughput with multiple processes.

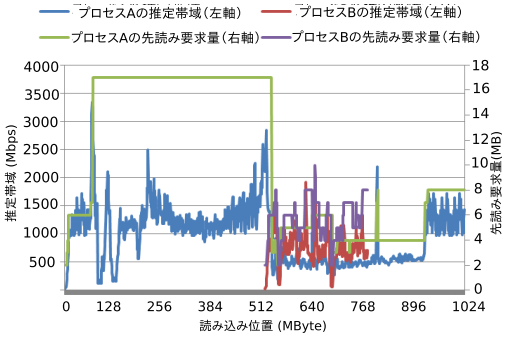

図 19　複数プロセスからの推定帯域と先読み要求量
Fig. 19　Estimated bandwidth and prefetch request size by multiple processes.

5.4　複数プロセスからの同時アクセス性能

最後に提案手法による適応的な先読みを，複数プロセスで実行した場合のファイル読み込みスループットを測定する．1 GB のファイルを用意し，あるプロセス（プロセス A とする）はシーケンシャルに 1 MB ずつファイルを読み込む．先頭 512 MB の読み込みが終わった直後にもう 1 つプロセスを立てて（プロセス B とする），プロセス A と同時に同じファイルを読み込む．プロセス B はファイルのオフセット 512 MB の位置から 256 MB を読み込んだのちに終了する．これらの読み込み 1 MB ごとに読み込み時間を測定し，そこからアクセススループットを算出したものを**図 18** に示す．また，この実験中の提案手法による帯域推定と先読み要求量の推移を**図 19** に示す．

プロセス A と B の間では同時アクセスが行われている間ほとんどスループットが同じ状態を遷移しており，限られた帯域を共有してデータ転送が行えていることが分かる．しかし，プロセス B が読み込みを終了してからプロセス A が元のスループットまで回復するまでにかなりの時間を要している．このことからもいったんスループットが落ちたあとに先読み量を増やしていく部分のアルゴリズムは改良の余地が残っていると思われる．

6.　おわりに

本稿では，分散環境での並列計算の基盤システムとして用いられる広域分散ファイルシステムを用いた高遅延環境でのアクセスでも，ネットワークの帯域を最大限に活かした読み込みスループットを得るための適応的な先読み手法を提案した．アプリケーションやユーザから特別な情報を得ることなくアクセスパターンを分類・検出し，先読みを適応的に低リスクで行う手法に関して述べ，評価を行った．その結果，提案手法を用いると，評価に用いた高遅延広帯域環境でのシーケンシャルアクセスの読み込みスループットを約 700～800%，ストライドアクセスの読み込みスループットを約 300～400%向上させることができた．

今回提案したアクセスパターン分類手法は実装やアルゴリズムの簡単化のため，ストライドアクセスに対して柔軟性に乏しく，少し予測と外れたところをアクセスするだけでまったく先読みを行わないという手法で，一定の決まった read，seek サイズでアクセスを行うアプリケーションには効果的だが，少し違ったアクセスをすると高遅延環境で読み込み性能が出ない．そのため今後アクセスパターンの検知アルゴリズムとして，もう少し柔軟性に富んだアルゴリズムを考案したい．また，今回の実験は国内のインターネット環境で実験を行っているためそのようなことはなかったが，RTT や利用可能帯域の変動する環境において本提案手法を適用する場合には少なくとも RTT を 1 回だけの計測にはせず，適度に測り直して調整する必要があると思われる．その他，本稿の提案手法は基本的にデータサーバにキャッシュが存在する場合に適した手法であるため，データサーバにキャッシュが存在せず，ディスクアクセスが発生する場合にも適した先読み手法を探ることも，今後の課題とする．

参考文献

[1] Montage: An astronomical image mosaic engine, available from ⟨http://montage.ipac.caltech.edu/⟩.
[2] 大辻弘貴，建部修見：Non-blocking rpc を用いた遠隔ファイルアクセスの実装と性能評価，研究報告ハイパフォーマンスコンピューティング（HPC），2011-HPC-132(16):1–5 (Nov. 2011).
[3] Tatebe, O., Hiraga, K. and Soda, N.: Gfarm grid file system, *New Generation Computing*, 28 (2010).
[4] Shvachko, K., Kuang, H., Radia, S. and Chansler, R.: The hadoop distributed file system, *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, pp.1–10 (May 2010).
[5] Schmuck, F. and Haskin, R.: Gpfs: A shared-disk file system for large computing clusters, *Proc. Conference on File and Storage Technologies*, pp.231–244 (Jan. 2002).
[6] Prost, J.-P., Treumann, R., Hedges, R., Jia, B. and Koniges, A.: Mpi-io/gpfs, an optimized implementation of mpi-io on top of gpfs, *SC Conference*, 0:58 (2001).
[7] Carns, P.H., Ligon, W.B. III, Ross, R.B. and Thakur, R.: Pvfs: A parallel file system for linux clusters, *Proc. 4th annual Linux Showcase & Conference*, Vol.4, p.28, Berkeley, CA, USA, USENIX Association (2000).
[8] Chen, Y., Byna, S., Sun, X.-H., Thakur, R. and Gropp,

W.: Hiding i/o latency with pre-execution prefetching for parallel applications, *Proc. 2008 ACM/IEEE conference on Supercomputing, SC '08*, pp.40:1–40:10, Piscataway, NJ, USA, IEEE Press (2008).
[9] Allcock, W.: GridFTP: Protocol extensions to FTP for the grid, *Global Grid Forum GFD-R-P.020* (2003).
[10] 伊藤建志，大崎博之，今瀬　眞：Gridftp-apt：データ転送プロトコル gridftp の並列 tcp コネクション数調整機構（インターネットの測定・性能評価技術及び一般），電子情報通信学会技術研究報告，情報ネットワーク，Vol.105, No.472, pp.19–24 (2005-12-08).
[11] Fuse, available from ⟨http://fuse.sourceforge.net/⟩.
[12] Intrigger platform, available from ⟨http://www.intrigger.jp/⟩.

堀内 美希　（学生会員）

1987 年生．2011 年東京大学工学部電子情報工学科卒業．同年より東京大学大学院情報理工学系研究科電子情報学専攻在学中．

田浦 健次朗　（正会員）

東京大学大学院情報理工学系研究科電子情報学専攻准教授．理学博士．1969 年生．1997 年東京大学大学院理学博士（情報科学専攻）．1996 年より東京大学大学院理学系研究科情報科学専攻助手．1999 年 7 月～2000 年 6 月米国カリフォルニア大学サンディエゴ校客員研究員．2001 年より東京大学大学院情報理工学系研究科電子情報学専攻講師．2002 年より現職．プログラム言語，並列・分散処理，システムソフトウェアに興味を持つ．ACM，IEEE，USENIX 等会員．