情報処理学会論文誌 Vol.53 No.2 724–731 (Feb. 2012)

分散ファイルシステムにおける ユーザコンテキストを利用したファイル配置アルゴリズム

岡田 耕司[1,a)] ロドニー バン ミーター[1] 村井 純[1]

受付日 2011年5月30日, 採録日 2011年11月7日

概要：仮想計算機環境の成熟にともない，ユーザは端末に依存することのないアプリケーション環境をネットワーク上に構築することが現実のものとなっている．それらのユーザ環境がネットワーク上に分散して配置されている際，既存のシステムではアプリケーション環境を構成する要素が，ユーザの利用場所と乖離して配置される．その乖離によりアクセス遅延が生じ，結果としてユーザの利用感を損なっている．アプリケーション環境の構成要素は，ユーザの享受するサービス，そしてサービスが利用する資源と定義することが可能である．本論文では，ユーザが接続ネットワークごとに求めるサービスに違いがあることに着目し，ネットワーク上にあるデータ配置の最適化を行うことを目的とする分散ファイルシステムの構築を行う．この目的のため，アプリケーション環境のうちユーザデータに焦点をあて，ユーザの行動様式を反映したデータの配置アルゴリズムを提案する．本論文で提案するアルゴリズムは，評価により他システムと比較しておよそ 80%のファイル転送量削減を実現した．

キーワード：広域分散ファイルシステム，プリフェッチ

File Placement Algorithm Utilizing User Context on Distributed File Systems

Kouji Okada[1,a)] Rodney Van Meter[1] Jun Murai[1]

Received: May 30, 2011, Accepted: November 7, 2011

Abstract: This paper presents an efficient data replication algorithm for distributed file systems. Virtual machine technologies support hardware-independent application environments on the global IP network allowing a user's "system" to migrate as the user moves from place to place. However, replication of the users' data to all possible destinations is expensive while on demand fetching gives unsatisfactory performance. We propose a semantic data placement method based on the human behaviors specific to the users' network locations. Our algorithm decreases file storage capacity notably while maintaining almost the same prediction accuracy as a more aggressive replication algorithm.

Keywords: global distributed file system, prefetch

1. はじめに

現在，ネットワーク上に高度に分散された計算機環境が構築されようとしている．近年，仮想計算機環境を実現するための技術が注目を集め，広く実運用されている．従来のコンピュータ環境において，ユーザにアプリケーション環境とユーザインタフェースを提供するための基盤システムである OS を仮想化する OS 仮想化技術の成熟により，ユーザは物理的ハードウェアに依存することなく，同一のアプリケーション環境を構築することが可能となった．また，iSCSI や iUSB など，計算機の物理構成要素であるパーツの物理バスとして IP ネットワークを利用することで，複数の OS から透過的に単一の物理パーツに接続，通信するための技術も提案，実装されている [1]．上記のように，次世代の計算機環境において，世界規模の広域分散ネット

[1] 慶応義塾大学
Keio University, Fujisawa, Kanagawa 252–0882, Japan
[a)] okada@sfc.wide.ad.jp

ワークであるインターネットを計算機内の共有バスとして用いることは，最も優先度の高い機能要件となっている．

一方で，クラウドコンピューティングに代表されるネットワーク型コンピュータの構造では，コンピュータを構成する資源を広大な IP ネットワークから透過的に参照するための仕組みが導入されている．ネットワーク型コンピュータにおけるユーザのアプリケーション環境を想定した際，アプリケーション環境はユーザが享受するサービスとデータから構成される．サービスとは，ユーザが計算機に期待する処理であり，OS 上ではプロセスとして抽象化されている．データは，ユーザサービスが扱うデータであり，OS 上ではファイルとして抽象化される．それらアプリケーション環境を構成する資源がネットワーク上に分散して配置された際，既存のコンピュータアーキテクチャでは問題とされなかった問題が発生する．

サービスはネットワーク上で高いレベルで抽象化され，ユーザはそれらのサービスを場所に依存することなく享受することが可能となった．さらに CDN などに代表されるオーバレイネットワーク技術の発達により，それらのサービス，そしてサービス基盤となる計算機資源も世界規模のネットワークに偏在することが予想される．本研究では，計算機資源が世界規模に分散したネットワーク型コンピュータにおけるアプリケーション環境のうち，ユーザデータに着目した問題点を整理し，その解決手法を示す．

2. 問題点

ネットワーク上にユーザデータを配置し，ユーザに透過的なユーザデータアクセスインタフェースを提供するシステムとして分散ファイルシステムがある．本研究では，ユーザデータストレージシステムとして分散ファイルシステムを前提とする．分散ファイルシステムを用いる際，ユーザがデータにアクセスする際のアクセス遅延はユーザサービス使用感に大きく影響する [2]．広域分散ファイルシステムにおいて，ユーザデータアクセス遅延を最小にするためには，利用可能性があるすべてのユーザデータをユーザが接続する可能性があるすべての拠点に複製するのが最も効率的である．しかし，すべての拠点に全ユーザデータを複製する場合，各拠点のファイルサーバ容量に大きな負荷をかけてしまうため，ストレージ容量の観点からはこのような手法は効率的とはいえない．

また，広域分散ファイルシステムにおいて，ファイルはファイルシステムの基盤として動作する IP ネットワークと同等の広がりを持って分散配置される．その際，ユーザのデータ利用場所と該当ユーザデータの格納場所のネットワーク距離が大きく離れていた場合，アクセス遅延が大きくなるとともに，データ配送時の経由リンク数も増えるため，通信の安定性が損なわれてしまう．通信の安定性の欠如は，フロースループットの不安定さ，ジッタとしてデータ転送フローに影響を及ぼす．したがって，データ転送時の経由リンク数を最小限とすることが必要となる．

本研究では，分散ファイルシステムにおいて，ユーザのネットワーク位置に着目し，場所に応じたユーザ行動に基づくデータ再配置手法として Data Preforwarding を提案する．本研究では，ユーザデータの管理に分散ファイルシステムを前提とするため，本論文ではこれ以降ユーザデータをファイルと表記する．

3. 関連研究

ファイルシステム上において，ファイル関連性を定義する際，ファイルのアクセス履歴を参照する手法が広く知られている．ユーザがあるファイルにアクセスする可能性は，直前にアクセスされたファイルから推測可能である [6], [7], [9]．たとえば，あるプログラムがいくつかの設定ファイルを読み込む際，その設定ファイル群を読み込む順序はプログラムによって固定的である．したがって，ファイルアクセス履歴に基づいて後にアクセスされると予想されるファイルを類推することは可能である．

上記のようなファイルシステム内の全イベントを包括的に参照するアクセスパターン予測手法は，分散ファイルシステムにおいては実効的でないとする研究も存在する [8]．これは，分散ファイルシステムでは，複数ユーザが各自に固有なパターンにより複数のプログラムを立ち上げてファイルにアクセスするため，単純にファイルシステム内のイベントを包括的に参照するだけではそれぞれのユーザに適したアクセス予測ができないためである．FARMER [12] は，分散ファイルシステムにおいて，ユーザ，プログラム，ファイルパスのそれぞれを参照してファイル関連性を定義し，ファイルアクセス履歴と統合することにより，プリフェッチ効率を高めることが可能であることを示している．また，Ellard らの研究 [10], [11] によると，NFS 環境において，ファイルアクセスモード（読み込み，書き込み，実行）やファイル名などのファイル属性により，ファイルのアクセスオペレーションを類推可能である，としている．さらに，類似ファイルをファイル属性から類推し，グループ分けすることにより関連ファイルを様々な粒度で定義可能である．

一方，ユーザのネットワーク上での振舞いを考えると，ユーザがネットワークへ接続する際，その接続形式には一定のパターンが存在することが示されている．Otiy [5] では，ワイヤレスメッシュネットワークにおける位置情報管理サーバへのノード移動情報更新を，ユーザ直近のサーバへ効率的に登録するための機構が提案されている．論文中では，ユーザのネットワーク移動を追跡した結果，ユーザには週ごとに一定の接続ネットワーク切替えパターンが存在することを示している．したがって，ユーザのネットワーク切断をイベントとし，直前複数週のアクセスパターンを

参照することにより，次接続ネットワークを判断することが可能となる．また，lcron [4] では，ユーザが接続ネットワークごとにアプリケーション挙動を変化させることに着目し，接続ネットワークごとにアプリケーション挙動を変化させるシステムを構築している．

本章では，ファイルシステムにおけるファイル関連性に関する先行研究と，ユーザのネットワークに依存した挙動についての関連研究を示した．本章で述べた関連ファイル予測手法は，本研究が想定するような超広域なファイル分散環境について考慮されておらず，Data Preforwarding に用いるには不十分である．一方，先行研究によるとユーザの接続ネットワーク予測は可能であり，また，接続ネットワークごとにアプリケーション挙動が変化することが示されている．本研究では，ユーザの接続ネットワークごとの挙動を意識したうえで，分散ファイルシステムにおける最適なファイル配置ならびにファイル転送を行うアルゴリズムの構築を行う．

4. ネットワークごとのユーザの振舞い

ユーザは，自身のファイルにアクセスする際，その現在場所に応じてその挙動を変化させる．たとえば，ユーザはオフィスにおいては，自身の職務に関連したファイルにアクセスする可能性が高い．一方，自宅においては，マルチメディアファイルなど，趣味性の高いファイルに対するアクセスが比較的に大きな割合を占める，という予想は直感的である．つまり，ユーザは場所に応じて求めるサービスに差異が存在し，それにともない，それらのサービスごとに利用するファイルに差異が発生する，と考えることができる．

ユーザにとって，場所ごとに求めるサービスに差異が存在するのであるならば，それにともなってサービスが求めるファイルにも差異が発生する．その際，ユーザが求めるサービスを，ユーザがその場所においてアクセスしたファイルの履歴から類推することが可能である．ユーザがファイルアクセスを行う際，そのファイルアクセスを予測するためには，そのユーザのファイルアクセス履歴を用いることが有効であることは 3 章ですでに述べた．場所ごとの必要サービスの差異を，ファイルアクセス履歴から導き出すために，本研究ではファイルアクセス履歴中のファイル拡張子の分布に着目する．場所ごとに求めるサービスに違いが存在するのであるならば，ファイルアクセス履歴中の拡張子の分布に偏りが見られるはずである．本研究では，場所ごとにファイルアクセス履歴を管理し，その拡張子の偏りに基づいたサービス特定を行う．

5. Data Preforwarding

Data Preforwarding では，すべてのファイルはネットワーク上に分散配置されたファイルサーバに格納されることを前提とする．そして，それらのファイルサーバは同一のストレージドメインネットワークに属することを想定する．この環境は，ユーザが接続する各アクセスネットワーク（**図 1** Local Network）上にファイルサーバを設置し，さらにそれらのファイルサーバ間をオーバレイネットワークを用いて接続することにより構築される．このオーバレイネットワークは，たとえば CDN のような，オーバレイネットワーク間で相互に秘匿性が確保されたネットワークを前提とする．ストレージドメインネットワークにおけるファイルの分散配置制御は Preforwarding Controller（PfC）によってなされる．ユーザクライアント PC は，アクセスネットワークに接続した際，ストレージドメインネットワークのエニキャストアドレスに対し認証要求を送信する（図 1：1)）．認証要求を受信したファイルサーバ（TFS: Tracker File Server）は，PfC に対して認証要求を転送すると同時に，自身の IP アドレスを付帯情報として PfC に送信する（図 1：2)）．PfC はユーザ認証情報を基にユーザを認証する．認証応答は TFS を経由しクライアント PC へと転送される（図 1：3)）．認証が成功した場合，クライアント PC はユーザが利用したファイル情報を，TFS を経由し，PfC へと定期的に送信する．同時に TFS はクライアント PC に対して Keepalive メッセージを送信することでユーザのネットワーク断を検知する．

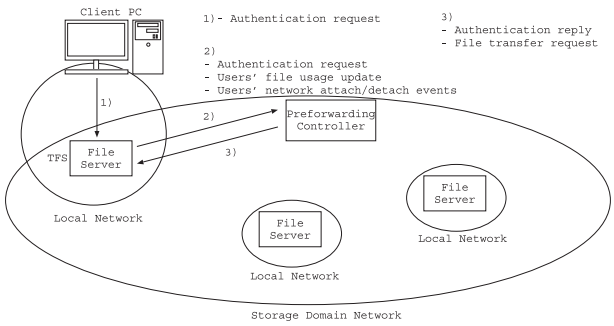


図 1 Data Preforwarding 概要図
Fig. 1 Data Prefowarding overview.

本論文で提案するファイル分散配置アルゴリズムである Data Preforwarding の機能要件は以下のとおりである．1) 場所に応じたユーザ行動の計測，2) 計測されたユーザ行動に基づく関連ファイル予測，3) 余剰転送の抑制，4) 場所に応じたファイル転送である．

Data Preforwarding では，ファイルを場所に応じて転送することが必要になる．そのためには，まず，ユーザ行動と場所を定義する必要がある．本研究では，「場所」をユーザ接続ネットワークとして定義する．ユーザの物理的移動にともない，ユーザ端末はその接続ネットワークを切り替える．その際，ファイルは，ユーザの物理的現在位置に近い場所に保存されるよりは，ユーザ端末の接続ネットワークにネットワーク位置として近い場所に保存されるべきである．上記の理由から本研究では，ユーザの「場所」を物

理位置ではなく，ユーザ端末の接続ネットワークとし，該当時間においてユーザ接続ネットワークに最も近いファイルサーバへとユーザファイルを転送することでファイルアクセス遅延を低減させる．ユーザの接続ネットワーク識別子として，本研究ではユーザ端末の最優先ネットワークインタフェースが接続するネットワークアドレスを用いる．

ユーザ接続ネットワーク情報を取得した後，Data Preforwarding は，ファイルシステムに対するユーザ行動を記録する．ファイルシステムに対するユーザの行動とは，ファイル読み込み，書き込み，作成などのファイルシステムイベントである．Data Preforwarding では，これらのファイルシステムイベントと場所情報としての接続ネットワーク情報を関連付け，ユーザがどの「場所」においてどのファイルを必要としているのかを把握する．本研究では，ある時点におけるユーザの「場所」とそれに応じたファイルアクセス履歴の対応付け情報を「イベント」と定義する．そして，イベント情報に基づいて，推測された必要ファイルを必要箇所にのみ複製/転送することにより，使用ストレージ容量を削減したうえでユーザのファイルアクセス遅延を解消することが可能となる．

Data Preforwarding では，イベントはユーザのネットワーク接続/接続断を基に作成される．その際に，PfC は各ユーザの各イベントごとに，接続ネットワーク情報，イベント開始時刻/終了時間，ファイルアクセス履歴，そして最近傍ファイルサーバを記録する．各イベントは次節で述べるイベントクラスタリングにより，グループ化されファイルアクセス予測を行う際の元情報として用いられる．

5.1 イベントクラスタリング

Data Preforwarding では，ユーザごとにイベントを管理し，イベントごとの相関を把握する．本研究では，イベント間の関連づけを行い，イベントグループ化を行う処理を「イベントクラスタリング」と定義する．イベントクラスタリングは，各ユーザごとに PfC により行われる．イベントクラスタとは，ユーザの作業種類に基づいて分類されたイベント群ならびにそれらの管理情報のことである．ここでいう作業種類は，ユーザが滞在する「場所」と「行動」で示される．「場所」は，イベントが発生した際にユーザ端末が接続されていたネットワークのネットワークアドレスのことである．前述のとおり，ユーザの作業内容はユーザの滞在場所によって大きく影響を受けるため，場所も作業種類を表す指標として用いる．「行動」とはユーザがイベント中に主にアクセスしたファイル拡張子の組合せである．

イベントクラスタリングを行う際，まず，各イベントの「場所」「行動」に関する情報をイベント管理情報として把握する必要がある．既述のとおり，場所はイベント中にユーザ端末が接続していたネットワークのネットワークアドレスにより示される．次に，イベント中にアクセスされたファイルアクセス履歴を分析し，総アクセスイベント数における各拡張子を持つファイルに対するアクセスの割合を計算する．イベントのファイルアクセス履歴中，アクセス割合が 10%を超える拡張子は，すべてイベントキー拡張子としてイベント管理情報に記録される．このイベントキー拡張子が該当イベント中にユーザがとった「行動」を示すことになる．そして，すでに登録されているイベントクラスタ管理情報とこれらのイベント管理情報を比較し，そのイベントをどのイベントクラスタに所属させるかを決定する．

イベントクラスタ管理情報は，そのイベントクラスタに属するイベント管理情報におけるネットワークアドレス，ならびに各イベントのキー拡張子の集合で示される．本研究では，イベントクラスタに属するすべてのイベントにおけるキー拡張子の集合のことをクラスタキー拡張子と定義する．あるイベントを新規にイベントクラスタに組み込む際，新規イベントのイベントキー拡張子と各イベントクラスタのクラスタキー拡張子を比較する．新規イベントのイベントキー拡張子のうち 5 割以上がクラスタキー拡張子として記録されているクラスタがあれば，新規イベントをそのクラスタに追加する．新規イベントのイベントキー拡張子のうち，イベントクラスタのクラスタキー拡張子に登録されていないものがあれば，新規にクラスタキー拡張子として記録される．イベントキー拡張子とクラスタキー拡張子の比較の結果，適切な所属クラスタが存在しない場合，新規イベントクラスタが作成され，イベントクラスタリストに追加される．その際のクラスタ管理情報は，ネットワーク場所を新規イベントのイベント発生ネットワークアドレス，クラスタキー拡張子を新規クラスタのイベントキー拡張子とし，所属イベントが追加対象イベントのみとする．

5.2 イベントクラスタ予測

イベントクラスタ予測は，予測対象日/時間に発生するイベントにおいて用いられるファイルを予測するために行われるものであり，ユーザごとに PfC によって行われる．ユーザ行動について考察すると，Otiy [5] に示されるとおり，ユーザは週単位で一定の行動パターンをとる可能性が高い．したがって，ある曜日のある時間におけるユーザ行動を把握することにより，次週の同曜日におけるユーザ行動を推測可能となる．イベントクラスタ予測は，このように，ある予測対象日があった際，前週の同曜日に発生したイベントを参照し，そのイベントがどのイベントクラスタに属するかを把握することで，予測対象日に必要とされるファイルを予測する．

しかし，ある曜日におけるファイルアクセス予測を行う際，直前週同一曜日の発生イベント情報のみに基づいてファイルアクセス予測を行う手法は十分ではない．なぜなら，1 週におけるユーザの作業種類についての時間割は，

	CID		CID		CID		CID		2010/11/12	CID	2010/11/13	CID	2010/11/14	CID
									15:58	9	1:32	10	22:10	11
								(*)	21:59	10	3:22	11		
2010/11/15	CID	2010/11/16	CID	2010/11/17	CID	2010/11/18	CID			CID	2010/11/20	CID	2010/11/21	CID
5:10	11	10:50	10	6:57	11	5:41	12				0:17	11	23:58	11
9:24	10	20:57	11	11:21	12	8:49	4				8:28	11		
12:50	12			12:10	10						16:22	12		
14:18	13										16:47	14		
15:35	10													
2010/11/22	CID	2010/11/23	CID	2010/11/24	CID	2010/11/25	CID		2010/11/26	CID	2010/11/27	CID	2010/11/28	CID
23:16	11	11:55	11	10:01	11	0:16	11		10:45	10	0:03	5	8:37	11
		23:33	11	11:40	12	11:55	12		15:48	9	3:08	11		
				18:05	10	12:27	10		21:58	12	11:39	11		
						12:47	9				14:47	12		
						16:18	10				15:15	10		
						18:55	12				23:32	11		
2010/11/29	CID	2010/11/30	CID	2010/12/01	CID	2010/12/02	CID		2010/12/03	CID	2010/12/04	CID	2010/12/05	CID
11:45	10	0:09	11	0:47	11	2:54	11		15:48	9	0:35	11	6:31	9
23:26	12	13:07	10	10:06	11	14:28	6		21:58	10	13:48	11	8:01	13
				12:15	10	22:48	10				23:38	10	9:18	11
				19:04	11								12:08	9
													20:43	9

図 2 ユーザイベントクラスタ例
Fig. 2 An example of user event clustering.

複数週をまたいである程度一定であることが予測されるとしても，その時間割内における作業はある一定期間をもって完了するためである．たとえば，あるユーザが月曜日日中，火曜日日中の 2 つの時間帯において，“主業務” を行うとする．そして，このユーザの主業務と分類される作業は主に見積り作成と提案資料作成であるとする．このユーザは週 1 火曜日日中においてある顧客を対象とした見積り作成を行っていた場合，直前週のアクセス履歴のみを用いて週 2 火曜日日中のファイルアクセス予測をする場合には，その顧客を対象とした見積り関連資料が多く予測されることになる．しかし，該当ユーザの週 2 における主業務が提案資料作成に移行していた際には，この予測は外れる．この際，直前別曜日に同一種類の業務を行う月曜日日中（同一のイベントクラスタイベントが予測された一日）のファイルアクセスも予測のソースとすることにより，提案資料に関連するファイルアクセスも予測対象とすることができるようになる．

Data Preforwarding では，イベントクラスタを用いたファイルアクセス予測を行う．**図 2** は，あるユーザのイベントクラスタリングの例である．計測は，1 週目木曜日から始まり，4 週目土曜日をもって終了している．図中 CID はイベントクラスタ識別番号（Cluster ID）を示し，各 CID 左隣はそのイベントの開始時間を示す．CID が 0 から開始されていないのは，イベントクラスタリングの過程でクラスタがマージされたためである．図 2 中で最も特徴的なのは，(*) で示した木曜日 16:00 直前から開始される CID:9 のイベント群である．これは，このユーザが毎週木曜日この時間に固定的なスケジュールを持っていることを示している．他曜日においても，同一週においては同一の CID に分類されるイベントが多く発生しているが，各イベントの開始時間はある程度以上分散している．

Data Preforwarding を行う際には，ある曜日における関連ファイルを類推する必要がある．そのためには，その曜日におけるイベントがどのイベントクラスタに属するものであるのかを予測する必要がある．**図 3** はイベントクラスタリングに基づいてイベント予測を行う際の手順である．

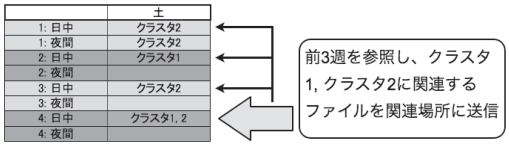


図 3 イベントクラスタ予測手法
Fig. 3 Event cluster prediction.

予測対象日と同一の曜日におけるイベントを参照し，直前 3 週の同一曜日におけるイベントを参照すると，第 1 週においてイベントクラスタ 2 に属するイベント，第 2 週においてクラスタ 1 に属するイベント，第 3 週においてクラスタ 2 に属するイベントが発生している．この際，第 4 週に発生するイベントとしてはクラスタ 1，クラスタ 2 の 2 つのイベントを推測可能である．Data Preforwarding では，この場合，2 つのイベント双方が発生するものとしてファイルの転送を行う．なぜならば，Data Preforwarding の最大目標は，ユーザのファイルアクセス遅延の最小化であるため，発生可能性が高いイベントすべてに対応してファイルを転送することが目標を満たす可能性が高いためである．

予測対象日における関連イベントが予測できた場合，関連イベントに属する直前 1 週分のアクセス履歴を参照し，同一のクラスタに属するイベントでアクセスされたファイル，別のクラスタに属するイベントで参照されたイベントキー拡張子を持つファイルがイベント関連場所に対して転送される．イベント予測は，基本的に前日（図 3 の場合は金曜日）までに完了させ，関連ファイルはイベント予測に従い関連場所にあらかじめ転送される．転送後に発生した類似イベントについては，イベント終了後（ユーザのネットワーク断時）にイベントクラスタ評価を行い，関連ファイルの更新/新規作成がある場合には他地点へ転送する．

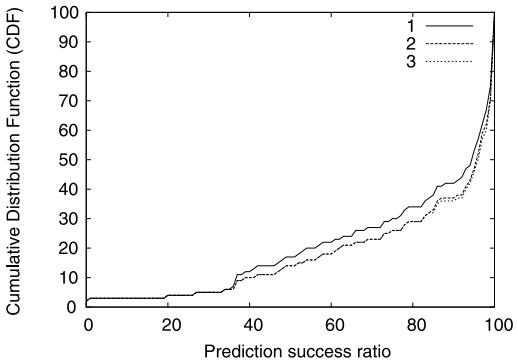


図 4 イベント参照週数（1，2，3）とファイルアクセス予測成功率
Fig. 4 Impact of varying number of weeks refering for file access prediction.

5.3 イベント参照週評価

既述のとおり，Data Preforwarding では，ある日時のファイルアクセスを参照する際，直前複数週の同一時間帯に発生したイベントからイベントクラスタを類推する．本節では，イベントクラスタ判定を行うために参照すべき週数を検討する．以下はあるユーザの 100 日間におけるファイルアクセス履歴に基づき，イベントクラスタ作成時に 1，2，3 週のイベント履歴を参照し，そのファイルアクセス予測成功率を示したものである．

図 4 の CDF で示されるとおり，1 週のみを参照し，イベントクラスタ予測を行ったうえでファイルアクセス予測を行う手法では，35%程度のファイルアクセス予測成功率が，2，3 週を参照するものに比べ大きな割合を占めている．その後，予測成功率 95%程度まで，1 週のみを参照するものと 2，3 週を参照するものとは差が開き続ける．次に 2 週を参照してアクセス予測を行う場合と，3 週を参照してアクセス予測を行う場合を比較する．ファイルアクセス予測率 85%の時点で 1%の差が認められるもののそれ以外のファイルアクセス予測率において同値をとり続け，予測成功率 100%の時点で 1%の差が吸収される結果となっている．したがって，ファイルアクセス予測を行う際，イベント予測のために 2 週参照する場合と 3 週参照する場合では，ファイルアクセス予測成功率に実効的な差はないといえる．一方，参照週が多くなればなるほど Data Preforwarding が行われる対象ファイルは多くなる可能性があるため，ファイル転送容量が大きくなる．以上の結果より，本研究では，イベント予測を行う場合の参照数デフォルト値として 2 週を選択する．

6. 評価

本章では，Data Preforwarding におけるアルゴリズムの評価を行う．本実験の目的は，ユーザ行動を把握し，転送量を制限する Data Preforwarding アルゴリズム（以下 DPA）が，妥当なファイルアクセス予測成功率を維持しつつ，ファイル転送サイズをいかに軽減できるか，を示すことである．評価では，DPA と，すべてのファイルサーバに対する複製アルゴリズム（STAR: Send To All Replication Algorithm）を比較する．評価項目はファイルアクセス予測率ならびに予測されたファイル群の合計サイズである．DPA，STAR 両手法とも本研究で提案するファイル複製手法である．STAR では，直近 1 週間においてユーザがアクセスしたすべてのファイルは，該当日にユーザが移動することが予測されるすべての地点に複製される．その際，すでにその拠点にファイルが存在する場合には転送は行われない．STAR が予測する「ユーザが移動する可能性がある拠点」は，直前 2 週同曜日のユーザ接続ネットワーク情報をもとに予測される．DPA は，ユーザ行動をイベントクラスタリングにより把握し，ファイルアクセス予測率を維持しつつ，ファイル転送量を低減する技術である．他関連技術との比較では，「場所に応じた行動パターン」を把握したことによるファイル転送抑制により，どの程度ファイル転送を効率化できたか，という評価を行うことができない．したがって，本論文では DPA と STAR の比較を行うことにより DPA の転送容量の効率を示す．

表 1 STAR と DPA の性能比較
Table 1 Performance comparison between STAR and DPA.

	STAR	DPA
ファイルアクセス予測成功率	73.561%	67.925%
累積ファイル転送サイズ（バイト）	1.36316E+14	2.31023E+13

本研究では，実際に 11 名の被験者の個人計算機上でファイルアクセス履歴，接続ネットワーク履歴を 1 カ月以上にわたって取得し，それらの履歴データを基にシミュレータ上に実装された STAR，DPA 各アルゴリズムの性能評価を行った．本実験に先立ち，ユーザには，個人 PC のホームディレクトリ以下のファイルに対するファイルアクセスとネットワーク接続情報を監視するデーモンを配布した．ファイルアクセス情報は長期にわたってホームディレクトリ外のログファイルに記録されている．また，デーモンはファイルアクセスが行われた際の，クライアント PC に設定されたネットワークアドレスも記録する．本評価では，シミュレーションの対象データとして，それぞれのユーザにおけるデータ取得開始から 6 週分のデータを選択した．そして，シミュレータでは，すべてのユーザ接続ネットワーク上に分散ファイルシステムを動作させたファイルサーバを設置した．シミュレーションにおける分散ファイルサーバ設置拠点数は 36 であり，各拠点はフラットに接続していることを想定する．被験者は慶応義塾大学，奈良先端科学技術大学院大学の学生，ならびに社会人で構成される．なお，抽出期間におけるファイルアクセスイベント総数は 920,257 である．また，実験において生成された 11 名全員のイベントクラスタ数の合計は 424 である．両アルゴリズムにおける性能評価の結果を**表 1** に示す．

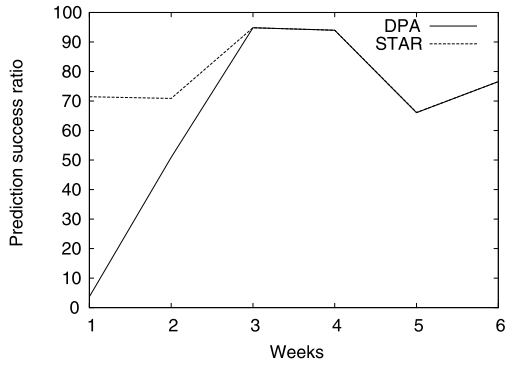


図 5 DPA/STAR におけるファイルアクセス予測成功率推移（1-6 週）

Fig. 5 Success ratio transition of access predictions by DPA/STAR (week 1-6).

まず，ファイル転送サイズを比較すると，DPA のファイル転送量は STAR の転送量のおよそ 17%にすぎない．しかし，両アルゴリズムの間には 6%の予測精度差が認められており，その要因を追求する必要がある．そのため，以下に STAR と DPA による 1 から 6 週目までのファイルアクセス予測率を比較する．

図 5 に 1 から 6 週目までの週単位でのファイルアクセス予測率の推移を示す．図 5 では，第 1 週目，第 2 週目において，DPA は STAR に比して著しく低いファイルアクセス予測成功率となっている．これは，STAR が直前発生イベントを用いて速やかにファイルアクセス予測を実現可能であるのに対し，DPA がユーザの場所ごとに必要とするファイル群を正確に認識しきれていないため予測が不完全なためである．学習が完了した 3 週目以降における予測成功率は STAR のものに類する結果となっている．これにより，上述の STAR と DPA 間の予測性能差は学習段階における差であることが分かる．初期段階における性能差は，システムを継続して利用し続けることにより解消可能であると考えられる．

Data Preforwarding では，上述のとおり，分散ファイルシステム上でのファイル転送を大きく削減することが可能となる．これは，4 章で述べたように，ユーザは場所（接続ネットワーク）に依存して求めるファイル種別が異なることが予想されるためである．この仮説を検証するため，**図 6** にユーザごとに構成されたイベントクラスタにおける，クラスタキー拡張子とクラスタ管理情報としての接続ネットワークの関係を示す．このグラフでは，11 名のユーザに対する全イベントクラスタを精査し，各ユーザのイベントクラスタキー拡張子が，接続ネットワークが異なるイベントクラスタのクラスタキー拡張子として登録されている個数の割合を示す．1 地点でのみ観測される場合は個数は 1 となる．

図 6 で示すとおり，61%のクラスタキー拡張子は，単一のネットワークに関するイベントクラスタにおいてのみクラスタキー拡張子として登録される．また，個別のユーザに着目し，ユーザが訪問する可能性があるすべてのネットワーク上に関するイベントクラスタにおいてクラスタキー拡張子として登録された拡張子数を分析した結果，該当する拡張子は全拡張子中 5.42%にすぎない．以上のことから，ファイルをユーザの全訪問ネットワークに複製するのは冗長であることが多く，拡張子で示されるファイル種別に応じて最適配置することが優位であることが分かる．Data Preforwarding はこのファイルの最適配置により，ファイル転送量を大幅に軽減することが可能である．

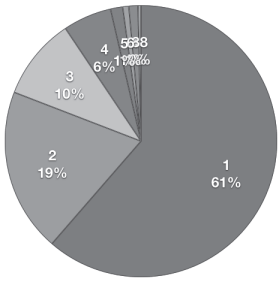


図 6 イベントクラスタにおけるキー拡張子とアクセス元ネットワーク数の分布

Fig. 6 The number of networks on which files with a key file extension were accessed (%).

7. 結論と課題

本論文では，分散ファイルシステムにおいて，ユーザの接続ネットワークにおける振舞いの違いに着目したファイル複製システム（Data Preforwarding）の提案を行った．Data Preforwarding では，ユーザが接続ネットワークに依存して必要とするファイルを，ファイルアクセス履歴におけるファイル拡張子の分布，場所，曜日の情報を基にイベントをクラスタ化し，ファイルの複製場所，転送を最小限におさえることが可能となる．

本論文では，シミュレータ上に実装した Data Preforwarding Algorithm を既存の完全複製アルゴリズムと比較し，アクセス予測成功率においてほぼ同一の結果でありながら，ストレージ使用率において 80%程度の削減を行えることをシミュレーションにより示した．また，DPA を用いることで，ユーザの振舞いを学習した後においては，ファイルアクセス予測率を飛躍的に向上させることができることが証明された．これにより，ユーザサービス使用感を維持しつつファイルサーバストレージ容量を削減することができた．

課題として，本論文で示した DPA ではファイルアクセス予測のアルゴリズムに関して単純にファイルアクセス履歴のみを参照している．したがって，アクセス履歴中に存在しないファイルに対するアクセス予測の向上に関して改善の余地が存在する．そのようなファイルに対するアクセス予測率の向上は，Ellard ら [10] や Xia ら [12] の研究に示

されるセマンティックを利用したファイルアクセス予測手法と DPA との融合により達成可能であると考えられる．

参考文献

[1] Van Meter, R., Finn, G.G. and Hotz, S.: VISA: Netstation's Virtual Internet SCSI Adapter, *ASPLOS VIII* (Oct. 1998).

[2] Riedel, E. and Gibson, G.: Understanding Customer Dissatisfaction with Underutilized Distributed File Servers, *Proc. 5th NASA Goddard Space Flight Center Conf. Mass Storage Systems and Technologies* (1996).

[3] Milojicic, D.S., Douglis, F., Paindaveine, Y., Wheeler, R. and Zhou, S.: Process migration, *ACM Computing Surveys* (*CSUR*), pp.241–299 (Sep. 2000).

[4] Heidemann, J. and Shah, D.: Location-aware scheduling with minimal infrastructure, *Proc. Annual Technical Conference on 2000 USENIX Annual Technical Conference*, San Diego, California, p.11 (June 2000).

[5] Boc, M., Fladenmuller, A. and de Amorim, M.D.: Otiy: Locators tracking nodes, *3rd CoNext 2007*, New York, NY, USA (Dec. 2007).

[6] Amer, A., Long, D.D.E., Paris, J.F. and Burns, R.: File access prediction with adjustable accuracy, *Proc. International Performance Conference on Computers and Communication* (*IPCCC'02*), Phoenix, AZ, USA (Apr. 2002).

[7] Kroeger, T.M. and Long, D.D.E.: The case for efficient file access pattern modeling, *Proc. 7th Workshop on Hot Topics in Operating Systems* (*HotOS-VII*), Rio Rico, Arizona, p.149, IEEE (Mar. 1999).

[8] Yeh, T., Long, D.D.E. Brandt, S.A.: Using program and user information to improve file prediction performance, *Proc. International Symposium on Performance Analysis of Systems and Software* (*ISPASS'01*), Tucson, AR, USA (Nov. 2001).

[9] Shah, P., Paris, J.F., Amer, A. and Long, D.D.E.: Identifying Stable File Access Patterns, *Proc. 21st IEEE / 12th NASA Goddard Conference on Mass Storage Systems and Technologies* (*MSST'04*), College Park, MD, USA (Apr. 2004).

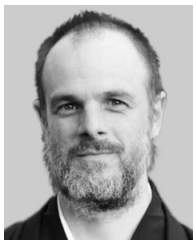
[10] Ellard, D., Mesnier, M., Thereska, E., Ganger, G.R. and Seltzer, M.: Attribute-Based Prediction of File Properties, Harvard Computer Science Group Technical Report TR-14-03 (Dec. 2003).

[11] Wang, F., Liao, C., Helian, N., Thompson, C., Wu, S., Deng, Y., Khare, V. and Parker, A.: accelerating linux/windows file systems by predicting access frequency, poster sessions of uk e-science all hands meeting (2007).

[12] Xia, P., Feng, D., Jiang, H., Tian, L. and Wang, F.: FARMER: A novel approach to file access correlation mining and evaluation reference model for optimizing peta-scale file system performance, *Proc. 17th International Symposium on High Performance Distributed Computing*, Boston, MA, USA (June 2008).

岡田 耕司

2003 年慶應義塾大学環境情報学部卒業．2005 年慶應義塾大学大学院政策・メディア研究科修了．ソフトバンク BB 株式会社，慶應義塾大学政策・メディア研究科特別研究員を経て，2007 年慶應義塾大学後期博士課程入学．移動体通信，ネットワークサイドコンピュータ，オーバレイネットワークの研究に従事．



ロドニー バン ミーター
（正会員）

1986 年カリフォルニア工科大学エンジニアリング･応用科学学士．1991 年南カリフォルニア大学コンピュータエンジニアリング修士．2006 年慶應義塾大学理工学博士．ストレージシステム，ネットワーク，ポスト・ムーアコンピュータアーキテクチャ，量子コンピュータの研究に従事．現在，慶應義塾大学環境情報学部准教授．

村井 純 （正会員）

1984 年慶應義塾大学大学院工学研究科後期博士課程修了．1984 年東京工業大学総合情報処理センター助手，1987 年東京大学大型計算機センター助手．1990 年慶應義塾大学環境情報学部助教授を経て，1997 年より同教授．1999～2005 年慶應義塾大学 SFC 研究所所長，2005～2009 年学校法人慶應義塾常任理事，2009 年より環境情報学部長．