ネットワーク分散型計算機環境における最適データ再配置手法

政策・メディア研究科
後期博士課程3年
岡田 耕司
okada@sfc.wide.ad.jp

研究概要

- 目的
  - 複数計算機環境における狭義のユーザデータ(ファイル)の効果的な配置手法
- 問題点
  - 広域分散ファイルシステムにおけるユーザデータ格納場所とユーザデータ利用場所との齟齬
- 提案手法
  - 分散ファイルシステムを前提とし、ユーザコンテキストに応じたユーザデータ転送手法の実現
  - Data Preforwarding

問題点

- ファイルアクセス遅延
  - “Understanding Customer Dissatisfaction With Underutilized Distributed File Servers”, Riedel, Erik Gibson, Garth, NASA-CP-3340-Vol- 2, 1996
  - 分散ファイルシステム環境ではファイル保存場所はユーザ利用場所とは異なる
- 「ネットワーク負荷」
  - 定義: スループット x 経由リンク数
  - 転送を行うので、それ自体では負荷の軽減を行われない
  - 負荷の効率化を行うことは可能
- 「ストレージ容量」
  - 全ユーザファイルを全拠点へ複製するのが最もアクセス遅延を低減可能
  - ストレージ容量は有限資源

先行研究と仮説

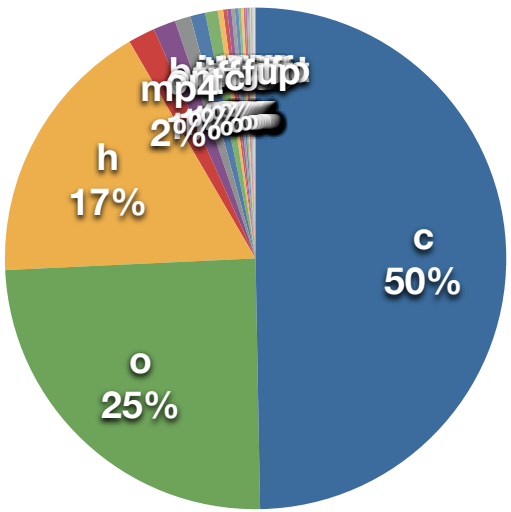
- ファイルアクセス予測
  - “Predicting File System Actions from Prior Events”
    - Thomas M. Kroeger and Darrell D. E. Long Usenix 1996
  - 単一ファイルシステム内のイベントの関係をトライ構造で示す
  - トライの中で示された優先度に基づいてprefetchすることで効率を向上
- ユーザ行動の把握と予想に基づいたサービス構築
  - “Otiy: Locators Tracking Nodes”, SIGCOMM 2007
  - ワイヤレスメッシュネットワークにおいて、ノードの位置情報を管理するlocator選択を、ユーザのネットワーク接続履歴に基づいて判断
  - ユーザ行動には週単位で行動パターンが存在

- 仮説: ユーザは「場所」毎に固有の予測可能なファイル利用パターンを持つ

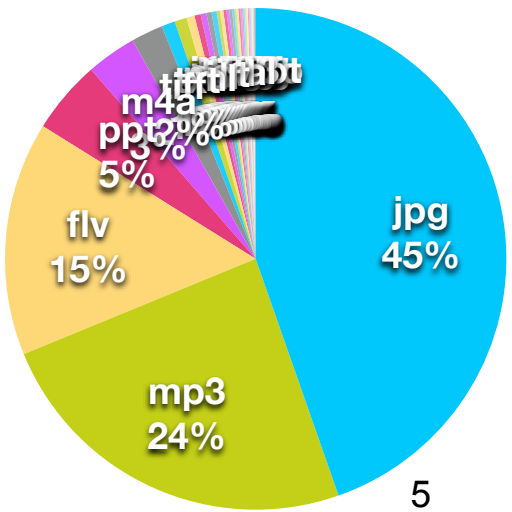
場所に応じた

ユーザデータアクセスパターン

- ユーザファイルアクセス傾向を拡張子に着目し検証
- ホームディレクトリ以下のファイルアクセス(open, modify, create)イベントを集計
- データ取得期間: 2009/09/28 - 2009/10/07



場所1



場所2

ユーザに対するデータの追従

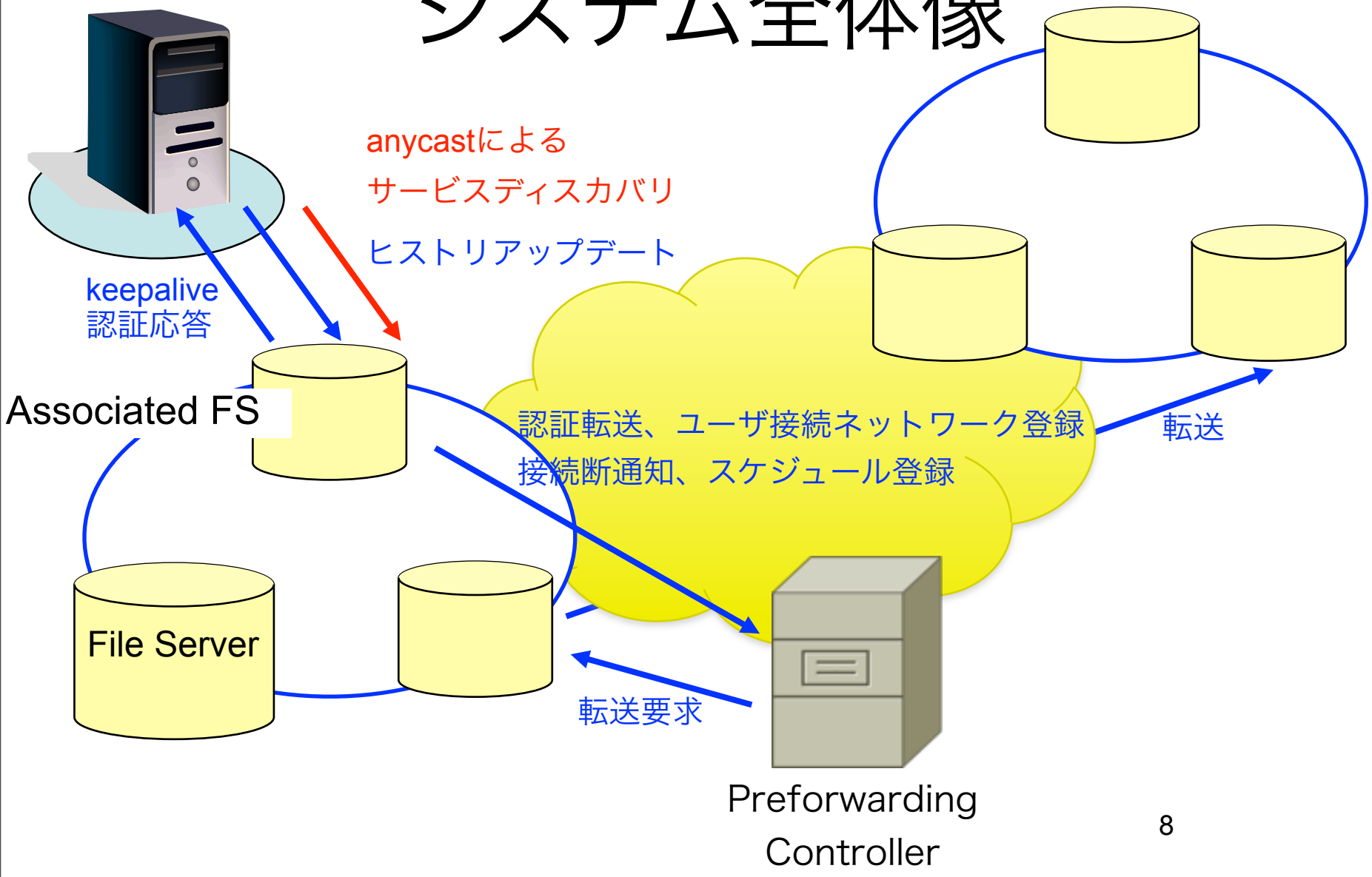
- Data Preforwarding -

- 目的
  - 分散ファイルシステムにおける”最適な”ファイル分散システムの構築
  - ユーザデータアクセス時の遅延を低減
  - データ転送によるネットワークへの負荷の軽減
  - ファイルサーバの使用資源を低減
- 手法
  - アクセス履歴に基づいた”場所”へのファイルの蓄積
  - ユーザの物理的移動に伴うデータ移動の実現
  - ユーザスケジュールに基づいて、ユーザのオフライン時にデータをあらかじめ必要場所に送信

Data Preforwarding手順

- ユーザデータ利用場所に基づくファイル配置最適化
  - ある場所において過去にアクセスされたファイルから未来にアクセスされる可能性の高いデータをあらかじめ転送
  - ユーザデータアクセス時の遅延を低減
  - ネットワーク負荷/ストレージ容量を考慮
- ユーザ移動に伴うファイル転送
  - データの「移動性」を実現

システム全体像


anycastによる
サービスディスカバリ
ヒストリアップデート
keepalive
認証応答
Associated FS
認証転送、ユーザ接続ネットワーク登録
接続断通知、スケジュール登録
転送
File Server
転送要求
Preforwarding
Controller

イベント作成

- 定例スケジュール
  - それぞれの情報は曜日単位で複数週管理
  - 場所情報: ユーザネットワークアドレス
  - 場所について最も近いファイルサーバ
  - ネットワーク接続/断時間
  - ファイルアクセス履歴
  - 該当時間において実現された転送速度
- 非定例スケジュール
  - アプリケーションからの入力により学習

イベントクラスタリング

- データ取得期間, 地点
  - 2009/09/29 - 2009/10109, 2地点
- 接続ネットワーク、アクセスファイル拡張子分布を基にクラスタリング
  - アクセスイベント中、10%を超えるアクセスがあったファイル拡張子の一致率を基にクラスタリング
- アクセスファイルからキーワードを抽出できる場合にはキーワードを優先

	日	月	火	水	木	金	土
1: 日中			クラスタ1	クラスタ1	クラスタ1		クラスタ2
1: 夜間			クラスタ2	クラスタ2	クラスタ2		クラスタ2
2: 日中	クラスタ2	クラスタ1	クラスタ1				クラスタ1
2: 夜間		クラスタ2	クラスタ2		クラスタ2	クラスタ2	
3: 日中	クラスタ3	クラスタ1		クラスタ1	クラスタ1		クラスタ2
3: 夜間				クラスタ2	クラスタ2	クラスタ2	
4: 日中	クラスタ2	クラスタ1					
4: 夜間		クラスタ2					

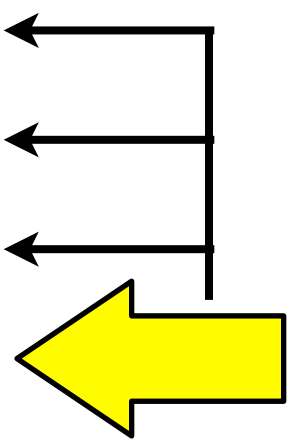
データ複製

- Data Preforwardingでは、複製データをユーザ移動に伴って転送
  - 該当時間における発生イベントを予測し、イベント関連ファイルを複製
  - 詳細は後述
- データ種別
  - Master File: 複数地点で利用されるファイルのオリジナルファイル
  - Replicated File: マスターファイルから複製されたファイル

イベントクラスタ予測

- 直前2週以上のイベントクラスタを参照し、該当イベントのイベントクラスタを決定
  - 下の例では、「4週目日中」のイベントとして、「クラスタ1」「クラスタ2」を予測
- 直前一週間において、同一クラスタと「予測された」イベントの関連ファイルを転送
  - アクセス上位ディレクトリ中、クラスタ関連拡張子ファイルを転送

	土
1: 日中	クラスタ2
1: 夜間	クラスタ2
2: 日中	クラスタ1
2: 夜間	
3: 日中	クラスタ2
3: 夜間	
4: 日中	クラスタ1, 2
4: 夜間	



前3週を参照し、クラスタ1, クラスタ2に関連するファイルを関連場所に送信

ファイル読み込み

- 単一地点から読み込まれるファイル
  - 一括読み込み
  - 順次読み込み
- 複数地点から読み込まれるファイル
  - 一括読み込み
  - 順次読み込み

単一地点から
読み込まれるファイル

- 一括読み込みファイル
  - 複製ファイルを利用予測地点へ転送
  - 実際にファイルにアクセスがあった場合にはマスターファイルを消去し、複製データをマスターファイルとして設定
  - アクセスがない場合は、複製データを消去
- 順次読み込みファイル
  - シーケンシャル読み込みデータについてバッファリング容量のみ複製
  - 確保可能な通信容量に基づいた容量評価
  - 過去に実現されたトラフィック量から推測可能
  - ユーザアクセスを基に順次転送開始
  - 完全転送が完了した段階でマスターファイル変更設定

通信予測手法関連研究

- 先行研究
  - “統計情報を利用したトラフィックバリエーションの見積もりに関する研究”, 原田義明, 岡村耕二, 信学技報
  - フローデータを地域毎に分析し、様々な粒度(AS単位、エッジネットワーク単位)での通信フロー変化を解析
  - 定常時のトラフィック傾向は安定しており、障害時においても障害の影響によるトラフィック変化は顕著とはいえない
- 手法
  - ネットワーク間の転送速度は統計的に学習可能

複数地点から
読み込まれるファイル

- 一括読み込みファイル
  - 概ね単一地点の場合と同様
  - マスターファイルがいずれの利用地点でもない場合に、マスターファイル設定変更
- 順次読み込みファイル
  - マスターファイルをそれぞれの利用地点の中間に配置
    - RTTによる計測
  - それぞれの利用場所に応じたバッファリング容量選択

書き込み

- ファイルに変更が発生した時点で、変更されたファイルをマスターファイルとし、他の複製ファイルを消去
  - ロック管理はPreforwarding Contorollerが行う
- 他の利用地点には、読み込みが行われる段階でPreforwardingの原理に基づきファイル転送

サーバクラスタリング

- RTTをベースとし、ファイルサーバをクラスタ化
- ファイル転送先優先度は、クライアントアクセスネットワークに最も近いファイルサーバが最高
- 最近傍ファイルサーバの容量が不足し、転送要求に応えられない場合、最近傍ファイルサーバと同一クラスタのファイルサーバへとデータ転送
- Preforwarding Controllerからのコントロールメッセージは、一度クラスタヘッドを介した上で各ファイルサーバに対して送信

評価方針

- 関連ファイル選択アルゴリズム検証
  - 複数人の場所毎におけるファイルアクセス履歴に基づいて検証
  - イベントクラスタの正当性、ファイルミスフェッチ率を評価
- システム評価
  - 分散ファイルシステム上にユーザデータを展開
  - 模倣環境によりアクセス遅延、ストレージ容量、ネットワーク負荷を評価

まとめ

- 分散ファイルシステムにおけるファイル移動性の実現
  - Data Preforwardingの提案
  - ユーザ行動の予測により、必要ファイルをあらかじめ必要場所へ転送
- ユーザ行動のプロファイリングにより必要ファイルを予測

今後の予定

- データ取得
  - 複数ユーザによる場所毎のファイルアクセスパターン取得中
- 評価
  - 得られたデータを基にアルゴリズム評価
  - 評価環境の整備
- 論文執筆
  - 投稿予定の決定