Title	ヘテロジニアスな環境における自律分散ファイルシステムに関する研究
Author(s)	渡辺，浩二
Citation	
Issue Date	2005-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1857
Rights	
Description	Supervisor:井口　寧，情報科学研究科，修士

ヘテロジニアスな環境における
自律分散ファイルシステムに関する研究

渡辺 浩二 (310125)

北陸先端科学技術大学院大学 情報科学研究科

2005 年 2 月 10 日

キーワード: グリッド Grid パリティ 分散ストレージ リードソロモン符号 GridFTP.

1 目的と背景

近年，高エネルギー物理やヒトゲノム解析等の大規模データ解析を必要とする分野ではグリッド技術がキーテクロノジーとなっている．グリッドコンピューティングでは多数のノードが接続されるため障害対策や効率的な資源の割り当てが必要不可欠である．また，グリッド環境は，さまざまなコンピュータから構成されているために，それぞれのコンピュータの信頼度は均一でないために，システム全体の信頼度を詳細に計算する必要がある．グリッド上では大量のデータが扱われるために，広域分散ファイルシステムを構築する手法が注目されている．データストレージ分野での先行研究では，データの保存，転送能力に主眼がおかれ，データの信頼性確保のための技術はレプリカ管理のみとなっている．レプリカ管理は冗長性の確保，負荷分散には有利になるが，ディスク容量の利用効率の面では不利になるといえる．

本研究ではグリッド上にデータの分散配置をするだけでなく，障害対策を盛り込んだ広域分散ファイルシステムについて研究を行う．先行研究のほとんどは，ターゲットとなるコンピュータがスーパーコンピューターなどの大型計算機であり，常時運転状態であることが前提で運用されていて信頼性も高い．それに対し，本研究で使用するコンピュータは一般ユーザが使用するコンピュータも含まれるために信頼性の面で不安定であり，個別に信頼性算出が必要となる．この信頼性に基づき，動的な分散配置を行い，データの配布先を変化させることで，信頼性が高く，容量に余裕のあるディスクだけにメンバを変更できる．よって信頼性を確保しつつ，容量の異なるディスクでも効率良く分散配置させることができる．これらの手法を用いて利用効率と障害対策を施した分散ファイルシステムについての構築，評価を行う．

2　提案システムの構成

システムの構成は，ファイルの分割や多重パリティ計算を行うマスタサーバ，データを格納するストレージノード，各ノードのディスク空き容量や空きメモリ状況を把握する資源管理サーバである．各ノードには GlobusToolkit2 (以下 globus) がインストールされている．マスタサーバでは，Perl と C 言語でプログラミングをした．資源管理サーバは，globus の MDS (Globus Metacomputing Directory Service) を使用し，各ノードの情報をまとめる．ストレージノードでは Globus の GridFTP サーバが待機しており，マスタサーバからのデータを待つ．分散配置されたデータの健全性は，確認プログラムを一定間隔で実行することで確認する．配布したファイル断片の中に一つでも異常があれば，一度データを再構築して正しいデータに復元したのち，再度エンコードを行い分散配置を行う．

3　システムの信頼性確保

データの分散配置では，データを分割し適切な多重度でパリティを生成する．このとき基準になるのはシステムの信頼度である．システムの信頼度は構成ノード個々の稼働率を求めた後，並列システムの m-out-of-n としてシステム全体の信頼度を算出する．M-out-of-n システムでは，個々の稼働率が異なる場合には故障ノードの全事象の組み合わせ計算を行う必要があるため，構成台数が多くなると計算の爆発が起きてしまう．そこで構成ノードを 3 つのクラスにわけ，各クラスに代表的な信頼度を割り当て，すべて同一の信頼度をみなして計算を行う．信頼度の高いクラスからノードを選択し使用する．データとパリティの多重度の割合は，要求されるシステム全体の信頼度を満たすように多重度を変化させ信頼度を確保する．

4　システムの性能

システムの性能評価では，マスタサーバ 1 台で処理を行う自己処理モード，負荷分散を考慮した外部処理モードを作成し実験を行った．予備実験として，ストライピングによる IO 性能の向上の確認と，リードソロモン符号のエンコーダー，デコーダーの基本性能の確認，NFS の性能確認を行った．ストライピングをして並列転送の実験をしたところ，最大で 740Mbps 程度の速度が得られた．これはほぼ Gigabit Ethernet の実行帯域をほぼ使い切っている性能といえる．Globus を使用してシステムを Grid へと展開した．ここで自己処理モードと外部処理モードの違いや，リードソロモン符号を扱う際のオーバーヘッドなどを測定し，評価した．リードソロモン符号を使用して多重パリティを生成処理にかかる時間は，元のファイルサイズに比例して大きくなる．しかしファイルサイズが小さいときは globus のジョブ終了検出に 30 秒程度の遅れが生じて，処理時間の短い場合のジョブ終了検出に時間がかかることがわかった．外部処理モードでデータの入出力を行った場合には，nfs でマウントされたディレクトリにファイルを置いておく必要があるために nfs

の性能がボトルネックとなり全体の処理時間が大幅に伸びる原因となった．パリティの計算時間自体には違いがないが，ファイルの書き出し，読み込み時間に差があったと考えられる．また，データの健全性を保つために，データが正しい状態で保存されているか，信頼性クラスが正しいかを判断して必要に応じてデータの再構築を行う機構について検討を行った．この実験では，復元されたデータが正しいかどうかを md5sum でチェックサムを計算し比較して確認した．

5　結論

本研究では，Grid 上への分散ファイルシステムを実現する際のシステム信頼性と，ディスク利用効率について解決方法提案と性能の評価を行った．Grid 構成ノードの稼働率を個々に調べることで，要求するシステム全体の信頼性を確保できることがわかった．また，多重パリティを使用することでレプリカ方式に比べてディスク使用効率の面で有利なことを示すことができた．たとえばストレージノードを 15 台使用し，パリティの多重度を 2 とした場合，最大で３７%のディスク容量節約することができた．最後に，今後の検討課題として，一時ファイルを利用せずに高速化を図る方式や，データが消失してしまうまでの時間 Mean Time To Data Loss (MTTDL) を求めて信頼性をあげる手段などに着いて挙げた．