Grid 環境における情報統合の実現

Implementation of Information Integration on the Grid environment

古澤 修　濱田 誠司　小椋 隆 †

日本アイ・ビー・エム株式会社 大和ソフトウェア開発研究所 ‡

Grid コンピューティングの実現のためには、分散データの有効利用を中心としたデータ Grid の技術が今後重要となってくる。データベースの概念を基本にした Grid 技術による情報統合を実現するために、IBM DB2 Information Integrator 製品上で、Globus Toolkit 2.2 を用いた情報統合に関して述べる。

1. はじめに

分散コンピューティング環境の普及に伴い Grid コンピューティングが注目されてきている。Grid は、VOs (Virtual Organizations)[1]の概念に基づいて、適切な認証により多様なリソースのダイナミックな共有を可能とするテクノロジーである。Grid 技術は、分散した CPU リソースの有効利用を中心としたプロセッシング Grid と、分散データの有効利用を中心としたデータ Grid に分類することができる。プロセッシング Grid 分野では、インターネットと PC を活用した SETI@home[2]などの研究利用をはじめ、従来技術であるクラスタリングをベースにした商用利用も進んでいる。一方、データ Grid の分野では、データ統合（共有）、レプリケーション、高速データ転送などといった研究が行われているが、データベースの利用を踏まえた事例、製品は数少ない。データベースを中心とした製品の一つとして、弊社の DB2 Information Integrator（以下 II)があげられる。II は分散された様々なデータに対して、あたかも単一のデータとしてアクセスできる機能を提供する。RDB による構造化されたデータに限らず、非構造化データに対しても Wrapper と呼ばれる処理コンポーネントを各データソースに用意することで SQL を介してアクセスを可能とする。ただし、現状では Grid に対応したデータソースへ直接アクセスを行うことはできない。しかしながら、Wrapper にそのための機能を補完すれば、データ Grid へ対応することが可能となる。

今回、データ Grid を用いたデータソースに対して情報統合を行うことを可能にする Wrapper の試作を行った。本論文ではその実装方法および評価、今後の適用方針に関して述べる。

2. 開発方針

2.1. Grid 構築ソフトウェアの選定

Grid 関連プロジェクトは、GGF(Global Grid Forum)[3]のような仕様策定のフォーラムから、ミドルウェアやアプリケーション開発プロジェクトまで幅広く存在する[4]。プロトタイプ作成のためのツールキットの選定にあたり次の点を考慮した。

- Grid を実現するインフラとして、既に広く使用されていること
- セキュリティ機能を備えていること
- データベースではデータの厳密性が要求されるため、提供されるコンポーネントの実績があること

この要件を満たすツールキットとして Globus Project[5]によって提供される Globus Toolkit 2.2 を選択した。Globus Toolkit は、もともと CPU リソースを共有して、膨大な計算を行うプロセッシング Grid の用途として開発されたものであり、現状ではデータ Grid 分野における機能が十分とはいえない。しかしながら、Grid 環境を構築するミドルウェアとしてデファクト・スタンダードであり、Globus へ対応することによって統一されたプロトコルやセキュリティ機能を利用することができるためその利点も大きい。さらに、ディレクトリサービスの利用も可能である。II、Wrapper、Globus Toolkit を使用した構成は図 1 のようになる。

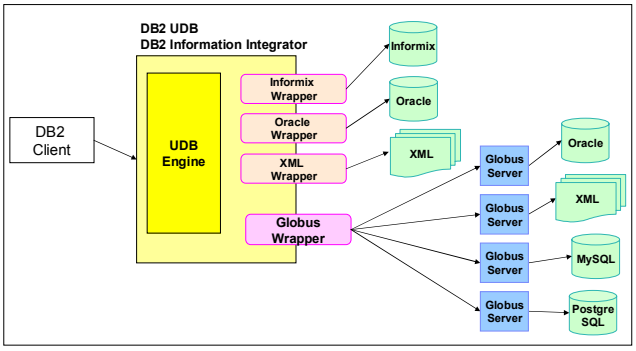

図 1: プロトタイプ・システム構成図

2.2. Globus Toolkit

Globus Toolkit は、Grid 環境構築のためのオープンな開発ツールである。Globus Toolkit 2.2 で提供される基本機能として以下の 4 つがあげられる。

- **GRAM (Grid Resource Allocation Management)**
  リモートのマシン上でジョブの実行、監視、制御などのリソース管理を行う
- **MDS (Monitoring and Discovery Service)**
  各 GRAM サーバー上のリソース情報を提供する LDAP ベースのディレクトリサービスを提供する
- **GridFTP**
  パラレル転送や第三者転送などの機能をもつ大量データ送受信用の FTP 拡張モジュールである
- **GSI (Grid Security Infrastructure)**
  SSL、公開鍵、x.509 証明書技術を元にシングルサインオンを可能とするセキュリティ機能を提供する

3. Wrapper への適用

Globus Toolkit は、統合されたアプリケーション・フレームワークといったものではない。サーバープロセスとしてサービスを行うモジュールは提供されているが、実体は機能単位で構成されたクラスライブラリである。そのため、Wrapper モジュール（以下 Globus Wrapper）への適用に際し、個々の基本機能に対する適用方法について順次述べていく。

3.1. GRAM の適用

GRAM はクライアントで指定したプログラムを GRAM サ

† Osamu Furusawa, Seiji Hamada, Takashi Ogura
‡ Yamato Software Development Laboratory, IBM Japan, Ltd.

ーバー上で実行し、その監視を行う。ただし、Globus はジョブを制御するための手段は提供しているが、クライアントサーバー間で送受信されるデータ形式に関しては規定していないため、それを独自に定義する必要があった。サーバー側で実行するプログラムは、標準入力として受け取った SQL 文をもとに検索を行い、その検索結果は標準出力に出力するという仕様とした。その仕様に沿って、GRAM クライアントである Globus Wrapper はデータへのアクセス処理を行うプログラムをリモート実行する。

3.2. MDS の適用

MDS を利用することで CPU、メモリー、ネットワーク負荷などの各データソースのマシン情報を動的に取得することができる。通常、II 上で Wrapper を使用する場合、DB2 とデータソースは静的に関連付けられるが、MDS を利用することにより、指定した条件に最適なデータソースを動的に切り替えることができる。この機能を実現するため、Globus Wrapper の内部処理においては、直接データソースのロケーションは指定せず、MDS サーバーのロケーションを登録し、MDS との連携よるデータソースの動的な選択を可能にした。今回の試作では、MDS がデフォルトで提供する CPU 負荷、メモリー使用量等の情報をもとに動的にデータソースを切り替える実装をおこなった。なお、MDS にはアプリケーション固有の情報を登録することもできるため、例えば、メタデータの利用によるデータ変換、QoS の確保、障害対応といった用途に使用することも可能となる。

3.3. GridFTP の適用

Globus Toolkit では Data Management 機能として大量のデータ転送用として GridFTP が提供されているが、適用に際し、オーバーヘッドによる処理時間の増加が想定された。大量データの転送が必要となるケースでは、SELECT 文および GRAM サーバー側での処理の最適化を行い、ネットワークを流れるデータの削減を図るアプローチのほうが望ましい。ゆえに、今回の試作では GridFTP の適用は行っていない。

3.4. GSI の適用

セキュリティ機構は、前述の 3 つの基本機能に対して利用されている。GSI はユーザー単位の認証であるため、DB2 上の認証ユーザーに対してマッピングを行えばよい。ただし、今回の試作では Wrapper 内での実装は行っていないため、DB2 のプロセス起動ユーザーの認証を GSI へ手動で行うことにより実現している。

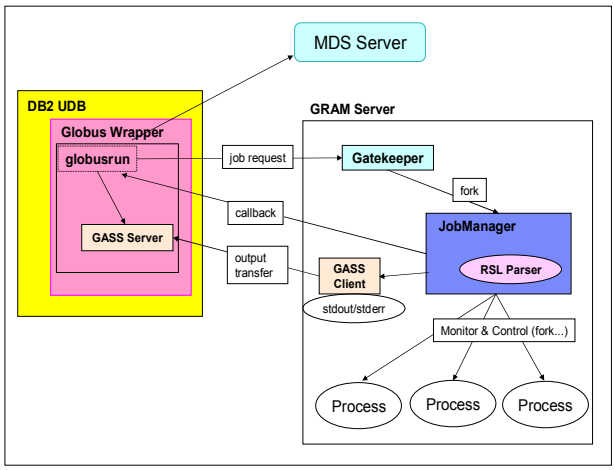

図 2: Globus Wrapper 構成図

適用結果による Globus Wrapper を用いた構成は図 2 のようになる。クライアントから要求があった場合、最初に MDS サーバーに問い合わせて、利用可能な GRAM サーバー名を取得する。続いて GRAM サーバーに対して SELECT 文を送りその結果を受け取る。

4. Globus Wrapper の評価

Globus Wrapper の適用例として 2 つのシナリオを実行してその有効性を評価した。

- 複数の GRAM サーバーに対しての情報統合 (Federated Search)
- CPU 負荷、メモリー使用量等のアクセスポリシーによって実際にアクセスを行うデータソースを動的に選択する

GRAM サーバー側には、専用プログラムを用意する必要はあるが、これは標準入出力に決められたデータ形式を扱う簡単なプログラムであり、GRAM サーバーに用意するだけで、Federated 検索の対象とすることができる。これにより、GRAM 対応のサーバーに対しては有効なデータ統合が行えた。

また、アクセスポリシーによる動的なデータソースの選択に関しては、負荷分散や障害対策などをはじめ、様々なソリューションに対して有効に利用できる仕組みであることが確認できた。

さらに、Wrapper として実装したことにより、UNION、JOIN といった SQL 文を用いた処理が、実績のある DB2 および II 本体側で行われるため、信頼性及びパフォーマンスの観点からも有効な手段であった。

5. まとめ

Wrapper 上で Globus をサポートすることにより、Globus に対応したデータソースに対する情報統合を可能とした。より汎用的なデータ Grid 対応を実現するためには多くの課題が残っているが、Grid 環境下において II を利用した情報統合を行うための方針の有効性は確認できた。

今後の方針として、Grid 対応したデータソース側との共通インターフェースのサポートが上げられる。例えば、OGSA (Open Grid Service s Architecture) の対応、および GGF のワーキンググループである DAIS (Database Access and Integration Services) によって策定されている仕様 [6]に対応していく必要がある。さらに、データ Grid においてインターネット上に分散した不特定のリソースに対してアクセスするためには、メタデータ管理、Semantic Web を実現するための技術要素であるオントロジーの利用などが必要となっていく。

参考文献

1) I. Foster, C. Kesselman, S.Tuecke “The Anatomy of the Grid” 2001 www.globus.org/research/papers/anatomy.pdf.
2) SETI@home, http://setiathome.ssl.berkeley.edu/index.html
3) Global Grid Forum, http://www.ggf.org/
4) M. Baker, R. Buyya, D. Faforenza “Grids and Grid Technologies for Wide-Area Distributed Computing”
5) The Globus Project, http://www.globus.org/
6) Grid Database Service Specification, 2003