ハイパフォーマンス
コンピューティング 92－8
（２００２．１０．２５）

# Quantum Chemistry Grid/Gaussian Portal Phase2

西川武志† 長嶋雲兵† 関口智嗣†

我々は、グリッド技術を用いて計算化学プログラム利用者のために Quantum Chemistry Grid/Gaussian Potal を開発している。Phase1 での基本的な機能に加え Phase2 ではセキュリティーポリシーの異なるネットワークゾーンでのシームレスな運用を可能とする機能を実装したことを報告する。これには、産総研グリッド研究センターで開発している仮想スーパーコンピュータセンタ利用環境（GridLib）のプロトタイプを用いた。

さらに、Phase2 では Gaussian の入力データをジョブ投入以前に解析し、アーカイブに蓄えられたすでに実行済みのデータとの高速完全一致類似検索機能を実装した。この機能により、すでに計算が実行してあればアーカイブ参照により計算機資源の節約が可能となり、さらにＣＰＵ時間やメモリ使用量を予想することが可能となった。これは、ジョブ内容に応じたスケジューリングを可能とする。類似検索機能は、平易な検索に重点を置いた高速検索と、より高度な一致度に注目した詳細検索の２つを用意した。

# Quantum Chemistry Grid/Gaussian Portal Phase2

Takeshi Nishikawa,† Umpei Nagashima† and Satoshi Sekiguchi†

Quantum Chemistry Grid/Gaussian Portal has been developed based on grid technology for easy and flexible use of Quantum Chemistry applications. We add a new function to realize seamless operation across network zones with different security policy by using virtual Super-Computing Center utility (GridLib) which have been developing by Grid Technology Research Center, AIST. A search function from the archives of computed results is also developed to equip to the portal in order to eliminate multiple execution of large scale computation. Two modes, quick and detailed, are available in this phase. A part of the function parsing Gaussian input file is also used to estimate requirement of CPU time, main memory space, disk space, job propertiy, and etc. Flexible job scheduling is possible using estimated properties.

## 1. はじめに

グリッド技術を用いると様々な計算資源に容易にアクセスすることが可能になる。これまでに Grid Application Service Provider（GASP）のフレームワークとして Quantum Chemistry Grid（QC Grid）を構築し、量子化学計算プログラム Gaussian[2] 利用者に計算資源について意識すること無く迅速に結果を得るためのポータルサービス、QC Grid/Gaussian Portal の設計と構築を行ない、新たな計算需要を掘り起こしたことを報告した[1]。

QC Grid/Gaussian Portal Phase 1 の設計基本方針は下記の通りである。

(1) グリッド技術を適用して計算資源の仮想化を行い、必要な計算サービスがいつでもどこからでも得られること。
(2) 高機能スケジューラにより計算内容に則した計算資源配分を行うこと。
(3) Web 技術を用いて簡単で使いやすいユーザインターフェースを構築すること。
(4) 多くの利用者を想定してスループットを重視すること。
(5) グループや組織単位の柔軟な課金のためアカウンティング機能をもつこと。
(6) コストパフォーマンスが良く、かつ安価なシステム構成であること。

QC Grid の利用者は、Web インターフェースを介して、物理的に分散して設置されている計算機上で公開運用されている応用プログラム（Gaussian）を、認証や計算機の機種の違いなどを考慮することなく、あたかも同一のシステムとして利用できる。QC Grid は、投入された入力データを解析し、すでに実行済みのジョブの結果が格納されているアーカイブを検索する。もし、入力データと類似のジョブの結果が有ればそれを利用者に報告する。報告された結果が利用者の望む結果である場合は、計算を実行しない。もちろん、

† 産業技術総合研究所 グリッド研究センター
Grid Technology Research Center, National Institute of Advanced Industrial Science and Technology(AIST)

結果の転送より実際に計算してしまう方が早く結果を得られるときは、実際の計算を行う。さらに QC Grid は、入力データを解析し、そのジョブの要求する CPU 時間やメモリ容量、ディスク領域の予測を行い、適切な計算機資源割り当てを行う。

これら、QCGrid の基本機能はすでに Phase1 で開発済みである。本報告では Phase2 として、分散資源の有効利用、使い勝手の向上を目的としたセキュリティ機能の拡充とメタスケジューラのアーカイブデータベース検索機能の改良について報告する。

2 章から Phase 2 に於ける機能強化点について述べる。2.1 節では GridLib について、2.2 節では高速検索について、2.3 節では詳細検索について、2.4 節では分子構造類似性検索および三次元構造表示について、2.5 節では遠隔地演算サーバの導入事例について述べ、3 章でまとめを述べる。

## 2. 機能強化点

Phase 2 では産総研グリッド研究センターにて開発の仮想スーパーコンピュータセンタ利用環境（GridLib）プロトタイプの導入とメタスケジューラによるデータベース検索機能の強化を行った。

GridLib プロトタイプの導入は、セキュリティポリシーの異なるゾーン毎での運用を可能とすることを目指し、QC Grid への計算機資源やアプリケーションの拡張とセキュリティ・アカウンテイング管理を容易にする。これを用いることで産総研つくば研究センターに設置された計算機とつくばから 300 km 離れた東北研究センター（仙台市）に置かれた計算機の計算資源をシームレスに利用することを確認した。GASP では計算資源がいくつかのサイトに物理的に分散化されているので、大量の負荷が発生しても、一つのサイトがピーク時の最大負荷に耐えられる設備を持つ必要はなく、GASP 全体として耐えられる設備を準備すればよい。何らかの事態で一部の計算資源の休止があったとしても GASP 全体としては運用が持続しており、可用性や信頼性が高くなる。

また、メタスケジューラのデータベース検索機能の強化はもれなく誤り無く高速な検索を可能とし、併せて、アーカイブに蓄積されたデータの保護を可能とした。アーカイブデータベースの検索では Gaussian 入力ファイルからキーワードを抽出してハッシュキーとし、キーワードの完全一致、分子構造は類似性の非常に高い結果を抽出する高速検索機能と、複数キーワードの指定や緩い分子構造の類似性により結果を抽出する詳細検索機能を実装した。分子構造の類似性は、入力データとアーカイブに保存されたデータの分子の主軸を重ね、構成原子の座標の自乗平均誤差を比較することで検出する。高速検索と詳細検索では自乗平均誤差のしきい値が異なる。このさらに利用者が最終的に

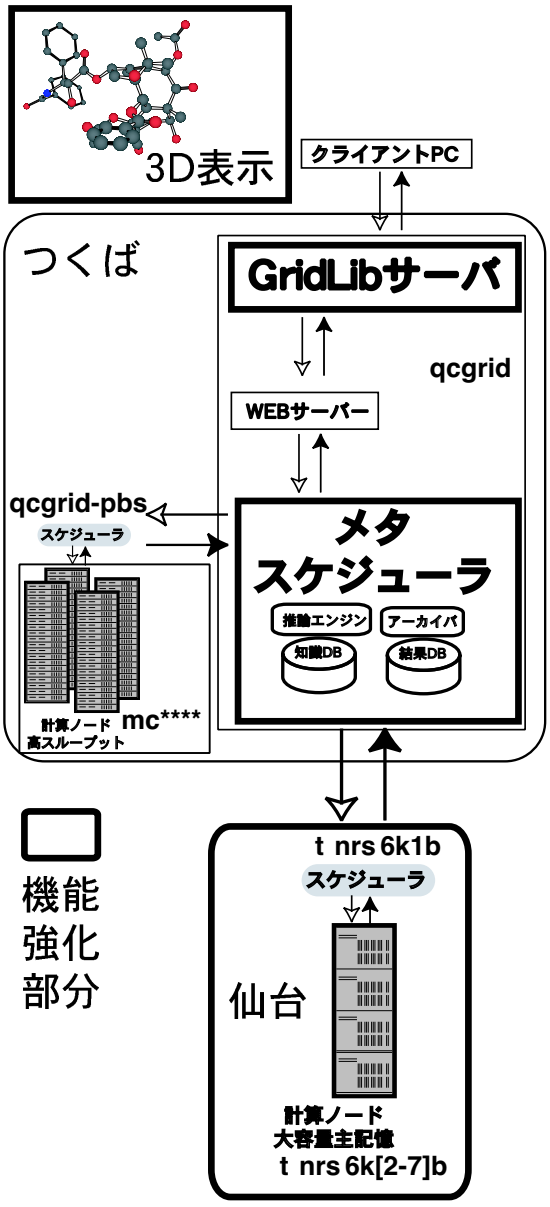


**図 1** Quantum Chemistry Grid/Gaussian Portal Phase 2 の強化点

確認することができるようアーカイブにあるデータの分子構造描画機能も加えた。

そのため、これら 2 つの機能強化に伴って必然的に付随する機能（Web インターフェース、入力ファイル評価機能、要求資源解析、スケジューラ、遠隔地計算資源等）が強化された（**図 1**）。

Quantum Chemistry Grid/Gaussian Portal Phase 2 のおおまかな処理の流れを**図 2** に示す。

### 2.1 GridLib

仮想スーパーコンピュータセンタ利用環境 GridLib は産総研グリッド研究センターで開発中の GASP を

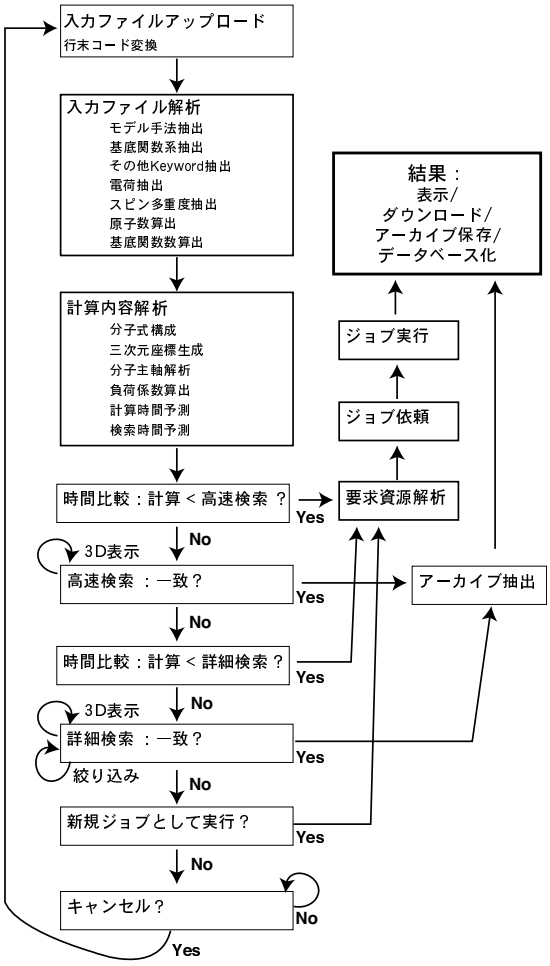


**図 2** Quantum Chemistry Grid/Gaussian Portal Phase 2 での処理の流れ

提供するフレームワークであり、

(1) 超大規模コンピューティング：複数の計算資源の計算能力を総合することで超高速かつ大規模メモリを必要とする計算の実行機能

(2) 高スループットコンピューティング：複数の計算に対して複数の計算資源を利用し結果を迅速に得る機能

(3) EasyComputing：スーパーコンピュータが具備するような高品質・高性能ライブラリやアプリケーションへの柔軟なアクセス手段の提供機能

これらの機能実現を目的として開発している。GridLib では**表 1** に掲げるような計算機遠隔利用における従来の問題点を解決しようとしている。

Phase 2 では GridLib をセキュリティ・アカウンティング管理機能の実現のために採用し、計算資源情報管理、アプリケーション情報管理、利用者情報管理、利用者認証に役立てる方針とした。**図 3** に QC Grid の最初の画面を示す。画面中央のリストには応用アプリケーションが列挙され、利用者は目的の応用アプリケーションをリストから選択し、利用可能状況を問い合わせるボタンを押す。利用可能なアプリケーションポータルへのリンクがリスト下部に列挙され、利用者目的のリンクを選択することで応用アプリケーションのサービスに入る。

**表 1** 従来の計算機遠隔利用における問題点
Table 1 The existent problems in the remote computing

| |
|---|
| ・機種、OS、ライブラリの違い |
| — 遠隔利用の共通インターフェースが無い |
| — ファイルや作業環境の共有ができない |
| — システム毎に於いて最適化の知識を要し作業が煩雑 |
| — 関数、ライブラリ毎のインターフェースが異なる |
| ・セキュリティ、アカウンティング、利用形態 |
| — センター毎に運用ポリシー、アカウントが異なる |
| — PCU 時間の提供の事実を隠蔽したい |
| — スケジューリング、負荷分散の簡便なインターフェース |
| ・ネットワークの効率的な利用 |
| — 通信量の最適化（不要データの削除、圧縮） |
| ・メンテナンスが煩雑（サーバ・クライアント型） |
| — サービス種が増えた際にサーバ・クライアント双方で管理業務発生 |

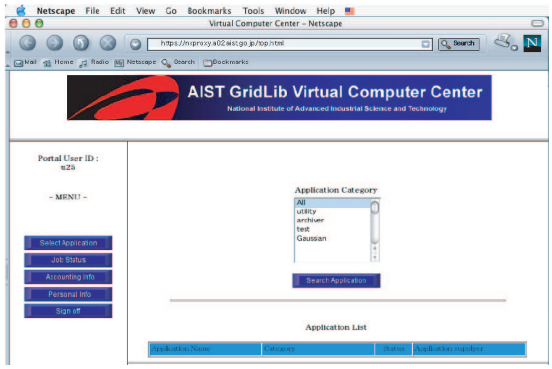


**図 3** Grid Lib: Virtual Computer Center

### 2.2 高速検索

高速検索では、分子式と計算手法と基底関数系を表す文字列からハッシュキーを作成し、分子式と計算手法、基底関数系が完全に一致するものを検索することとした。Phase 2 から、検索キーワードの内部表現法を変更し、速度は劣るが検索キーの表現を人が読み易いように、入力ファイルの分子定義を解析して得られた分子式と Gaussian がサポートする全計算手法と全基底関数系の文字列をハッシュキーとして直接用いることにした。高速検索では分子式と計算手法と基底関数系の三要素が完全に一致する結果のみが以下の項目を持つ表形式で出力される。

- No.
  検索結果のシリアル番号（1,2,3,...）
- Methods
  Gaussian での計算手法（HF,MP2,B3LYP,...）

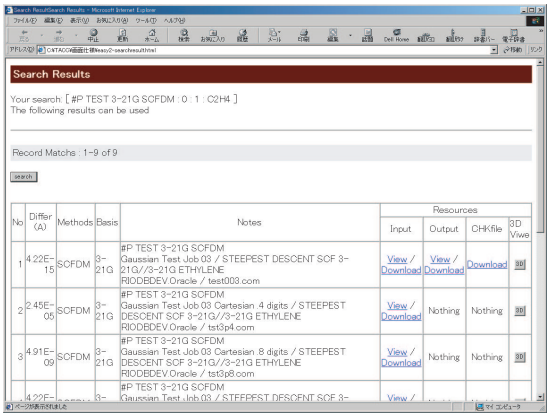


**図 4**　高速検索画面

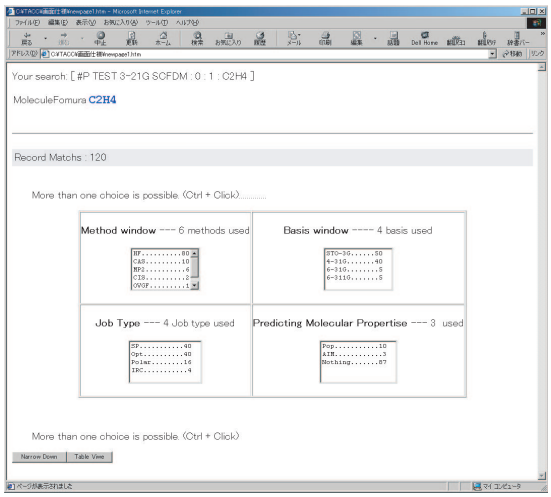


**図 5**　詳細検索画面

- Basis
  Gaussian での基底関数系（STO-3G,6-31G(d),cc-VQZ,. . . ）
- Notes
  備考（Gaussian のその他コマンドや入力ファイルのコメントの抽出）
- Input
  Gaussian 入力ファイルの表示/ダウンロードへのリンク
- Output
  Gaussian 出力ファイルの表示/ダウンロードへのリンク
- Gaussian Check point ファイルの有無
  存在すればファイルのダウンロードへのリンクを表示
- 3D View
  分子の三次元構造表示へのリンク

さらに検索結果ウィンドウには「詳細検索画面」に移行する「Detailed search」ボタン、検索結果を引用しないでアップロードした入力ファイルで新規ジョブとして実行する「Submit as a new job」ボタン、キャンセルしてメイン画面に戻る「Cancel & Return」ボタンが表示される（**図 4**）。

高速検索機能のキーワードの保持方法を変更した結果、高速検索するよりもジョブを投入して計算した結果を得る方が速いかどうかを判定する値がこれまでよりも大きくなったが、これまでの目標値 10 秒以下に収まった。例外的に小さな系でなければ Gaussian の計算で 10 秒以下で終了するものは無いのでほとんど高速検索を経ることとなる。

### 2.3　詳 細 検 索

詳細検索では、高速検索よりも条件を緩くして、最初に分子式が同一である結果が抽出され、次に列挙する各項目毎に複数の種類のレコードがさらにスクロールボックス中に列挙され、その件数が併記される。

- Methods
- Basis Sets
- Job Type
- Predicting Molecular Properties

それぞれの項目毎にレコードを一つ以上選択することが可能であり、選択した条件で絞り込むことが可能である。選択に際しては Web ブラウザのリストに於ける複数選択機能が利用可能である。複数条件を指定し詳細検索を実行した結果表示は高速検索と同じように、No.、Methods、Basis、Notes、Input、Output、Check point、3D View の各項目を持つ HTML テーブルとして返される（**図 5**）。

### 2.4　分子構造類似性検索および三次元構造表示

Phase 2 では分子の三次元構造を化学的性質を考慮してアーカイブデータベースを検索する機能と、利用者が最終的に同一分子かどうかを判断するために分子の三次元構造を分子構造表示プログラム RasMol[3] により画像ファイルとして表示する機能を強化した。Gaussian の入力ファイルでは原子の入力順は自由であるが、分子構造入力範囲のすべての原子を数え上げれば分子式☆ が判明する。しかしながら分子式が同じであっても異性体☆☆が存在し、さらに異性体には構造異性体☆☆☆、立体異性体☆4があり、立体異性体には幾

☆ 分子式：分子の原子組成と組成比を元素記号を用いて表したもの。物質の名称とともに分子量（1 モル）も表している。たとえば，水素→ $H_2$，酸素→ $O_2$，水→ $H_2O$，二酸化炭素→ $CO_2$，など。

☆☆ 異性体：分子式は同じであるが，構造が異なるため，化学的・物理的性質が異なる化合物。たとえば，エタノールとメチルエーテルの分子式はどちらも $C_2H_6O$ であるが，構造をしめす構造式（示性式）は，エタノール $C_2H_5OH$，メチルエーテル $CH_3OCH_3$ と異なっている。

☆☆☆ 構造異性体：原子結合の順、官能基の位置が異なる。

☆4 立体異性体：原子の三次元座標が異なる

何異性体☆と光学異性体☆☆がある。従ってアーカイブデータベースの検索に於いては異性体を区別し、同じ分子のみを検索抽出する必要がある。かつ外場が存在せず孤立して存在する光学異性体の性質には異性体間の差はないので一方の光学異性体の構造によってもう一方の光学異性体の結果が検索できた方が良い。以上の条件を満足するように分子構造類似性検索では分子の主軸を検出し入力データの三次元座標数値は異なるものの回転並進操作により主軸を一致させ類似性を比較する。検索結果は HTML テーブルで返され、主軸を一致させた後の座標間で算出した自乗平均誤差が表示され、分子の三次元構造表示のための「3D」ボタンも表示される。「3D」ボタンをクリックすると新規ウィンドウが開き、分子構造を示した画像ファイルが表示される。三次元構造表示ウィンドウには表示される分子の向きを変更するために x 軸、y 軸それぞれに 30 度おきで回転角度を指定できるプルダウンリストが存在し、「Rotation Viwe」ボタンを押すと、表示角度を変更して画像ファイルが生成され再表示されるようになっている（**図 6**）。このように利用者が分子の画像や座標間自乗平均誤差を比較して分子の類似性の最終的な判断を下す。

Web インターフェースで分子の三次元構造を表示し、回転させ様々な方向から眺めるための仕組みとして、分子の座標を Protein Data Bank(PDB)[4] フォーマット提供し Web ブラウザのプラグインや RasMol 等のヘルパーアプリケーションで表示することが考えられ、これらの方が様々な角度から眺める際の自由度、リアルタイム性等で優れている。しかしながら Phase 2 にて分子三次元構造の画像による提供の背景には、タンパク質等の生体高分子や対称性の高い分子等では、アーカイブデータベースに既存の計算結果が存在していなくとも、入力ファイルに定義されている分子の三次元座標定義自体に大きな価値が存在し、その価値の保護のため三次元座標そのものを外部に出したくないという要望があるからである。

### 2.5 遠隔地演算サーバの導入事例

GridLib の採用によりセキュリティポリシーの異なるゾーン毎での運用が可能となる仕組みを実装し、遠隔地の資源、今回は Phase 1 にて構築した IBM xSeries を中心とするシステムが設置してある茨城県つくば市から 300km 離れた宮城県仙台市に設置された IBM pSeries を Phase 2 での追加の計算資源として導入した。つくばと仙台に設置された機器の主要諸元は以下の通り。

- 産総研つくばセンター（茨城県つくば市）
  - Portal Node(qcgrid): x220, Pentium III 1.4GHz 1-Way, 512MB
  - Jobmanager Node(qcgrid-pbs): x342, Pentium III 1.4GHz 1-Way, 1GB
  - Computing Nodes(mc****): x330, Pentium III 1.2GHz 2-way SMP, 2GB,HDD 36GB 108 nodes
- 産総研東北センター（宮城県仙台市）
  - Jobmanager Node(tnrs6k1b): pSeries(7043-260), POWER3 200MHz 2-way SMP, memory 2GB, HDD 9.1GB
  - Computing Nodes(tnrs6k[2-7]b): pSeries(7043-260), POWER3 200MHz 2-way SMP, memory 2GB, HDD 9.1GB, 7nodes

Portal Node にはメタスケジューラ、Web サーバ、ファイルサーバ、データベースサーバ等の主要機能が稼働しており、Jobmanager Node では各ローカル計算資源のためのバッチジョブサーバが稼働し、Computing Nodes ではバッチジョブクライアントと Gaussian が稼働している。相互には日本テレコムが提供する広域専用 IP 網を用い 10Mbps で接続されている。

このように 300km 隔てて設置されている計算資源であるが Phase 2 の利用者からは計算資源間の物理的距離は全く認識されない。ジョブ制御ウィンドウを開いても現在実行中のキュー名が表示されるのみでシームレスに追加計算資源を導入することに成功した。

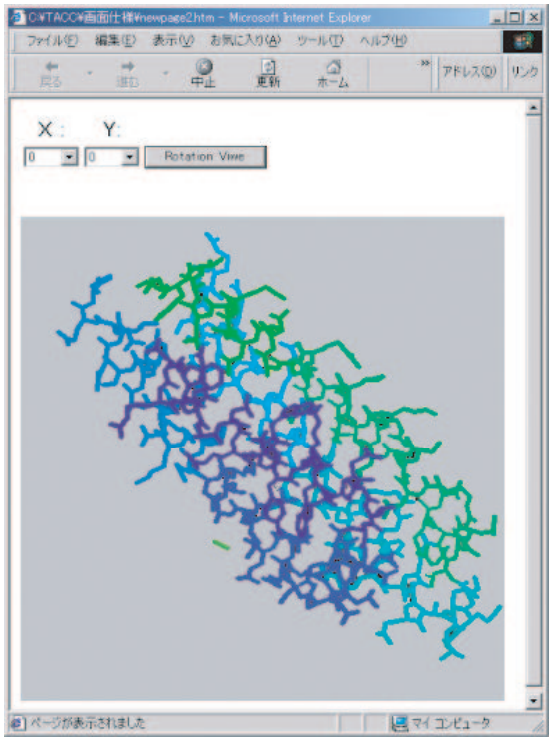


**図 6** 三次元構造表示ウィンドウ

☆ 幾何異性体：代表例は二重結合炭素へ結合する基の配置の違いでシス形、トランス形がある

☆☆ 光学異性体：代表例は 4 つの異なる基が結合する炭素原子（不斉炭素）による鏡像関係にあるもの。生命に関わる有機化合物はほとんどが光学異性体の一方のみからなる。

このように資源が分散化されているので、Gaussian Portal に対し大量の負荷が発生しても、一つのサイトがピーク時の最大負荷に耐えられる設備を持つ必要はなく、Gaussian Portal 全体として負荷に耐えられる計算資源を準備すればよい。また Portal Node までも多重化しそれぞれのサイトに設置すれば、定期メンテナンスや災害によりいずれかのサイトで計算資源の休止があったとしても Gaussian Portal 全体としては運用が持続しており、可用性や信頼性が高くなる。

## 3. ま と め

Quantum Chemistry Grid/Gaussian Portal Phase2 では、仮想スーパーコンピュータセンタ利用環境（GridLib）プロトタイプの導入とメタスケジューラのアーカイブデータベースの検索機能の改良に主眼を置き、グリッドインフラウェア部分には GridLib を採用し異なるセキュリティポリシーのゾーン毎での運用が可能となる仕組みを実装し遠隔地の資源をシームレスに導入することに成功した。アーカイブデータベースの検索では完全一致類似の結果を抽出する高速検索機能の強化と、同一分子式での結果を広範囲に抽出する詳細検索機能の強化を図った。さらに分子構造類似性検索では分子配向に依存せず、類似の分子構造を検索する機能を実現した。分子の三次元構造表示ではアーカイブデータベースに蓄積された分子の三次元座標を直接提示すること無く、すなわち知的財産を守りながら、さまざまな角度からの三次元構造表示と比較検討を可能にした。

今回の機能強化、異なるセキュリティポリシーでの運用を可能にする仕組みの採用、結果検索機能の強化、知的財産保護の仕組みによって産業界への QC Grid/Gaussian Portal の提供の第一歩を踏み出したと言える。

## 4. 今後の課題

QC Grid/Gaussian Portal は実際に運用することで計算結果を蓄積し、その結果を元に計算時間の予測機能に利用する。したがって今回、新機能や新しい計算資源を導入した結果、計算時間や資源配分の精度に経験不足から来る不備がある。実用に値するシステムとしての完成度を高めるには実際に運用して多種多様な Gaussian ジョブを処理し、結果をアーカイブデータベースに蓄積しなければならない。

加えて日進月歩で性能向上を遂げる計算資源に対応した計算時間予測、それに伴うジョブスケジューリングや、計算資源追加・削除が柔軟にできる資源の動的管理などを検討しなければならない。知的財産保護の観点からは、計算結果の公開、非公開の制御、非公開ならば期間や対象の制御の仕組みも検討しなければならない。

**謝辞** GridLib は科学技術振興事業団の計算科学技術活用型特定研究開発推進事業「仮想スーパーコンピュータセンタ利用環境 GridLib の構築」の一部である。GridLib 周りの設計、実装に関してアドバイスをいただいたグリッド研究センター建部修見博士ならびに田中良夫博士と SRA の平野基孝氏に感謝する。共同で開発作業を行った日本 IBM の寒川光博士をはじめとする開発チームにも感謝する。

## 参 考 文 献

1) 西川武志, 長嶋雲兵, 関口智嗣:Quantum Chemistry GRID/Gaussian Portal の構築, 情報処理学会研究報告, 2002-HPC-90-8,pp.43-48(2002)
2) Gaussian 98 (Revision A.9), M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. G. Johnson, W. Chen, M. W. Wong, J. L. Andres, M. Head-Gordon, E. S. Replogle and J. A. Pople, Gaussian, Inc., Pittsburgh PA, 1998.
Gaussian Inc.: http://www.gaussian.com/
3) Roger Sayle and E. James Milner-White, "RasMol: Biomolecular graphics for all", Trends in Biochemical Sciences (TIBS), Vol. 20, No. 9, p. 374.(1995,)
http://www.rasmol.org/
4) H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook and C. Zardecki, The Protein Data Bank, Acta Cryst. D58, pp.899-907,(2002).
Protein Data Bank: http://www.rcsb.org/pdb/