報道発表資料

2005 年 11 月 29 日
独立行政法人 理化学研究所

# タンパク質解析用高速処理計算機環境構築システムの完成

## - 専門的技術知識なしでもタンパク質解析が簡単に -

◇ポイント◇
- わずか 10 分で構築できる簡便なクラスター計算システム構築環境を実現
- “Knoppix for InterProScan4.1 High Throughput Computing Edition”（高速処理計算機用クノーピクス・インタープロスキャン）を開発
- バイオインフォマティクスやタンパク質解析等のライフサイエンス研究に不可欠な IT 環境を、専門家がいなくても構築可能になることで研究の効率化に貢献

独立行政法人理化学研究所（野依良治理事長）は、タンパク質の機能部位検索を高速に実行する環境を簡便に構築と撤収ができるポータブルな新しいシステムとして「タンパク質解析のための高速処理計算機環境構築システム」（Knoppix for InterProScan4.1 High Throughput Computing Edition:高速処理計算機用クノーピクス・インタープロスキャン）を開発しました。理研ゲノム科学総合研究センター（榊佳之センター長）、ゲノム情報先端技術研究グループ（小長谷明彦プロジェクトディレクター）の小西史一研究員らにより設計・構築されたものです。

開発した新システムは、高度なクラスター計算機構築技術を持たない人でも、簡便に複数の計算機を統合して働かせる環境の構築と撤収を行うことができます。従来は 32 台で構成されるクラスター計算機環境を構築に 6 時間以上かかっていたのが、専門でない人も約 10 分で構築が可能になりました。同研究グループはこのシステムの検証として、実験的にタンパク質機能部位検索（InterProScan※14.1:インタープロスキャン 4.1）に対して実施、性能を確認しました。今後はこの技術を他のアプリケーションやデータベースに対して広げることにより、一般的な利用者にも、高度な計算機利用による効果が得られることが期待できます。

開発したシステムは、ウェブサイト（http://big.gsc.riken.jp/index_html/Members/fumikazu/htc）にて、12 月 1 日から CD-ROM ISO イメージ※2 のダウンロードが利用可能となります。

本成果は、11 月 29 日（火）の ComSys2005（情報処理学会）で発表されます。

### 1. Knoppix for InterProScan4.1 High Throughput Computing Edition とは

Knoppix※3（クノーピクス）は、従来のようにハードディスクにインストールすることなく CD だけで利用することができる Linux※4（リナックス）ディストリブーション※5 です。このシステムを利用した後、計算機を再起動することで、元の計算機の状態に簡便に戻すことができるため、一過性の作業などにその用途は広がっています。

このような特徴のシステムを、European Bioinformatics Institute（EBI）※6 が開発した「InterProScan（インタープロスキャン）」という「タンパク質の機能

ドメイン検索を行うアプリケーション」に適用できるようにしました。また、検索を高速に実施するユーザのことを考慮して、複数のネットワーク機器（ノード）に分散処理させるアプリケーションである「Condor（コンドル）※7」などの作業管理機能を盛り込み、高速実行環境を簡便な手順で構築することができるシステムとして開発しました。

今回のシステムは、従来のようなアプリケーションプログラムだけの公開ではなく、オペレーティングシステム、ミドルウェア、そしてアプリケーションといった実行に必要な全ての環境そのものを提供することができます。このため、特にクラスター計算機などの規模の大きなシステムを構築する際のユーザ利用環境によるソフトウェアプログラムの再構築や、システム変更などの不便を解消することができるという特徴を発揮します。

## 2. 研究手法と成果

クノーピクスで構成されたシステムは、書き込みが必要な情報を、計算機のメインメモリで構成されたラムディスクに対して書き込みます。書き込み変更の必要のない読み出しだけのプログラムファイルなどは、CD-ROM 上から直接読み出します。そのため、システムをハードディスクにインストールする必要のない計算機利用環境を提供することができます。クノーピクスは、デバイスの認識率が高いことで知られており、多くの計算機に対して利用することができますが、基本的にラムディスク上で書き込み情報を利用することになるので、その容量はハードディスクと比較して非常に限られています。そのため、バイオインフォマティックスで用いられる膨大なデータベースを利用するタイプのアプリケーションに対しては不向きであるとされてきました。

研究チームは、構成されているワークノードが持つラムディスクに、「PVFS2※8」や「Gfarm（ジーファーム）※9」等の並列クラスターファイルシステムを適用し、ディスク不要な計算機システムとすることで解決しました。その結果、膨大な量のデータを利用するデータベース利用型のアプリケーションでも十分に実用レベルのシステムを簡便に構築することができるようになりました。具体的には、数十ノードの計算機システムを 10 分程度の時間で構築設定することが可能となりました。

これは、熟練したシステムインテグレータが、システムのインストールと調整を行うよりも遥かに早く構築できるということです。さらに、設定に必要なほとんどの操作が、Web ブラウザから必要事項を選択確認することができ、大規模なシステムインテグレートに関する専門的な知識が無くても利用することができます。

このように、バイオインフォマティックスにクノーピクスを利用する際に起きる諸問題を解決し、実践的に活用できるシステムの簡便な構成を実現しました。特に、構築されたシステムの再起動後に元のシステム状態に復元することができるメリットが生まれたことにより、既存の計算資源を安心して活用する手段を提供することができるようになります。

## 3. 利用方法

開発したシステムは、理研横浜研究所ゲノム科学総合研究センター（GSC）ゲノム情報先端技術研究グループ広域分散情報統合研究チームのホームページにおい

て、CD-ROM ISO イメージのダウンロードによるサービスを準備中で 12 月 1 日から利用が可能となります。（http://big.gsc.riken.jp/index_html/Members/fumikazu/htc/）また、CD-ROM のメディアによる配布は、送料のみの負担で利用することができます。

対象ユーザは、インタープロスキャンへの大量検索の実施を希望する研究者で、クラスター計算機のワークノードの総メモリ量として、6 ギガバイト以上のメモリを確保することができる方です。一般的なクラスター計算機用ノードのメモリ実装量を 1 ギガバイトとすると、6 台以上の計算機でシステムの構築が可能です。

**4. 今後の展開**

現在、理研から公開・提供されているデータベースやアプリケーションプログラムなどに対して、本システム構築で培った技術を適用していき、より簡便で柔軟なデータベースやアプリケーションの提供などに応用していきますが、これは、デリバリーインフラストラクチャー※10 としてグリッド技術と同様に、情報のより高度な共有と利用につながっていきます。さらに、バイオインフォマティクスやライフサイエンス研究の発展に貢献できるように、さらなる研究開発を推進します。

（問い合わせ先）
独立行政法人理化学研究所　横浜研究所
ゲノム科学総合研究センター　ゲノム情報先端技術研究グループ
広域分散情報統合研究チーム
研究員　　小西　史一
Tel：045-503-9602 / Fax：045-503-9613

研究推進部　企画課　　溝部　鈴
Tel：045-503-9117 / Fax：045-503-9113

（報道担当）
独立行政法人理化学研究所 広報室
Tel：048-467-9272 / Fax：048-462-4715
Mail：koho@riken.jp

## <補足説明>

**※1 InterProScan（インタープロスキャン）**

European Bioinformatics Institute(EBI)により開発されたタンパク質のドメイン機能検索のためのアプリケーションソフトウェア。入力として与えられるタンパク質配列情報から、既知の様々な機能的な部分を検知することができることから、未知のタンパク質配列に適用して入力タンパク質の機能同定をサポートするアプリケーション。(http://www.ebi.ac.uk/interpro/)

**※2 CD-ROM ISO イメージ**

世界標準化を促進するための規格機関である、International Organization for Standardization（略名:ISO）により規格化された規定に沿って、CD の内容を 1 つのファイルとして変換したもの。

**※3 Knoppix（クノーピクス）**

ドイツで開発された CD から起動することができる LiveOS 型の Linux ディストリブーション。システムの起動を CD 等のデバイスから行い、各種デバイスなどを自動的に認識する能力に長けており、プログラムなど改変が必要のないプログラムは、直接 CD 上のファイルを利用し、改変の必要な設定ファイルなどは、主記憶装置から分割された RAM ディスク上に保管することにより、計算機上のハードディスク上に直接インストール無しに、計算機利用環境を構築することができる機能をサポートしている。日本国内の Knoppix 情報は、http://unit.aist.go.jp/itri/knoppix/）で情報が公開されている。

**※4 Linux（リナックス）**

Linus Torvalds 氏によって開発された、UNIX 互換の OS。既存の他の OS のコードを使わずに、各国ボランティアの手によって改良が続けられている。その特徴は、ネットワーク機能やセキュリティーに優れ、また非常に安定している。

**※5 ディストリブーション**

インターネットなどから入手したソフトウェアを、さらに別のサイトなどで公開し、第三者が入手できるように配布すること。

**※6 European Bioinformatics Institute（EBI）**

バイオインフォマティクスの研究とサービスを行っているセンター。DNA、タンパク質の配列および構造等の生物学データの保守管理を行う。

**※7 Condor（コンドル）**

米国ウィスコンシン大学マディソン校の Prof..Miron Livny がリーダーとなり、1985 年からスタートした Condor Research Project で配布されている HTC(High Throughput Computing)を実現するために開発された高機能ジョブスケジューラ（作業管理）システム。(http://www.cs.wisc.edu/condor/)

**※8 PVFS2**

米国 Clemson 大学とアルゴンヌ国立研究所により、オープンソースとして開発された並列クラスターファイルシステム。複数のノードが IO ノードとなりディスクスペースを提供し、ファイルをその複数のノードに細切れに格納することで、大規模なディスクストレージシステムを構築することができる。
（http://www.pvfs.org/pvfs2/）

**※9 Gfarm（ジーファーム）**

Gfarm ファイルシステム。産業技術総合研究所により、オープンソースとして開発されている並列クラスター／グリッドファイルシステム。複数の計算ノードのディスクスペースを利用し、ファイルをそれぞれの計算ノードに格納すること、ファイルの複製を複数ノードにもつことにより高性能、高信頼性を実現する。
(http://datafarm.apgrid.org/)

**※10 デリバリーインフラストラクチャー**

配布のための通信基盤整備

Knoppix for InterProScan High Throughput Computing Editon のシステム構成概要