コンテナ型仮想環境によるタンパク質間相互作用予測システムの大規模並列化

青山 健人 [1,2)]　山本 悠生 [1,2)]　大上 雅史 [1,3)]　秋山 泰 [1,2,3)]

1) 東京工業大学 情報理工学院 情報工学系
2) 東京工業大学 情報生命博士教育院
3) 東京工業大学 科学技術創成研究院 スマート創薬研究ユニット

1. 概　　要

軽量かつ高速な仮想化技術の 1 つであるコンテナ型仮想化は，ソフトウェア環境の可搬性やデータ解析の研究成果の再現性に優れており，研究機関の大規模並列計算環境へ導入されはじめている．本研究では，生命情報科学分野のアプリケーションに対するコンテナ型仮想化の導入と分散基盤の構築のため，大規模並列環境を想定したタンパク質間相互作用予測システム “MEGADOCK” [1] に対して，コンテナ型仮想化を利用した分散処理システムをクラウド環境上に構築した．これにより，様々な計算機環境でアプリケーションの依存環境を意識せず，迅速かつ容易にタンパク質間相互作用ネットワーク予測のための並列計算環境を構築可能とした．

2. MEGADOCK

2.1 MEGADOCK の並列実装の概要

MEGADOCK は分散メモリの大規模並列計算環境を想定し，大量のタンパク質ペアの相互作用予測計算を MPI/OpenMP によるハイブリッド並列で実装している（図 1）．対象タンパク質のペアを MPI によりノード間に割り振り，各タンパク質ごとの相互作用予測の計算を OpenMP によってスレッドごとに並列計算する．相互作用予測の計算では高速に並列処理が可能な GPU，MIC も利用可能である．

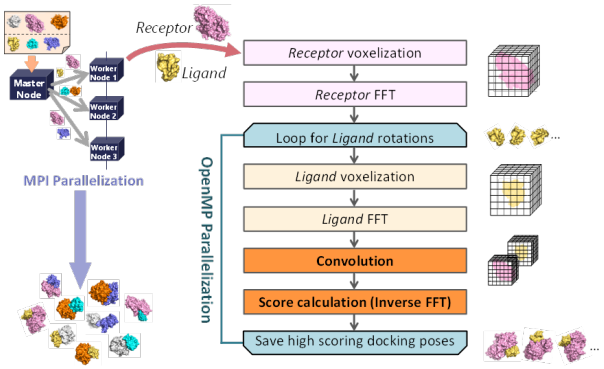

図 1　**MEGADOCK** のワークフロー概要

2.2 タンパク質のドッキング計算

MEGADOCK はタンパク質のドッキング計算において，ボクセル化モデルに基づいた手法を採用している．評価関数では形状相補性，静電相互作用，疎水性相互作用を考慮しており，高速フーリエ変換（FFT）により，これらを同時に計算・評価する．現在は FFT ライブラリとして `FFTW`（GPU 版では `CUFFT`）を用いている．MEGADOCK の処理は約 6 割が FFT によるドッキング計算，約 3 割が計算結果の出力であり，プロセス間・スレッド間の同期・通信の影響は小さく，計算とファイル I/O が律速となるアプリケーションである．

2.3 ソフトウェア公開情報，および実績

本報告ではクラウド環境 Microsoft Azure 上の実行結果を示す．MEGADOCK は東京工業大学の TSUBAME 2.5，理研 AICS の「京」コンピューターで応用研究にも利用されている．MEGADOCK のソースコードは次の URL から取得可能である．`http://www.bi.cs.titech.ac.jp/megadock`

3. 評 価 実 験

3.1 システム構成

クラウド環境 Microsoft Azure の仮想マシン上で Docker により MEGADOCK の実行環境とコンテナ間ネットワークを構築した．仮想マシンのインスタンスには `Standard_D14v2`（CPU 16 コア，RAM 112GB，SSD 800GB）を採用した．各仮想マシンでは MEGADOCK の計算コンテナを起動し，コンテナ内で MPI プロセスを起動する．1 VM あたり 4 MPI プロセス，合計 16 スレッドにより並列処理を行った．

3.2 並列性能測定

仮想マシン 30 台（ワーカーコア数 39.7 倍）のとき並列処理による高速化率は，仮想マシンで Docker コンテナを利用した並列処理では 35.5 倍（強スケール 0.895），仮想マシン上の並列処理では 36.6 倍（0.924）であった．

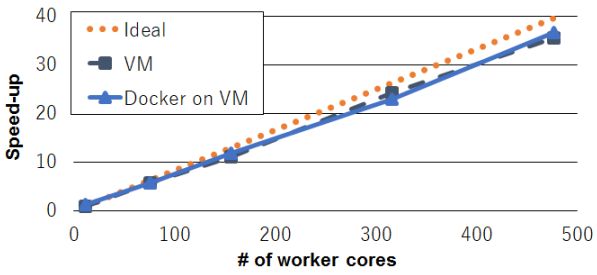

図 2　並列性能測定（強スケール）
`VM`: 仮想マシン上の並列処理の高速化率
`Docker on VM`: 仮想マシンのコンテナ上の並列処理の高速化率

4. 結　　論

コンテナ型仮想化によりクラウド環境上に迅速にタンパク質相互作用予測システムの並列処理環境を構築した．生命情報科学分野に多い計算律速なアプリケーションではコンテナ型仮想化の性能低下の影響は小さく，迅速かつ容易な計算環境の構築と合わせて応用研究の促進に有用であることが示唆される．

文　　献

[1] M. Ohue, T. Shimoda, S. Suzuki, Y. Matsuzaki, T. Ishida, Y. Akiyama., “MEGADOCK 4.0: An ultra-high-performance protein-protein docking software for heterogeneous supercomputers,” *Bioinformatics*, vol. 30, no. 22, pp. 3281-3283, 2014.

[2] 青山健人, 山本悠生, 大上雅史, 秋山泰. “コンテナ型仮想化による分散計算環境におけるタンパク質間相互作用予測システムの性能評価”, 情報処理学会研究報告 バイオ情報学（BIO）, 2017-BIO-49(3), 1-8, 2017.