e-Science Journal
東京工業大学 学術国際情報センター
Vol. 10

TSUBAME ESJ.

TSUBAME2.5への進化
The TSUBAME2.5 Evolution

非エバルト法に基づくGPUで加速された G蛋白質共役型受容体の分子動力学計算
Molecular Dynamics Simulation Accelerated by GPU for GPCR with a non-Ewald Algorithm

巡回セールスマン問題に対する 反復局所探索の大規模並列アルゴリズム
Large-scale Parallel Iterated Local Search Algorithm for Traveling Salesman Problem

http://www.gsic.titech.ac.jp/TSUBAME_ESJ

TSUBAME2.5への進化

松岡 聡
東京工業大学 学術国際情報センター

2010年11月より東京工業大学・学術国際情報センターが開発・運用するTSUBAME2.0は、2013年9月より、TSUBAME2.5にアップグレードされた。理論最高性能は倍精度演算では2.4ペタフロップスから5.7ペタフロップスへ、単精度演算性能では4.8ペタフロップスから17.1ペタフロップスへと3倍以上になり、2013年9月現在ではこの尺度では我が国最速のスパコンとなった。大幅な性能向上にもかかわらず、平均電力は二割程度削減され、ソフトウェアの上位互換性も保持され、今後も我が国の学術スパコンの全国インフラであるHPCIにおいて、リーディングマシンの一つとして活躍することが高く期待される。本稿では、TSUBAME2.0から2.5へのアップグレードの詳細と、2015年度後半に稼働が期待されるTSUBAME3.0への繋がりを示す。

はじめに 1

本センターのTSUBAME2.0は、3年近くの長きにわたり、我が国初のペタフロップスマシン・HPCIにおけるリーディングスパコンとして活躍してきた。それが2013年9月よりTSUBAME2.5にアップグレードされ、理論最高性能は倍精度演算では2.4ペタフロップスから5.7ペタフロップスへ、単精度演算性能では4.8ペタフロップスから17.1ペタフロップスへと3倍以上になり、2013年9月現在ではこの尺度では我が国最速のスパコンとなった。しかし、そのアップグレードは初期の計画にはなく、決して平坦だったわけではない。本稿の目的は、そのアップグレードの様々な経緯を解き明かすものである。

図1 2013年9月10日稼働のTSUBAME2.5

TSUBAME2.0 に関して 2

TSUBAME2.0は、東京工業大学学術国際情報センターにて設計、NEC/HP/NVIDIAなどの企業連合体と共同開発され、2010年11月に運用開始されたスーパーコンピュータである。国内で初のペタフロップス越えを達成し、2010年11月のTop500ランキングにおいては世界4位、同Green500ランキングにおいては世界2位および“Greenest Production Supercomputer”賞を獲得した。また、2011年11月には、京コンピュータと並び、スパコンの実アプリケーションの最高峰の賞であるACM Gordon Bell Awardを獲得した。2013年7月末の段階で、登録ユーザは1万人近いが、実際のスパコンユーザは2000人ほどで、常時100名近くが利用している。メンテナンスや震災による一時的停止、および夏季のピークシフト運転の昼間の一部のノード停止を除き、基本的に24時間稼働し続けており、多くの科学的な成果を生み出している。

TSUBAME2.0は、他の7大学の全国の情報基盤センターのスパコンらとともに、全国共同利用センターの主要スパコンとしての役割を果たしており、全国の学術機関から広く利用が可能である。また、理化学研究所・計算科学研究機構（Riken AICS）の京コンピュータや、先の全国共同利用の主要大学のスパコン、並びに海洋研究機構（JAMSTEC）の地球シミュレータ等の主要国立研究所のそれらとともに、我が国の学術研究の共通スパコン基盤であるHPCI（High Performance Computing Infrastructure）の一員となっており、一定の計算資源はその審査過程に基づき配分される。更には、企業利用も積極的に進んでおり、先端研究基盤共用・プラットフォーム形成事業などを通じて、100社以上の民間企業が用いている。

TSUBAME2.0は最先端のスパコンとしてNEC・HP・NVIDIA・DDNなどの多くの世界的なスパコン関連の企業と共同開発され、種々の最先端の技術を内包している。その技術的特徴は以下の通りである：

図2 新型GPU NVIDIA Kepler K20Xが搭載された TSUBAME2.5の計算ノード

高演算性能・高バンド幅を持つ計算ノード：

主要部はThinノードと呼ばれるノード1408台から成り、各ノードはマルチコアCPU2基とGPU3基、およびメモリ54GBまたは96GBを搭載したHP社HP Proliant SL390s G7である。CPUとしてはIntel社製6コアXeon（Westmere-EP）2.93GHzを搭載し、GPUとしてはNVIDIA社製 Tesla M2050を搭載する。各M2050 GPUには448個のCUDAコア、および3GBのメモリが搭載され、その理論ピーク性能は515GFlops、理論メモリバンド幅は150GB/sであり、Thinノード一台の理論ピーク性能は約1.6TFlops以上、合算メモリバンド幅は500GB/s近くに達する。また、ノード内で特に大容量の共有メモリを必要とするアプリケーションのために、メモリの多いMedium/Fatノード（メモリ量128GB、256GBまたは512GB）も合計40台存在する。

階層型大規模ストレージ：

TSUBAME2.0のストレージは、様々なI/O要求を行う多種アプリケーションのため階層化され、（1）ノードローカルSSD、（2）ディスクによる共有ストレージ、（3）テープライブラリから成る。

（1）各計算ノードはローカルストレージとして、ハードディスクの代わりに120-240GBの容量のsolid state drive（SSD）を持つ。チェックポイントなどの用途でこのSSDを利用することにより、共有ストレージへの負荷を軽減する。ノードあたりのSSDのバンド幅は300MegaByte/s以上、マシン全体では1/2TeraByte/sに達する。
（2）共有ストレージは合計7.2ペタバイト（PB）の容量（raw capacity）を持つ。このストレージは6つのファイルシステムに分割され、1つがホーム領域、5つが並列ファイルシステム領域として利用される。各ファイルシステムにおいてデータを蓄積するのはDDN社 SFA 10000ストレージシステムである。ホーム領域は高信頼性に焦点をおきつつ、高速なNFS性能、およびCIFS、iSCSIなどの複数プロトコルに対応する。並列ファイルシステム領域はスケーラビリティを焦点に設計されており、ファイルシステムとしてLustreおよびGPFSを採用する。
（3）さらにバックアップを主目的とした三次領域として、計8PB（圧縮）の既存のSL8500テープライブラリと接続されている。階層的ファイルシステムの利用により、このテープライブラリと並列ファイルシステム領域間で透過的なデータアクセスを提供する。

これらストレージシステムは、過去のTSUBAME1.0においてストレージがしばしばSingle Point of Failureになった経験から、大幅に多重化され、かつ自動回復されるようになっている。例えば、ストレージに対するInfiniBandでのアクセスパス・コントローラ・サーバ群は全て多重化されており、かつ障害時には自動的にfailoverする。HDDはRAID6であることは勿論で、かつ個体HDDの障害時には自動的にスペアドライブからリビルドが行われる。ファイルシステムはNFS、Lustre, GPFSと多重化され、致命的なバグ時にもシステムソフトウェアレベルで多重化されている。

フルバイセクションネットワーク：

1400以上の計算ノードおよびストレージを結合する高速インターコネクトとして、4× QDR InfiniBandを採用している。インターコネクトはほぼ同形の二つのInfiniBandネットワークから成り、それぞれがフルバイセクション・ファットツリーと呼ばれるトポロジを成す。ファットツリー構造の上部にはコアスイッチとして324ポートの大規模InfiniBandスイッチを計12個採用し、コアスイッチと下部のエッジスイッチ間は光ファイバーで結線されている。ファイバーの本数はシステム全体で3500本、総延長100km程度である。ノードあたりのインジェクションバンド幅は理論的には80GigaBpsであるが、実測でもノード間のバンド幅は7.5GigaByte/s、レーテンシは2μ秒以下である。システム全体のバイセクションバンド幅は220TeraBpsであり、全世界のインターネットの全体のトラフィックの総和の平均を超える。

TSUBAME2.0 から 2.5 へのアップグレード計画 3

TSUBAME2.0は、その先進性もあいまって、高い利用率を誇ってきたが、近年ではその容量が限界に達してきていた。特に、年度の後半の繁忙期では、その利用率が90%を常時超えて、100%に達することもあり、我が国2位の計算速度にも拘わらず、その容量不足は明らかとなっていた。また、世界的に鑑みてもTSUBAME2.0の相対的な計算能力やランキングは経年的に低下してきていた。10年間で1000倍近い性能向上を果たすスパコンの世界では仕方がないことであるが、それでも2010年11月の稼働時には日本一位、世界4位だったのが、2012年11月段階では日本3位、世界20位程度に低下し、その競争力の低下はあらわであった。

TSUBAME2.0は本来4年間の運用を前提に契約・調達したので、このような性能低下は2014年11月にTSUBAME3.0が実現できれば問題なかった。しかしながら、昨今の半導体プロセスの進化のスローダウンにより、予定していた半導体プロセス、およびそれによるCPUやGPUなどが用いられないことが各メーカーとの協議で明らかになってきた。本来はIntel、NVIDIAなどのメーカーはそれでも2年毎に新しいプロセスとアーキテクチャを採用してきたので、本来4年ならばTSUBAME2.0から2世代先のプロセッサが使えるはずであったが、一世代のプロセッサの変革では、大幅な性能向上を果たすのは困難であり、TSUBAME3.0の性能目標を満たすことが不可能なることが予測された。

さらに2011年3月11日の未曾有の大震災により防災対策の重要性が改めて認識され、スパコンも防災・環境・医療・ものづくりなど、国民の直接の関心事に直接応えていくことが要求されるようになってきた。TSUBAME2.0ではすでにこれらのアプリケーションは多々実行されており、実際地震のシミュレーションは国のハザードマップの

同定に大いに貢献した。しかしながら、特に繁忙期に100%近くなる利用率により、これらのアプリのタイムリーな計算容量の確保が困難になってきていた。よって、単に性能向上だけでなく、これらのアプリケーションを優先してスケジューリングする方式が求められた。

最後に、HPCIのインフラとして、上記の国民の関心事の高いアプリケーションが、TSUBAMEのみならず、京や他の基盤センターの複数のスパコンで相互に連携してスムースに動作させる必要があった。スパコン間の連携をスムースにするためにHPCIでは全国統一のHPCIアカウントや、そのための認証システムなどを整備しているが、特にスパコン間で連携が重要なのは共通のストレージであり、2012年度時点では東には東京大学の情報基盤センター、西は理研AICSに数十ペタバイトの共通HPCIストレージが確保されているが、TSUBAMEもそれらと連携するストレージが必要であることが明白になった。

以上のような現状から、本センターではTSUBAME2.0の運用期間の延長と、それによって生じる余剰資金でTSUBAMEの部分的なアップグレードを計画するとともに、フルシステムのアップグレードを2012年9月ごろに文部科学省に提案し、かつ同時に政府調達（いわゆるスパコン調達）も開始した。

調達開始後、しばらくしての2012年度後半に我が国に新政権が誕生し、景気浮揚と震災対策のために大規模な補正予算が施行され、HPCIも強化の申請をすることとなった。その中で本センターは、「TSUBAME2.0スパコンの防災シミュレーション増強」との名称で、M2050GPUの全面アップグレードと、その他上記の強化を提案した。それがほぼ2012年冬に全面的に認められ、結果としてTSUBAME2.0→2.5のフル実装のアップグレードが可能となった。最終的にはNECがNVIDIA、HPと組んだ連合体がTSUBAM2.5を2013年7月12日に入札し、夏の稼働中および夏季停止期間に入れ替え作業を行って、2013年9月16日にTSUBAME2.5として稼働を開始した。

TSUBAME2.5 へのアップグレードの技術詳細 4

2013年9月にTSUBAME2.0のM2050GPUを最新のKepler世代のTesla K20XへGPUアクセラレータを置き換えることにより、理論ピーク性能5.76PetaFlops（倍精度）、17.1PetaFlops（単精度）を持つTSUBAME2.5にアップグレードされた。また、メモリバンド幅は理論ピーク値では1.16 PetaByte/s、実測でも約0.8 PetaByte/s に増速された。図3がアップグレードの概要、表1 がTSUBAME2.0と2.5の主な比較表である。

補正予算が措置され、資金的には大幅な増額となったが、TSUBAMEのアップグレードに際し、新たな計算ノード群の追加は、電力やスペースの限界から、最初に選択肢から除外された。他の方式も検討したが、やはり当初の計画通りアクセラレータを新型にアップグレードする事に決めた。しかし、TSUBAMEは数千人もユーザがいる運用スパコンであるので、種々の技術的な問題が露呈した：

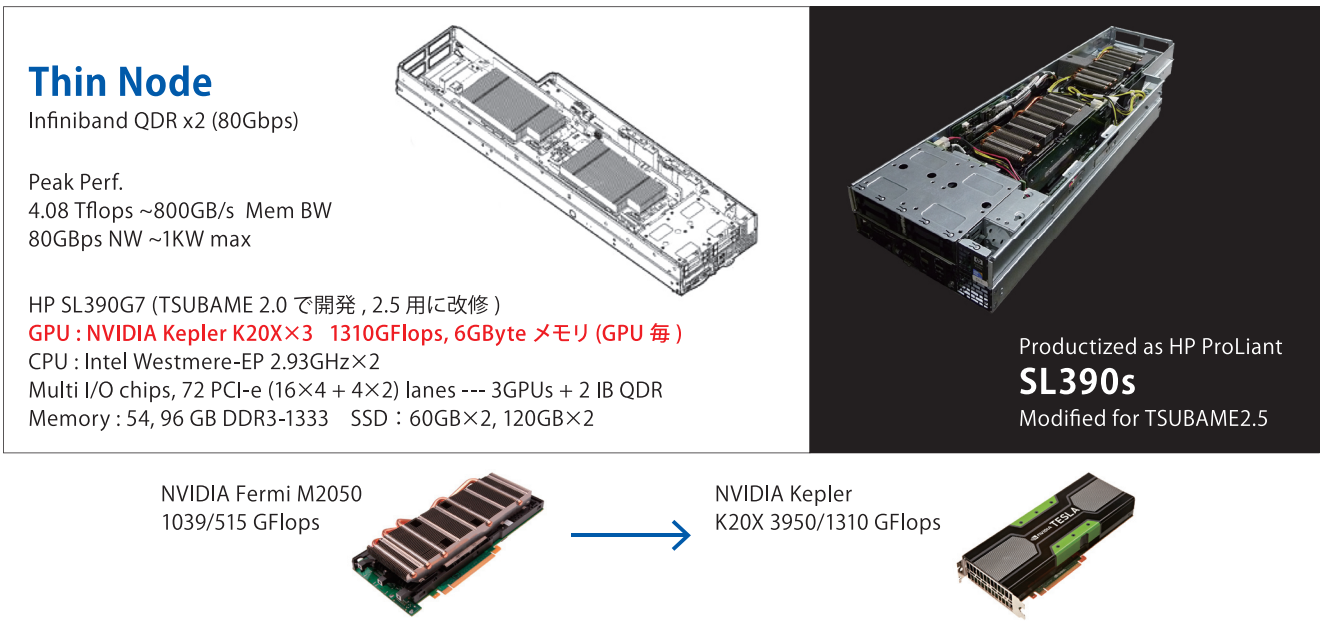

図3 TSUBAME2.5 の計算ノードの進化

(1) **どのメニーコアアクセラレータか：**
単純性能だけでなく、今までのTSUBAME2.0のソフトウェアスタック上のアプリケーションとの上記互換性は確保できるのか？特に、アップグレードはTSUBAME2.0調達後に計画されたもので、当初の運用計画にはない。よってユーザにとって予定外のシステム変更は受け入れ難いのは自明であり、上位互換性の確保が必須となる。

(2) **新たなアクセラレータTSUBAME2.5上で動くのか：**
ハードウェア上の互換性はあるのか、もし電力増加があったらノード・ラック・さらには全体の電源系や冷却系は耐えられるのか？自作パソコンを秋葉原でパーツを買ってアップグレードするのとは異なり、大規模スパコンでは複数年に渡る24時間の高負荷運用でも耐えうる信頼性が必要であり、アップグレードされたアクセラレータがシステム全体の安定運用を損なわないかの検証は単純ではない。

(3) **効率は出るのか：**
特にネットワークや他の部分はボトルネックにならないか？
また候補アクセラレータは、単体性能がM2050の2~3倍になると予想されたので、相対的にPCI-eやネットワーク、さらにはストレージがボトルネックとなり、下手をすると全く性能向上が達成されない可能性すらあった。

(4) **単精度演算の高速化は意味があるのか：**
上記と関連するが、一部の候補アクセラレータは単精度演算ピーク性能が倍精度演算のそれと比べて、M2050の2倍から3倍や4倍に比率が増速されるが、それは実際のアプリケーションで意味があるのか。Intelのx86アーキテクチャのそれは2倍であるが、IBM BlueGene/Qや京/富士通FX/10では倍精度も単精度もピーク性能は同一である。TSUBAME2.0では多くのアプリケーションで単精度性能の高さは性能向上に役立っているが、その倍率が更に高まった場合、メリットがあるかどうかの検証が必要であった。

(5) **運用上ユーザに影響を与えずにアップグレードが可能か：**
最高で1408計算ノード上の4224枚のGPUを新型アクセラレータに交換するが、それらは各ノードに基本的にノードの停止、交換、動作テストを伴い、また交換中はシステム全体として旧型のM2050と新型の候補アクセラレータが混じる事になる。これらの作業期間は一か月以上に渡ると予測されたので、これらの制約下で、長期の予定外のシステム停止などをせずに、運用上ユーザにほとんど影響を及ぼさずにアップグレードが可能か、は大きな課題となった。

以上のような技術的な課題を、時間をかけて我々は解決していき、またそれらはTSUBAME2.5の設計、特に調達仕様に反映された。それらの詳細を述べるのは本稿のスペースでは無理だが、それぞれどのような解決にあたったかの概要を述べる。

	TSUBAME2.0	TSUBAME2.5
Thin Node x 1408 台		
Node Machine	HP Proliant SL390s	← 変更なし No change
CPU	Intel Xeon X5670 (6core 2.93GHz, Westmere) x 2	← 変更なし No change
GPU	NVIDIA Tesla M2050 x 3 ● 448 CUDA cores (Fermi) ➢ 単精度 SFP 1.03TFlops ➢ 倍精度 DFP .515TFlops ● 3GiB GDDR5 memory ● ~90GB/s STREAM BW 実測メモリバンド幅	NVIDIA Tesla K20X x 3 ● 2688 CUDA cores (Kepler) ➢ 単精度 SFP 3.95TFlops ➢ 倍精度 DFP 1.31TFlops ● 6GiB GDDR5 memory ● ~180GB/s STREAM BW実測メモリバンド幅
ノード性能 Node Performance (incl. CPU Turbo boost)	● 単精度 SFP 3.40TFlops ● 倍精度 DFP 1.70TFlops ● ~300GB/s STREAM BW 実測メモリバンド幅	● 単精度 SFP 12.2TFlops ● 倍精度 DFP 4.08TFlops ● ~570GB/s STREAM BW実測メモリバンド幅
TOTAL System		
理論演算性能 Total Peak Performance	● 単精度 SFP 4.80PFlops ● 倍精度 DFP 2.40PFlops ● 実測メモリバンド幅 ~440TB/s	● 単精度 SFP 17.1PFlops (x3.6倍) ● 倍精度 DFP 5.76PFlops (x2.4倍) ● 実測メモリバンド幅 ~803TB/s (x1.8倍)

表1 TSUBAME2.0とTSUBAME2.5のThin Node群のスペック比較表

(1) **どのメニーコアアクセラレータか**：
TSUBAME2.0開発の段階ではスパコンで用いることができるメニーコアアクセラレータは現実的にはNVIDIA社のFermiシリーズのTesla GPU（M2050）のみであった。実際、3年にも渡る運用において、4224枚のM2050は、若干のトラブルを除けば、おおむね安定したメニーコアアクセラレータとして動作し、近年ではその利用率は3-5割にものぼり、全系での利用も十分に安定に可能であった。しかしながら、TSUBAME2.5の候補アクセラレータとしては、Fermiの次の世代のKepler GPUのみならず、同じGPUとしてAMD社のFireStream、そして何よりIntel社のGPU由来のメニーコアでx86命令セットと互換性があるXeon Phiも選択肢に入った。特にPhiはGPUとの上位互換性はないものの、ノードのCPUのXeonとほぼ上位互換性があり、重要な選択肢となった。しかしながら、種々のテストで、PhiはCPUとの命令互換性はあるものの、TSUBAME2.0上でFermiGPU実行されるアプリケーションの性能にマッチするにはかなりのチューニングが必要である事も判明した。そこで、TSUBAME2.5のスパコン調達においてはアップグレードを二つに分け、前者はTesla Fermi GPUのアプリの高い性能上位な互換性を要求し、後者はソフトウェア上の互換性のみとした。結果として、両者ともNVIDIA社のKepler K20Xにアップグレードする提案が調達で公募採択された。

(2) **新たなアクセラレータTSUBAME2.5上で動くのか**：
K20XやXeon Phiなどの候補アクセラレータはカタログスペック上のTDPはTSUBAME2.0のFermi M2050の225Wを上回る235-300Wであり、電源系や冷却系の増強が必要となった。

また、TSUBAME2.0のHPと開発したThin Node計算機であるHP SL390G7は、Keplerで更新された電源制御のプロトコルをサポートせず、実質接続しても動作しない状況にあった。これらの問題はエンジニアリング的ではあるが、HP社が他社と連携して解決する必要があり、決してトリビアルとは言えない問題であった。幸い、粘り強い努力の末、これらの問題を解決する開発を行うこととなり、新型アクセラレータも無事安全に動作することとなった。

(3) **効率は出るのか、特にPCI-Eやネットワークや他の部分はボトルネックにならないか**：
メニーコアアクセラレータを用いたスパコンアーキテクチャの最大の問題点は、CPUと比較してその数倍高い演算とメモリバンド幅に対し、I/Oやネットワークの性能が十分確保できるかである。TSUBAME2.0ではそれらが十分になるように設計され、マルチI/Oハブや複数レールのQDR InfiniBandを装備したが、候補アクセラレータの2-3倍の性能向上に対して、それらが果たして十分となるか、という点が問題となった。例えば、Linpackの初期の測定においては、SL390の旧世代のIntel Westmere Xeon + Tylersberg IOHの組み合わせは、Fermi M2050 GPUでは十分な性能を発揮するも、Kepler K20XではHP SL250等に用いられている新世代のSandy Bridge Xeonと比較するとPCI-e間の合算バンド幅が十分でなく、ノード単位では半分程度の性能が出ないことが判明した。(図4)

同様の問題は、InfiniBandネットワークにおいてもTSUBAME2.0のQDR世代では二つのレール合算で7GigaByte/s強だが、これはFDR世代の2レールで達成可能な12~13GigaByte/sの高々半分強であった。

そこで、I/Oやネットワークが律速にならないように、種々の研究開発を進行させた。Linpackに関しては、PCI-eの性能に左右されづらいアルゴリズムに変更した。また、ネットワークの詳細な研究により、ルーティングのボトルネックを解消したり、あるいはレーテンシ隠ぺいによってオーバーヘッドを解消するアルゴリズムを進化させたりした。これらにより、完ぺきではないものの、性能は新世代のI/Oやネットワークのそれにかなり近づいた。

(4) **単精度演算の増強が有効か**：
HPCアプリが倍精度計算をしているのは、本質的に精度を必要とする場合もあるが、多くの場合は昔のスパコンにおいてCPUの倍精度計算の性能が単精度と変わらなかったので、簡便性のために単精度を精度的に包含する倍精度計算が行われ、それが受け継がれている場合が多い。しかし、昨今の単精度の性能の伸びから、単精度を積極的に利用する研究が色々行われ、実アプリにも多く反映されている。例えば、地震波伝搬・津波・気象/気候で主流である、陽的な時間積分のアプリケーションは(長時間積分の場合でも)ほぼ単精度計算が可能である。また、同様の理由でものづくりにおける電磁波解析や、飛行機・自動車の空力解析、創薬などでの分子動力学なども、多くは単精度で解析が可能であり、TSUBAME2.0上の多くの大規模アプリで実証され、単精度・あるいは混合精度演算は倍精度と比べ倍近い性能を示している。

問題は、それ以上のピーク性能比があった場合、アプリケーションでそれを活かせるか、である。2:1以上の比率では、当然計算密度が高くなるので、そのような高い密度でありつつ、高精度である必要が少ないアプリケーションである必要がある。所謂N体(重力多体)問題や、力場計算にFMM(高速多極展開法)を用いる分子動力学計算などはそれに十分当てはまるが、今後他のアプリケーションでも検討の必要がある。

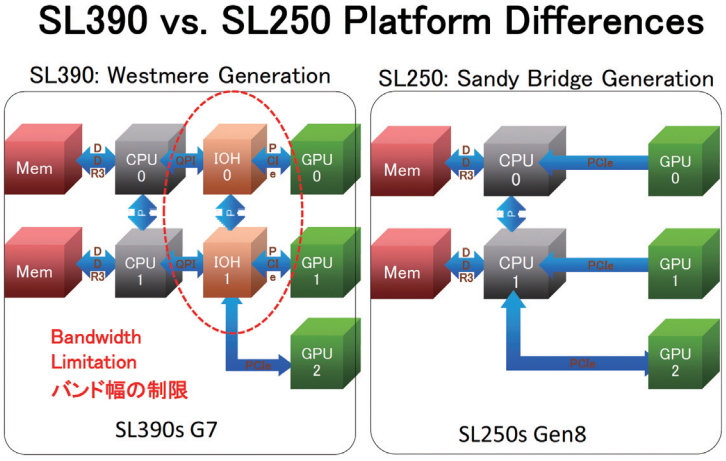

図4 SL390とSL250のI/Oハブのアーキテクチャの比較

(5) **運用上ユーザに影響を与えずにアップグレードが可能か**：
TSUBAME2.0は震災以降に、文部科学省「グリーンスパコン」プロジェクトの元に、昼間に一部のノードを自動停止し、夕方復活させ、ジョブスケジューリングもそれに追随させる「ピークシフト運転」の技術を開発し、7月から9月の夏季の昼間に実際に運用してきている。今回のTSUBAME2.5へのアップグレードを夏に行うことにより、このピークシフトを活用し、停止されるノードを順繰りに変更し、停止時に交換を行うことにより、影響を最小限とした。また、夏季の全キャンパスの停電期間も積極活用し、マシンが停電＋メンテナンスでダウンしている時期に集中的に作業を進めた。マシン構成が混合する際には、GPU利用率が高いバッチキューはなるべく一気に停電期間中にアップグレードが行われるようにし、それ以外のキューでも、ユーザが各ノードが旧型なのか新型なのかがわかり、それをスケジューリングに生かせるような方式を通知した。結果として特にトラブルもなく作業は7月下旬より進み、予定より2週間早い8月末には全GPUのKepler 20Xへの交換が終了した。

以上により、TSUBAME2.0 → TSUBAME2.5へのアップグレードは達成され、大幅に性能向上 し、運用期間が一年以上伸びることとなった。現状のGPU中心のベンチマークテストでは予定通り2~3倍程度の性能向上が達成されている。具体的には以下のような表2のような性能向上である。

全般的に予定通りの性能向上だが、特にGreen500の性能向上が大きい。これはKeplerのアーキテクチャが性能向上のみならず、省電力化も果たしたことと、その以外の部分のTSUBAME2.0の電力効率も元々良い事が大きい。一方これらにおいてネットワークの律速の顕在化も観測されており、それらを解消する事でTSUBAME2.5の性能向上は今後より大きくなる可能性がある。無論、他のアプリやベンチマークでも比較ベンチマークを行っていく予定である。

Application	TSUBAME2.0 Performance	TSUBAME2.5 Performance	Boost Ratio
Top500/Linpack 4131 GPUs (PFlops)	1.192	2.843	2.39
Green500/Linpack 4131 GPUs (GFlops/W)	0.958	3.069	3.20
Semi-Definite Programming Nonlinear Optimization 4080 GPUs (PFlops)	1.019	1.713	1.68
Gordon Bell Dendrite Stencil 3968 GPUs (PFlops)	2.000	3.444	1.72
LBM LES Whole City Airflow 3968 GPUs (PFlops)	0.592	1.142	1.93
Amber 12 pmemd 4 nodes 8 GPUs (nsec/day)	3.44	11.39	3.31
GHOSTM Genome Homology Search 1 GPU (Sec)	19361	10785	1.80
MEGADOC Protein Docking 1 node 3GPUs (vs. 1CPU core)	37.11	83.49	2.25

表2 TSUBAME2.0からTSUBAME2.5への性能向上

おわりに – TSUBAME3.0 に向けて 5

TSUBAME2.5へのアップグレードにより、TSUBAME2の運用期間は5年以上に伸びた。現在、我々はTSUBAME3.0のための諸技術を研究開発し、その設計を行っているが、先に述べたように各ベンダーの新世代のプロセッサの開発スケジュールに作用される。現在、TSUBAME3.0は2015年度末（2016年3月末）までの稼働を目指しているが、その場合TSUBAME2.5としては2年半以上、TSUBAME2全体としては5年半近い運用期間となる。このような長期運用でも、TSUBAME2.5の運用の最後の方の時点でも安定な運用と十分な競争力を発揮していることが強く期待される。これはTSUBAME2.0の設計が長期運用に耐える堅牢なものであったことと、中間アップグレードが可能なアーキテクチャであった事に由来する。

TSUBAME3.0は様々な新しいテクノロジを装備し、25-30ペタフロップスと更に大幅な性能向上を果たしながら、TSUBAME2と同程度のサイズと消費電力が達成され、そのために更なる冷却を含む進化が見込まれる。また、ビッグデータ時代のために、I/Oや耐故障性が大幅に強化される。これらの詳細は本ジャーナルの将来の記事で明らかにしていく予定である。

非エバルト法に基づくGPUで加速されたG蛋白質共役型受容体の分子動力学計算

真下 忠彰*† 河内 隆行* 福西 快文** 神谷 成敏*** 鷹野 優***
福田 育夫*** 中村 春木***
*一般社団法人バイオ産業情報化コンソーシアム **独立行政法人産業技術総合研究所 ***大阪大学蛋白質研究所

医薬品の開発では、医薬品の結合するタンパク質の立体構造の理解が欠かせないが、タンパク質の立体構造シミュレーションは、極めて計算量が多い。最も時間のかかる計算は2原子間の静電相互作用の計算である。そこで、静電相互作用を極めて効率よく計算できる新手法（Zero-dipole summation法）を開発し、これを実装したGPUを利用する高速分子シミュレーションソフト（myPresto/Psygene-G）をTSUBAMEに移植した。このGPU版は、いくつかのタンパク質に適用した結果、CPU版MDソフトウェアよりも30倍の高速性能を達成した。これを創薬において重要なターゲットであるG-Protein Coupled Receptor（GPCR）の遮断薬や作動薬の結合時における構造変化の解析に適用し、このタンパク質の動的構造が解析できることが示された。

はじめに 1

新薬の開発には多大なコストが掛かり失敗も多いため、コンピュータによる医薬品設計は利益が大きい。医薬品の約半数は、Gタンパク質共役受容体（GPCR）を標的分子として結合するため、GPCRと医薬品の結合状態（複合体）の計算シミュレーションは、医薬品開発において極めて重要である。GPCRは、体内に数百種類もある重要な創薬標的であるが、そのうち数種類について立体構造が近年、解明されつつある。その結果、作動薬や遮断薬の結合によって、GPCRは独特の構造変化（induced-fit）を示し、その動的な構造変化によって機能を発揮していることが分かってきた。induced-fitやタンパク質の動的挙動を考慮するには、分子動力学（MD）計算を利用した構造サンプリングが有効である。

ところで、膜や溶媒分子を含めたリアルな系におけるタンパク質の立体構造シミュレーションは、極めて計算量が多い。最も時間のかかる計算は、長距離までその影響が届く2原子間の静電相互作用の計算である。そこで、静電相互作用を極めて効率よく計算できる新手法（Zero-dipole summation法：ZD法）を開発し[1-5]、本研究では、これを実装したGPUを利用する高速分子シミュレーションソフト（myPresto/Psygene-G）をTSUBAMEに移植した。このZD法では、遠距離的な静電相互作用を12 Å程度の比較的短い距離でカットオフする一方、遠距離からの効果をカットオフ球上に置いたイメージ電荷で繰り込むため、高精度の計算を可能とすると同時に高い並列性を発揮できる。本研究では、このGPU向けのMDプログラムmyPresto/Psygene-GをTSUBAME上で稼働させ、薬物の結合したGPCRのモデル構造を分子シミュレーションにより構築し、その動的構造を解析した。

電気双極子中和（Zero-dipole summation）法 2

クーロン静電相互作用計算では、精度やMDの安定性等の点から、距離による単純な相互作用カットオフは許されず、適切に考慮された処理が必要になる。我々が目標としたのは、*N*粒子系のエネルギーを高速かつ高精度で、そして実際上のインプリメントが容易で、並列計算に有効であり、さらに物理的アーティファクトの小さい手法にて求めることである。

この目標の実現のため、多くの方法の着目点である「二体ポテンシャル関数値が粒子間距離 r_{ij} と共に減衰する」という、距離のみに焦点を当てた考え方を変え、代わりに、電気的性質、即ち、各電荷 q_i 及びそれらが粒子配置で作る構造に着目する。生命体を担っている環境では、真空中に散在した電荷の体系とは異なって、1つの荷電粒子の周りには、たくさんの分子、イオン等がひしめいていると考えられる．その際、高エネルギー現象は通常は起きない、即ち、電気的相互作用がうまくほぼキャンセルされるように荷電粒子の集団は振るまっていて、その結果、実際は $1/r$ よりも早く遮蔽されている。このような物理的洞察に基づき、我々は、電荷と電気双極子モーメントの中性条件を考え、これを二体相互作用和で扱う新規な方法を開発した[1]。この計算法は、次の2つの戦略からなる：（A）各粒子 i に対する全ての粒子からの寄与を算出する和を、中性部分集合 $M_i \subset N_i$ が存在すると仮定し、M_i に関する和で置き換える（図1）；（B）この新たに定義した和に関するエネルギーを（直接に M_i を算出する事無く）、元のような単純なペア毎和で表わせるように、二体ポテンシャル関数形の方を変える。

*† 現所属：株式会社情報数理バイオ

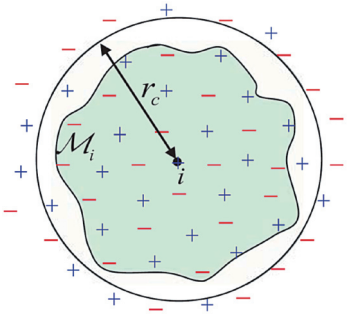

図1　中性条件を表す概念図. 中性条件の満たされた緑色の部分からの寄与のみを取り込む

こうして求められた系の静電相互作用エネルギーは、次のように表わされる[1,2]：

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j(\neq i)\\ r_{ij}<r_c}} q_i q_j \left[u(r_{ij})-u(r_c)\right]-\left[\frac{u(r_c)}{2}+\frac{\alpha}{\sqrt{\pi}}\right]\sum_i q_i^2,$$

$$u(r)=\frac{\mathrm{erfc}(\alpha r)}{r}+\left[\frac{\mathrm{erfc}(\alpha r_c)}{2r_c}+\frac{\alpha}{\sqrt{\pi}}\exp\left(-\alpha^2 r_c^2\right)\right]\frac{r^2}{r_c^2}$$

ここで、α は遮蔽パラメタであり、erfc は補誤差関数である。一般の分子系に適用するためには、多少の修正が必要である[2]。この方法では、全エネルギーが二体相互作用項と定数付加項のみで表されるので、プログラムの移植も簡単であり、高性能計算機においても低コスト化が期待できる。粒子数Nが大きくなり、通常のカットオフ距離に比べ系のサイズが大きくなると O（N）計算が可能になってくる。これは、波数部を含むEwald法等と差別化される。さらに、そもそもEwald法等において仮定される周期境界条件を生体系に用いることのアーティファクトも指摘されている[3]。ZD法では、このような系の厳密な周期性は仮定しない。他方、原子ベースのカットオフ和とスムースな関数を使っていることの恩恵として、シンプルということ以外にも、系の（a）重心の保存、（b）エネルギー（一般には運動方程式の不変量）の保存、という、重要な物理的性質が保証されるため、これらの性質も他手法と差別化される。例えば、particle mesh Ewald法、fast multipole法は、エネルギーを良く近似する方法であるが、その代償を、各々、（a）、（b）にて払う必要がある。ZD法の精度については、既に、種々の系[1,2,4,5]において、十分な値を有することが検証されている。例えば、後述するGPCRの系[4]では、12 Åのカットオフ距離で、エネルギー誤差0.04%の精度を達成した。従って、本章冒頭で述べた目的を全て達成するためには、高効率並列化の実現が課題となる。

空間分割による並列化　3

TSUBAMEは、多数のGPUアクセラレータを搭載した分散メモリ型計算機によって構成されている。したがって、TSUBAME向けのMDは、アクセラレータによる加速を受けつつ並列にプロセスを実行し、プロセス間のデータ交換を高速なネットワーク通信を介して行う、CPU・アクセラレータ複合型の実装となっている（図2）。

空間分割MD（myPresto/Psygene-G）はNVIDIA社製GPUをアクセラレータとした超並列計算機向けに開発されたGPU・MPI複合型の並列プログラムである。このプログラムは、系を構成する原子を座標空間で分割し、その部分空間に所属する原子のMD計算を並列計算機の構成要素（MPIプロセス）に割り当てる。割り当てたMD計算のうち２原子間の相互作用計算はGPUによって計算される。また、この相互作用計算で必要な隣接する部分空間の原子情報の相互伝達、部分空間をまたがる原子の移動、および温度・圧力制御に必要なデータの伝達はMPI通信で実現されている。遠距離的な静電相互作用を12Å程度の短い距離でカットオフできるZDは、空間分割による高速化に適しており、このプログラムに同時に搭載した。

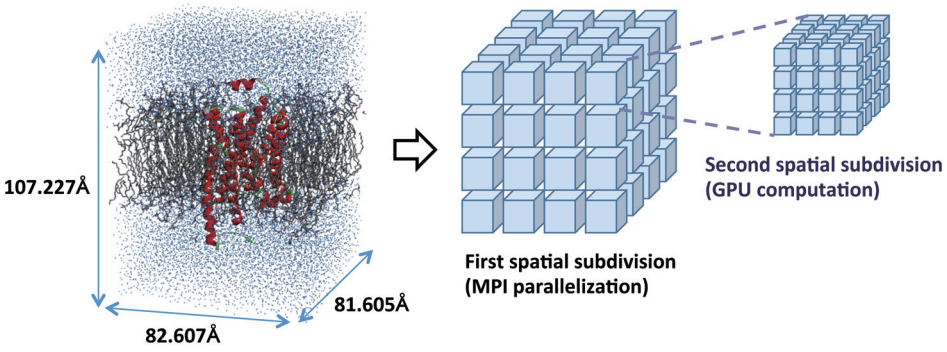

図2　myPresto/Psygene-Gにおける空間分割の概要

非エバルト法に基づくGPUで加速された G蛋白質共役型受容体の分子動力学計算

各セル内での計算 4

MDの静電相互作用とvan der Waals相互作用の二体相互作用計算では、近距離（カットオフ内）にある近傍原子からの影響を考慮した力の計算を行うが、近傍原子の探索には原子数の二乗の計算時間が掛る。この近傍探索を効率化する手法として、空間分割法がある。空間分割MDで採用している空間分割法は、探索空間を等間隔格子で分割し、当該及び隣接する分割領域に含まれる原子間の二体相互作用計算のみ行う事によって、計算量を低減する方法である。

空間分割MDでは、分割した系の構成要素を部分空間（Cell）としている。Cellは直方体の系を三軸に沿って任意の数で分割する事によって得られる。各Cellは領域の境界、構成原子数、各構成原子の情報（座標・速度など）を管理しており、全てのCellを統合する事によって系全体が構成される。また、Cellは並列計算機のCPUコアが担当する計算領域の最小単位でもあり、1Cellは1CPUコア（MPIプロセス）に割り当てられる。

一般にMD計算の1ステップは、力の計算および運動方程式の積分計算に分けられ、力の計算は結合項および非結合項の計算に分けられる。更に非結合項の計算は、van der Waals力項、近距離静電力項および遠距離静電力項の三つの計算に分割される。遠距離静電力項の計算はZD法によって近似計算される。空間分割MDでは、MD計算1ステップの処理は各Cellに割り当てられ、並列に実行される。その際、計算時間の大半を占める二体相互作用計算はGPU上で計算される。

プロセス間通信 5

各Cellで実行される近距離二体相互作用計算では、当該Cellの原子と隣接するCellの原子との相互作用を計算する必要がある。したがって、3次元の空間分割の場合、Cellは周囲26隣接Cellと原子座標の送信・受信を行う必要がある。

空間分割MDでは原子情報をCell間で交換し、かつCell間をまたがる原子の移動を実現するため、MPI通信によるプロセス間通信を行っている。MPI通信では、通信時間低減のために非同期型の一対一通信を優先的に使用している。

表1に、こうして開発したmyPresto/Psygene-Gの、種々の大きさの蛋白質に対するMD計算の演算性能を示す。CPUの列は、GPUを用いない場合にMPI並列化を行った場合の1ステップあたりの演算速度を示す。

タンパク質の系 (溶媒，膜を含む原子数)	セル数	1 ステップの速度 (ms)	
		GPU	CPU
EGFR キナーゼ (38,452)	1	73.28	3631.44
	8	10.57	456.24
β_2 アドレナリン受容体 (56,120)	1	128.05	6151.17
	8	17.21	773.60
アクアポリン-4 (104,414)	1	229.57	14708.82
	8	26.94	1827.38
	27	17.13	368.77
ダイニン (1,004,846)	1	2352.85	248148.26
	27	96.89	9581.55
	64	51.27	2664.80

表1 myPresto/Psygene-GのMD演算性能。

GPCRへの応用 6

GPCRを標的とする医薬品には、GPCRを活性型に変形させ下流のシグナル伝達をうながす作動薬（agonist）、逆にGPCRを不活性型に変形させてシグナル伝達を止める遮断薬（inverse agonist）、その他、遮断薬と作動薬の中間の性質を持つ部分作動薬（partial agonist）があるので、作動薬結合時と遮断薬結合時のGPCR（本研究では、β_2-アドレナリン受容体）の構造変化について解析した。

GPCRは生体膜に埋め込まれた系であり、系の構成原子数は約56,000原子となる。また、薬剤結合時の構造変化は遅い過程であり、最低でも数十nsecの分子シミュレーション計算を行わねばならない。

GPCRの構造は、図3のような模式図で表される、7本のα-ヘリックス構造からできている。GPCRは膜に埋まっており（図では膜、溶媒は省略されている）、上が細胞外、下が細胞内であり、薬物は、GPCRの上から約1/3程度のところに結合する。薬物結合時の、薬物結合ポケット周辺の構造変化はわずかであると考えられているが、GPCRの下部では、その動きは大きくなり、細胞内に信号が伝達される。

そこで、TSUBAME上で我々のmyPresto/Psygene-Gを生体膜に埋め込まれたGPCRの系（膜や溶媒分子を含んで約56,000原子からなる）へ適用し、20～50 nsecに及ぶMD計算を従来の約30倍の速度で行った。この計算により、作動薬、遮断薬の結合時に、タンパク質全体の構造はどうなるかを調べた。

β_2-アドレナリン受容体に薬物をドッキングして生体膜に埋め込み、溶媒を付加して、圧力を1気圧、室温でのMD計算を行った。系は、図4のように緩和され、自然な状態でのβ_2-アドレナリン受容体の運動を観察した。

MD計算では、多数の薬剤の解析を行ったが、ここでは、遮断薬Carazololと作動薬Formoterolの結合した場合のMDシミュレーションの結果を示す。

タンパク質全体の動きを７本ある各ヘリックスのRMSDの時間変化で表したのが、図５，６である。若干、遮断薬結合型の方（図５）が、構造のゆらぎが作動薬結合型（図６）より大きくなっているように見え、化合物の機能によってGPCR全体に構造変化が起こることが観察された。

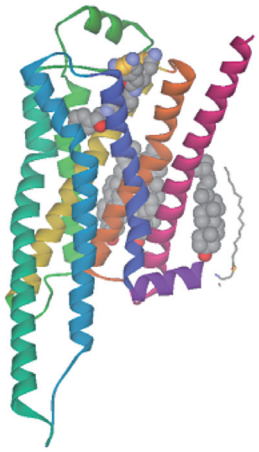

図３
GPCR（β_2-アドレナリン受容体）の構造模式図。異なるα-ヘリックス構造は、色で区別してある。膜分子、溶媒分子は省略。

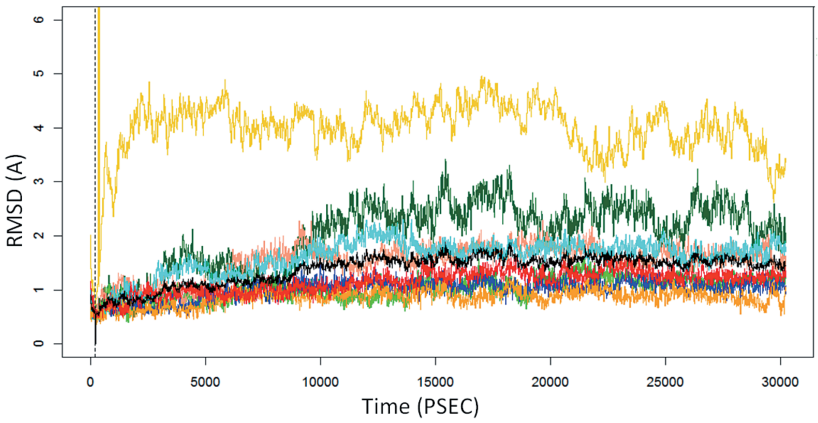

図５
遮断薬Carazolol結合型β_2-アドレナリン受容体の７本の膜貫通ヘリックス（TMH1-TMH7）の動き：薄黄色：ループを含むGPCR全原子のRMSD。
深緑：TMH1、緑TMH2、青：TMH3、濃オレンジ：THM4、
空色：TMH5、黄：TMH6、赤：TMH7、黒：TMH1-TMH7の全原子。

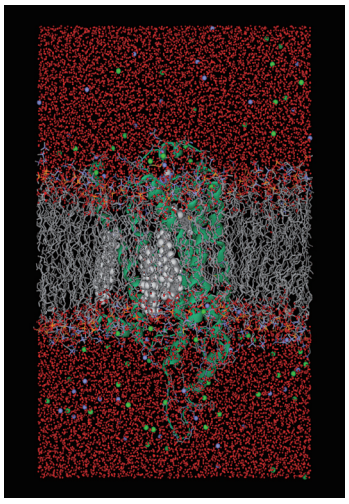

図４
β_2-アドレナリン受容体のMD計算

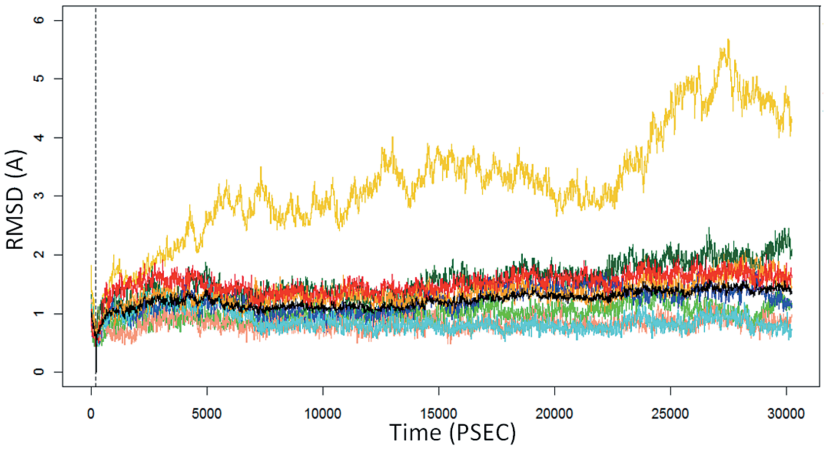

図６
作動薬Formoterol結合型β_2-アドレナリン受容体の７本の貫通ヘリックス（TMH1-TMH7）の動き：薄黄色：ループを含むGPCR全原子のRMSD。
深緑：TMH1、緑：TMH2、青：TMH3、濃オレンジ：THM4、
空色：TMH5、黄：TMH6、赤：TMH7、黒：TMH1-TMH7の全原子。

まとめ 7

分子シミュレーションプログラムmyPresto/Psygene-Gの非結合項の二体相互作用計算を、専用計算機（マルチコアアクセラレータ）であるTSUBAMEのGPUに移植した。GPUに実装された相互作用計算は、CPU実装版による結果と高い精度で一致し、力計算・エネルギー計算の両面において十分な精度を保っていた。GPU版のZD法を導入したmyPresto/Psygene-Gは十分な計算速度を実現し、更にスケーラブルな拡大を可能とする事が確認された。

GPCRに、遮断薬、作動薬などを結合し、生体膜中での系を構成し、MDシミュレーションを行い、遮断薬、作動薬において、GPCRのアミノ酸残基や構造の運動の違いが現れることが示唆された。

謝　辞

本計算はTSUBAMEグランドチャレンジ大規模計算制度の元で平成24 年度に実施させて頂いたもので、学術国際情報センターの方々に深く感謝の意を表する。本研究の一部はNEDO「創薬加速に向けたタンパク質機能解析基盤技術開発」から支援を頂いた。ここに記して謝意を表す。

参考文献

[1] I. Fukuda, Y. Yonezawa, and H. Nakamura: Molecular Dynamics Scheme for Precise Estimation of Electrostatic Interaction via Zero-Dipole Summation Principle, J. Chem. Phys., Vol. 134, 164107 (2011)

[2] I. Fukuda, N. Kamiya, Y. Yonezawa, and H. Nakamura: Simple and Accurate Scheme to Compute Electrostatic Interaction: Zero-dipole Summation Technique for Molecular System and Application to Bulk Water, J. Chem. Phys., Vol. 137, 054314 (2012)

[3] I. Fukuda and H. Nakamura: Non-Ewald methods: Theory and Applications to Molecular Systems, Biophys. Rev., Vol. 4, pp.161-170 (2012)

[4] N. Kamiya, I. Fukuda, and H. Nakamura: Application of Zero-dipole summation method to molecular dynamics simulations of a membrane protein system, Chemical Physics Letters, Vols. 568–569, pp.26–32 (2013)

[5] T. Arakawa, N. Kamiya, H. Nakamura, and I. Fukuda: Molecular Dynamics Simulations of Double-Stranded DNA in an Explicit Solvent Model with the Zero-Dipole Summation Method, PLoS One, Vol. 8, e76606 (2013)

巡回セールスマン問題に対する反復局所探索の大規模並列アルゴリズム

Kamil Rocki　須田 礼仁
東京大学 情報理工学系研究科 コンピュータ科学専攻

複雑な組み合わせ最適化問題を解くために高性能並列アルゴリズムが重要であることは論を待たない。数百万ものヘテロな計算コアを有する計算システムが登場し、今後の計算機にも適したアルゴリズムを開発することは極めて重要である。今のところ、このような問題を解くのに最も有効な手法はメタヒューリスティクスによる近似解探索である。中でも最も汎用的で成功しているアルゴリズムは反復局所探索（Iterated Local Search、ILS）と呼ばれている。これは近似解を徐々に改良してゆくアルゴリズムで、時間を掛けるに従ってよい近似解が得られる。我々は汎用性を犠牲にしたり問題固有の知識を求めたりせず、ILSのもともとの仮定を維持したまま、利用可能な並列性を使えることを示す。本稿で示す並列反復局所探索（並列ILS）アルゴリズムは、高い効率で分散的に組み合わせ最適化を行うことができる。本手法はシンプルかつ高水準であり、従来のILS法で解ける任意の問題に対し、逐次コードを多少修正するだけで適用可能である。巡回セールスマン問題に適用した実験では、本手法は、より手の込んだ局所探索の並列化よりも高性能であった。ノード間並列にMPIを用いた並列ILSは、TSUBAME2.0スーパーコンピュータの256ノードを用いて、逐次計算にくらべて90倍の高速化を達成した。

1 はじめに

組み合わせ問題は情報科学や、人工知能、OR、バイオ情報などの多様な分野で表れる。有名な問題だけでも、最適スケジューリング問題、命題論理のモデルを探すSAT、グラフをたどる巡回セールスマン問題（TSP）、Quadratic Assignment Problem（QAP）などがある。これらの問題には、離散集合の上で、特定の条件や制約を満たす分割・順序・割り当てを探索するという共通の特徴がある。これらは離散変数で解が表現される組み合わせ最適化（CO）と呼ばれるクラスの問題を含む。すなわち、解は有限または加算無限個の整数の集合や、部分集合、組合せ、グラフ構造などである。多くの組み合わせ最適化問題において、解候補の集合は問題サイズに対して指数的である。組み合わせ最適化問題に対して開発されてきた多数のアルゴリズムは、大別して厳密解法と近似解法（ヒューリスティクス）に分けられる。厳密解法は有限サイズの任意の問題に対してある時間で最適解を出す保証がある[1][2][3]。近似解法は構成的解法と局所探索法とにわけられる。構成的解法は、ゼロから要素を追加してゆき、一つの解を完成させる手法である。局所探索法は、ある初期解から出発して、適当に定義された現在の解の近傍の中からよりよい解を選ぶことを繰り返す[3][5]。局所探索法は、十分によい解に到達する前に局所最適解に陥る可能性があるため、膨大な全探索空間を網羅的に探索するには向かない。しかし、高速に局所最適解に収束するので、多数の局所探索解を探索できる。これを打開する新たなアルゴリズムとして現れたのがメタヒューリスティクス[7][8]である。これらの手法では、基本的な発見的手法を高水準の枠組みで組み合わせることにより、探索空間を効率的かつ効果的に探索することを目指す。Ant Colony Optimization（ACO）[11]、Genetic Algorithm（GA）[10]、反復局所探索（ILS）[6]、Simulated Annealing（SA）[9]、Tabu Search（TS）[7]などが含まれる．本論文では，単純ながら強力なメタヒューリスティクス[12][13][16][17]であるILS（図1、図2）に焦点を絞る。

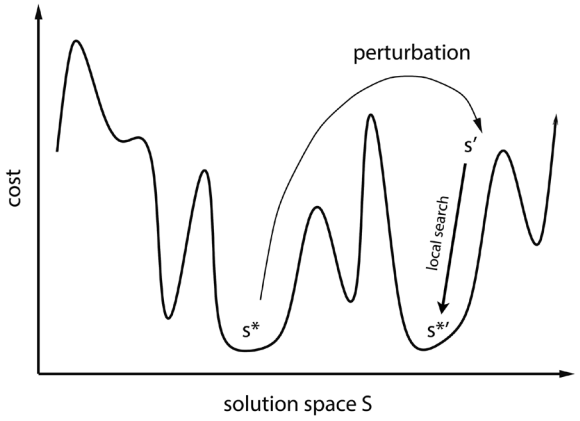

図1　反復局所探索の解空間イメージ

```
1: procedure ITERATED LOCAL SEARCH
2:     s0 := GenerateInitialSolution()
3:     s* := 2optLocalSearch (s0)
4:     while (termination condition not met)
5:         s′ := Perturbation (s*)
6:         s*′ := 2optLocalSearch (s′)
7:         s* := AcceptanceCriterion (s*, s*′)
8:     end while
9: end procedure
```

図2　反復局所探索のアルゴリズム

巡回セールスマン問題に対する反復局所探索の大規模並列アルゴリズム

2 反復局所探索

反復局所探索（ILS）は確率的局所探索（SLS）の一種で、何らかのヒューリスティクスにより近似解の系列を生成するが、そのヒューリスティクスをランダムに用いたよりも格段によい結果を与える手法である[5]。ここでは、局所最適解が相互に隣接しているという暗黙の仮定を置いている。すなわち、最小化の場合、よい局所最小解を探すには、ランダムな初期値から出発するより、小さい値を与える局所最小解から出発する方が容易であるという仮定である。そこで、ILSでは、（1）現在の局所最適解に摂動を加え、（2）修正された解に局所探索を適用する、ということにより局所最適解を次々に得る。摂動は、異なる局所最適解に至るattraction basin に trajectory を移動できるだけの強さが必要であるが、ランダムな初期値と同等なほど強くてもいけない。適切な摂動は、局所探索の手法同様、問題依存である。我々はTSPに対して2-opt moveと呼ばれる局所探索を用いている。

2．1 2-opt 局所探索

2-optアルゴリズムは巡回路から2本の辺を削除し2本の辺を追加する。これにより、部分巡回路が反転して再接続される（図3）。これは通常2-opt move と呼ばれ、巡回路が短くなる限り反復的に適用される。2-opt moveで巡回路が短くならなくなれば局所最適解である。これは大域的な最適解であるとは限らない[10]。そこで局所最適解から脱出する手法、すなわち探索空間の中でより悪い解への移動が必要である。

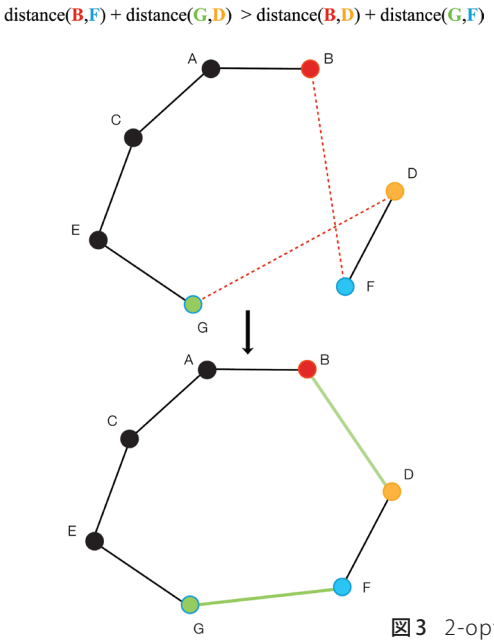

図3 2-opt move

3 並列反復局所探索

反復局所探索の基本的な並列化手法は前述のマルチスタートである。すなわち各スレッドがランダムな出発点から独立に探索するもので、異なるスレッドあるいはプロセスは速度の異なる計算機、さらには異なるアルゴリズムで実行することができる。これが提案手法のキーポイントである。我々は、積極的な通信を行うようにアルゴリズムを改良した並列反復局所探索（あるいは通信つきのマルチスタート局所探索）を提案する。逐次コードは基本的に手を付ける必要はなく、必要なのはスレッドカプセル化、メモリ同期（critical section）の追加、大域的なベスト解の書き込み・読み出しである。また、解とそのコストを共有するためのスレッド間の通信が必要であるが、その実装はシステムに依存する。アルゴリズムの全体をListing 1 に示す。

```
Algorithm :  Multi-start Local Search with Communication
 1: procedure  Parallel Iterated Local Search
 2:     S* <- ∞                          //Best Known Solution
 3:     s^n_0 <- GenerateInitialSolution ()  //Random or Heuristic
 4:     s^n* <- ParallelLocalSearch (s^n_0)
            //All processes
 5:     while termination condition not met do
 6:         if Cost(s^n*') > Cost (S*) then
 7:             s^n* <- S*
            //Read the Best Global Solution
 8:         end if
 9:         s^n* <- Perturbation (s^n*)
10:         s^n*' <- ParallelLocalSearch (s^n*)
    //Intra-thread parallelism (SIMD)
11:         if Cost(s^n*') < Cost (S*) then
12:             S* <- s^n*'      //Update the Best Global Solution - Critical Section
13:         end if
14:         if AcceptanceCriterion (Cost(s^n*')) == TRUE then
15:             break
16:         end if
17:     end while
18: end procedure
```

Listing 1 通信付きマルチスタート局所探索の疑似コード

まず、各プロセスは状態空間の異なる点に対応するランダムな初期解から開始する。続いて、最初の降下を行う。その後、各プロセスが探索と摂動というILSのサイクルを、どれかのプロセスが許容可能な解に至るまで反復する。この間、どれかのプロセスがよい解を見出し次第、他のプロセスに送信する。7行目から12行目までがデータ交換ステップに当たる。これは共有メモリ上で、あるいは分散メモリの場合はMPIで行う。

3．1 共有メモリの場合

ノード内でメモリを共有していれば、大域変数を用いてスレッド間のデータ交換ができる（図4）。単純かつ最も効果的な実装である。

3．2 分散メモリの場合

スレッド並列ではなくプロセス並列の場合は、スレッドとして実行される関数に計算を詰め込む必要がなくなり、コーディングはやや簡単になる。他方で、プロセス間通信が必要になり、よりコードは複雑で、多くの場合遅くなる。我々はMPI（Message Passing Interface）を用いてアルゴリズムを実装し、評価したところ、ノード内は共有メモリ通信を用いてデータ転送し、プロセス間通信には１スレッドのみが参加する実装が最速であった（図5）。

結果 4

我々のアルゴリズムの共有メモリ版をGeForce GTX 680 GPU で実行した。さらに分散メモリ版をTSUBAME 2.0 の256ノード（各ノード3 GPU搭載）まで用いて実行した。性能比較を図６に示す。この図では、時間をx軸に、解の質（値が低いほど結果はよい）をy軸に取ってプロットしてある。期待通り、解はILSにより徐々に改善されてゆく。２つの結果は同等の解が得られるまでの時間を見ることで比較できる。このようにして高速化率を図より、768 GPUすべてを用いて約90倍高速に実行できた。スケーラビリティが完全でない要因としては、ノード間通信の時間およびGPUとのデータコピーの時間が考えられる。

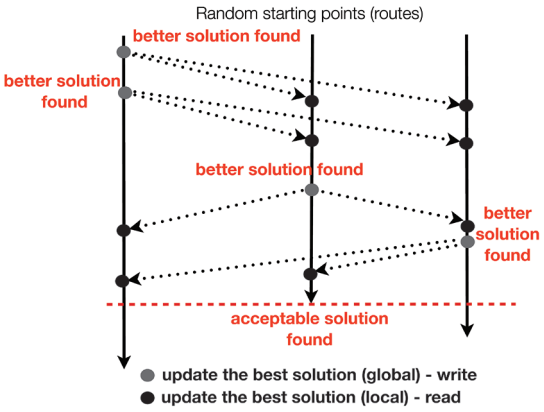

図4　共有メモリ並列化スキーム

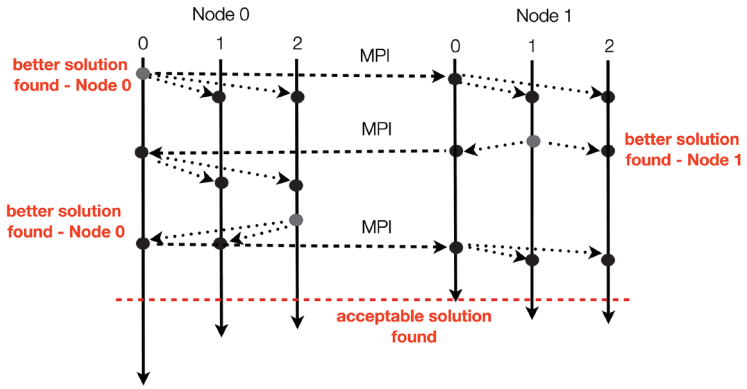

図5　分散メモリ並列化スキーム

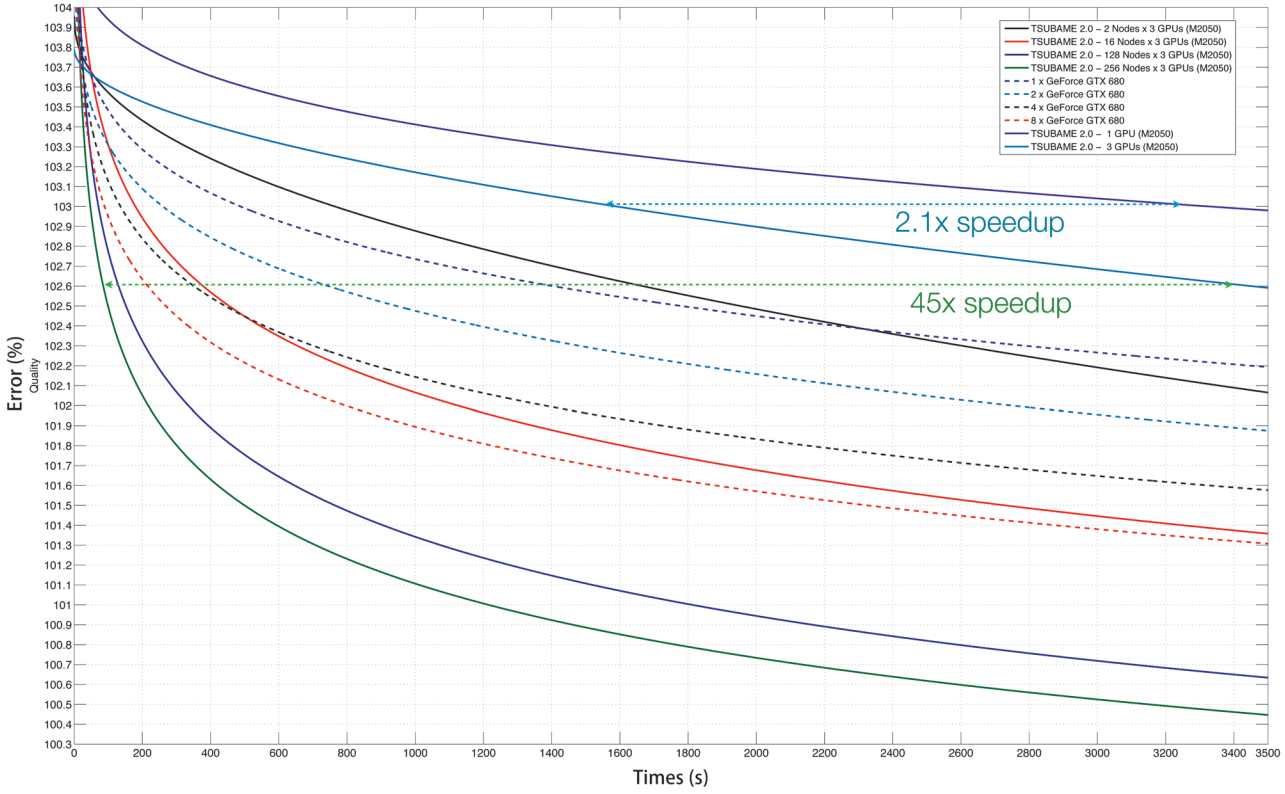

図6 TSUBAME 2.0 を用いた並列ILSの結果

まとめ 5

本論文では、複数GPUを用いた2-opt局所探索による巡回セールスマン問題の高性能実装について論じた。我々の貢献の主たる部分は並列反復局所探索にあり、これは他の局所探索法と組み合わせて使うこともできる。本手法により、TSUBAME 2.0 のような分散GPUシステムを用いて巨大な問題を解くことができる。すなわち、ある意味ブルートフォースなやり方であるが、非常に高い並列性を実現でき、全体としての速度により巨大なTSP問題に適用が可能である。性能評価においては、TSUBAME 2.0 の単一ノードのGPU実装に比べて、単一探索操作の所要時間を1/90に減らすことができた。我々の手法は強スケーリング性を有すると考えられるが、ネットワークおよびCPU-GPU 間通信により制限される。

謝辞

本研究の一部は、JST CREST および文科省科研費の支援を受けています。

参考文献

[1] Applegate, D.L., Bixby, R.E., Chvatal, V., Cook, W.J.: The Traveling Salesman Problem: A Computational Study. Princeton University Press, Princeton (2007)

[2] Lawler, E.L., Lenstra, J.K., Rinnooy Kan, A.H.G., Shmoys, D.B.: The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization. Wiley, Chichester (1985)

[3] Johnson, D. and McGeoch, L.: The Traveling Salesman Problem: A Case Study in Local Optimization. Local Search in Combinatorial Optimization, by E. Aarts and J. Lenstra (Eds.), pp. 215-310. London: John Wiley and Sons, 1997.

[4] Garey,M.R.andJohnson,D.S.ComputersandIntractability:A Guide to the Theory of NP-Completeness. San Francisco: W.H. Freeman, 1979.

[5] Croes G. A.;A Method for Solving Traveling-Salesman Problems, Operations Research November/December 1958 6:791-812;

[6] M. A. O'Neil, D. Tamir, and M. Burtscher.: A Parallel GPU Version of the Traveling Salesman Problem. 2011 International Conference on Parallel and Distributed Processing Techniques and Applications, pp. 348-353. July 2011.
[7] Dorigo, M. and Gambardella, L.M.: Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem. IEEE Transactions on Evolutionary Computation, Vol. 1, No. 1, pp. 53-66. April 1997.
[8] Fujimoto, N. and Tsutsui, S.: A Highly-Parallel TSP Solver for a GPU Computing Platform. Lecture Notes in Computer Science, Vol. 6046, pp. 264-271. 2011.
[9] Reinelt, G.: TSPLIB - A Traveling Salesman Problem Library. ORSA Journal on Computing, Vol. 3, No. 4, pp. 376-384. Fall 1991.
[10] Rego,C.and Glover,F.: Local Searchand Metaheuristics. The Traveling Salesman Problem and its Variations, by G. Gutin and A.P. Punnen (Eds.), pp. 309-368. Dordrecht: Kluwer Academic Publishers, 2002.
[11] Lourenco, H. R. Martin, O. C. Stutzle, T.: Iterated Local Search, International series in operations research and management science, 2003, ISSU 57, pages 321-354
[12] Karp, R. Reducibility among combinatorial problems: In Complexity of Computer Computations. Plenum Press, pp. 85- 103. New York, 1972
[13] Tsai, H.; Yang, J. Kao, C. Solving traveling salesman problems by combining global and local search mechanisms, Pro- ceedings of the 2002 Congress on Evolutionary Computation (CEC'02), Vol.2, pp. 1290-1295.
[14] Pepper J.; Golden, B. Wasil, E. Solving the travelling sales- man problem with annealing-based heuristics: a computational study. IEEE Transactions on Man and Cybernetics Systems, Part A, Vol. 32, No.1, pp. 72-77, 2002
[15] NVIDIA CUDA Programming Guide http://docs.nvidia.com/cuda/index.html
[16] Helsgaun, K.; An Effective Implementation of the Lin-Kernighan Traveling Salesman Heuristic, European Journal of Operational Research, 2000, vol 126, pages 106-130
[17] Nilsson, Ch.; Heuristics for the Traveling Salesman Problem, Linkoping University, pages 1-6

● **TSUBAME e-Science Journal vol.10**
2013 年 11 月 18 日　東京工業大学 学術国際情報センター発行 ©
ISSN 2185-6028
デザイン・レイアウト：キックアンドパンチ
編集：　TSUBAME e-Science Journal 編集室
　　青木尊之　ピパットポンサー・ティラポン
　　渡邊寿雄　佐々木淳　仲川愛理
住所：　〒 152-8550 東京都目黒区大岡山 2-12-1-E2-6
電話：　03-5734-2085　FAX：03-5734-3198
E-mail：tsubame_j@sim.gsic.titech.ac.jp
URL：　http://www.gsic.titech.ac.jp/

TSUBAME

vol. 10

TSUBAME 共同利用サービス

**『みんなのスパコン』TSUBAME共同利用サービスは、
ピーク性能 5.7PFlops、18000CPUコア、4300GPU搭載
世界トップクラスの東工大のスパコンTSUBAME2.5を
東工大以外の皆さまにご利用いただくための枠組みです。**

課題公募する利用区分とカテゴリ

共同利用サービスには、「学術利用」、「産業利用」、「社会貢献利用」の3つの利用区分があり、さらに「成果公開」と「成果非公開」のカテゴリがあります。
ご利用をご検討の際には、下記までお問い合わせください。

TSUBAME 共同利用とは…

他大学や公的研究機関の研究者の**学術利用**［有償利用］

民間企業の方の**産業利用**［有償・無償利用］

その他の組織による社会的貢献のための**社会貢献利用**［有償利用］

共同利用にて提供する計算資源

共同利用サービスの利用区分・カテゴリ別の利用課金表を下記に示します。TSUBAMEにおける計算機資源の割振りは口数を単位としており、1口は標準1ノード（12 CPUコア, 3GPU, 55.82GBメモリ搭載）の3000時間分（≒約4ヵ月）相当の計算機資源です。
1000 CPUコアを1.5日利用する使い方や、100 GPUを3.75日利用する使い方も可能です。

利用区分	利用者	制度や利用規定等	カテゴリ	利用課金（税抜）※
学術利用	他大学または研究機関等	共同利用の利用規定に基づく	成果公開	1口：105,000円
産業利用	民間企業を中心としたグループ	「先端研究基盤共用・プラットフォーム形成事業」に基づく	成果公開	トライアルユース（無償利用）
				1口：105,000円
			成果非公開	1口：391,000円
社会貢献利用	非営利団体、公共団体等	共同利用の利用規定に基づく	成果公開	1口：105,000円
			成果非公開	1口：391,000円

※　平成 25 年度の利用課金です。最新の利用課金については、下記 URL をご参照ください。
http://www.gsic.titech.ac.jp/node/58

産業利用トライアルユース制度（先端研究基盤共用・プラットフォーム形成事業）

東工大のスパコンTSUBAMEを、より多くの企業の皆さまにご利用いただくため、初めてTSUBAMEをご利用いただく際に、無償にてご試用いただける制度です。

（文部科学省 先端研究基盤共用・プラットフォーム形成事業による助成）

詳しくは、下記までお問い合わせください。

お問い合わせ

●東京工業大学　学術国際情報センター　共同利用推進室
●e-mail　kyoyo@gsic.titech.ac.jp　Tel. 03-5734-2085　Fax. 03-5734-3198
詳しくは http://www.gsic.titech.ac.jp/tsubame/をご覧ください。

GSIC
Global Scientific Information and Computing Center