e-Science Journal
東京工業大学 学術国際情報センター
Vol. 2

TSUBAME ESJ.

TSUBAME 2.0始まる
TSUBAME 1.0から2.0への長い道のり（前編）

次世代気象モデルのフルGPU計算
— TSUBAME 2.0の3990GPUで145TFlops —

MEGADOCKによるタンパク質間相互作用予測
～システム生物学への応用～

TSUBAME 2.0始まる

TSUBAME 1.0から2.0への長い道のり(前編)

松岡 聡*
*東京工業大学　学術国際情報センター

2010年11月、日本初のペタコンTSUBAME 2.0がいよいよ稼働する。
しかしその前任のTSUBAME1より更に前に始まる技術的道のりは平坦ではなかった。
前編では、まずTSUBAME1のスパコンとしての技術的特徴を述べ、
そこからTSUBAME 2.0へ継承された利点、並びに改善された欠点、
更にどのように2.4ペタフロップス・30倍もの性能向上が僅か4年半で達成されたかを論じる。

はじめに 1

2010年10月初旬—かつては本センターの基盤総務で多くの人々が事務作業をしていた—階の部屋のドアを開けると、今は別世界の風景が目に飛び込んでくる（図1）。TSUBAME 1.0の冷却の暴力的な騒音に慣れた耳にとっては遥かに抑えられた、しかし確実な存在感を伴い、最新の密閉式の水冷冷却ラックを空気が循環する音が聞こえてくる。TSUBAME 2.0の奥行きの深いラックが並び立つその光景は一見はまるで事務倉庫のようで、とても世界最先端の技術を結集した日本最速のスパコンを内包しているようには見えない。しかし、そのラックのドアを開けると、TSUBAME 2.0用に新開発された計算ノード群が理路整然と並び、高々数本のケーブルがそれぞれから突出しているのが見える。大型の冷蔵庫のようなラック僅か一本に内包される総合性能は50テラフロップスで、僅か8年前に世界一位だった、大型の体育館のような施設全体を占め、600ラック以上が都市の高層ビル群のように林立していた地球シミュレータと同程度の性能である。TSUBAME 2.0全体では総合性能2.4ペタフロップス—2010年の時点では我が国の全ての公共機関のスパコン性能を全て合算したより高速—であるが、その高密度さは自らが技術的にそう意図した所によるものであることが分かっていながら、実際に目にするとやはり感動を憶える。

11月の稼働を前に、既にTSUBAME 2.0の訪問・見学は後を絶たない。しかし、単にマシンを見ただけではその中身がわからないし、仮に分かってもそれだけでは不十分だ。スパコンとして、何故そのように設計されているのか—何故個々のコンポーネンツが選択されているのか、何故それがそのように組み合わさっているのか、TSUBAME1や他のスパコンの経験がどのように技術的に生かされているか—は必ずしもはっきりしない。ユーザにとって、あるいは今後のスパコンの発展の為には、これらの点が明確になり、はたしてTSUBAME 2.0でそれらに纏わる種々の技術的な課題がきちんとクリアできたか、評価する必要がある。単なる安定運用だけでなく、そのような科学技術的先端性の評価がされ、それが公知にならなければ、国立大学の全国共同利用センタースパコンとして30億円以上の税金が投入される価値はない。

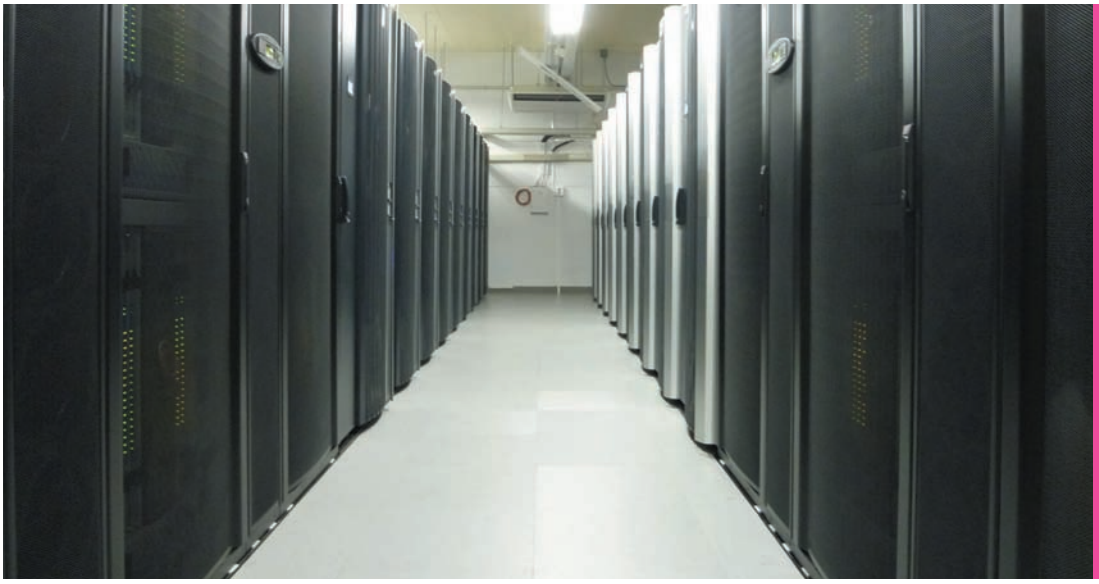


図1
GSIC105号室：
TSUBAME 2.0
計算機ルーム

それらの点で、TSUBAME 1.0 のスパコンの立ち位置は日本のそれとしてはユニークであるかもしれない。特に、GSICセンターとしてはメーカー製品のスパコンを単純にカタログから買うわけではなく、地球シミュレータや後のT2K と同様メーカーと共同の「開発型」の調達を行っている。のみならず、常に ①研究室での方式やシステムソフトウェアの基礎研究⇒②GSICセンターにおける実験運用とそのアセスメント⇒③プロダクションマシンとしての開発・仕様決定と①②へのフィードバックという、ウォーターフォール開発モデルの段階を常に踏んできた。TSUBAME 1.0 における①は松岡研におけるクラスタのシステムの基礎研究や、他の東工大の研究室におけるクラスタのアプリケーション基礎研究であり、②は東工大キャンパスグリッドにおける4年間に渡る合計400ノード以上に及んだクラスタ群での実験運用である。それらの経験を元に③であるTSUBAME1 の仕様が決定された。TSUBAME 2.0 においては、①基礎研究はTSUBAME1 を含む種々の資源で行うことが可能となり、②は後述のようにTSUBAME1 に実験用設備を増設してかつユーザにも開放し、一体実験運用することで実現した。その結果、③後継機種としてのTSUBAME 2.0 の仕様が決定した。今後はTSUBAME 3.0 においてTSUBAME 2.0 が②の役目を担う番である。

前号では、TSUBAME 2.0 のそれぞれの構成要素の特徴を述べた[1]。本稿では、前後編構成でその裏にある技術的な理由づけを述べる。前編ではまずTSUBAME1 の設計と、その経験がTSUBAME 2.0 の設計に影響したかを手短に述べる。後編は、ペタスケールからエクサスケールに至る最新の技術動向が、そのプロトタイプとしてのTSUBAME 2.0 にどのような技術要件として反映されたかを述べる予定である。より完全な技術紹介とその評価は、TSUBAME 2.0 稼働後に独立した本の刊行を要求するだろうが、ひとまずTSUBAME 2.0 始動にあたり各人の参考になれば大変幸いである。

TSUBAME 1.0を振り返って 2

TSUBAME 2.0 の前任であるTSUBAME 1.0（図2）は2006 年4 月に産声を上げた。東工大における長年にわたるクラスタ計算機の種々の研究と、GSIC のスパコンセンターとしての経験を生かした東工大キャンパスグリッド（2002 年4 月~2006 年3 月）プロジェクトの種々の運用実験の成果を元に、各種設計が行われ、約80 ペタフロップス・655ノード・10,000CPUコア・21テラバイトのメモリ・1.1ペタバイトのハードディスク・それらを全てInfiniBand で相互結合・360 枚（後に648 枚に増設）のClearSpeed アクセラレータによる密行列演算系の加速、等種々の特徴を兼ね備え、後継機種のテンプレートとなった。特に、当時各ノードにおいて、1ノード2~4CPUコアでメモリが高々1~4GByte が主流だった時代に、Sun x4600 と、その内蔵するAMD Opteron の800シリーズのマルチソケットサーバ用CPU の特性を生かして、各ノード16CPU コア・32GByte メモリ・（80+80）ギガフロップスのピーク性能・最大20ギガbps のネットワークと、いわゆる「ファットノード」構成を実現したのは、TSUBAME1 のスパコンとしての安定性・利便性のみならず、他の環境で実行できないアプリを実行できる「スパコンならでは」の特徴として大変有利なものであった。2010 年の今でもスパコンのノード計算機として十分普通に通用していることがその証である。

これらの特徴を元に、TSUBAME 1.0 は、我が国のスパコンとして初めてTop500 においてかつての世界王者だった地球シミュレータを破り、アジアNo.1 のスパコンとして1 年半君臨した[2]。また、東工大内外から、企業を含み2000人近いユーザが集まり、TSUBAME 1.0 は人気のある「みんなのスパコン」として多様なユーザ層に大いに活用され続けた。

それらの点から、TSUBAME1 は客観的に鑑みても「成功」であるとみなして差し支えないであろう。Top500などの「グランドチャレンジベンチマーク性能」もさることながら、その高性能・大容量のみならず従来と専用設計のスパコンと比較して、スパコンとしての諸性質を保ちながら遥かに使いやすい環境を実現した。これによりユーザ数の大幅な増加・高い利用率だけでなく、GSIC および東工大に数々のメリットをもたらした。文部科学省Global Center of Excellence プログラムの一つである「計算世界観」の採択、国立大学法人の所謂スパコン基盤センター連合である「学際大規模情報基盤共同利用・共同研究拠点」の構成センターとしての旧帝大系以外からの初めての参加、企業利用を促進する「先端研究施設促進事業」の採択とそれによる50近い有名企業からの利用、米国マイクロソフト社やNVIDIA 社からの" Center of Technical Innovation "，" CUDA Center of Excellence "の認定を含み、数々の学術的成果や栄誉が国内外に認定されたのである。

TSUBAME 1.0の4年以上に渡る運用経験は、特に我々やユーザが大規模ベンチやアプリで多大な負荷をかけていた時だけでなく、通常の数百人ものユーザが同時利用している平常時まで、慎重にかつ網羅的に記録され、TSUBAME 2.0 の設計に繋がってきた。以下はTSUBAME1 において「上手くいった点」「上手くいかなかった点」の列挙である。当然ながら上手くいった点はTSUBAME 2.0 に引き継がれており、一方上手くいかなかった点はTSUBAME 2.0 における解決課題となった。

図2 TSUBAME1 計算機ルーム

TSUBAME 1の光 3

上記に述べたようなTSUBAME1 の成功を語るのは短い本稿内では困難ではあるが、TSUBAME 2.0 へ引き継がれる技術的成功要因を挙げることはまず必要だろう。これは以下の点を含む：

(ア) **高性能x86プロセッサ・Linux等業界標準技術の大幅な採用：** 莫大なPCおよびサーバの産業的「エコシステム」に支えられたPCクラスタのメリットは各所で語りつくされている：時勢に合わせた（Just-in-Timeな）高性能・低コスト・（大量生産による）品質の安定性は多く語られているが、それに加え研究室の通常のPC・ワークステーション・クラスタや、他の多数の同型のクラスタスパコンとのソフトウェアの連続性が大きなメリットである。研究やそれを支えるシミュレーション等のソフトウェアの複雑化により、単一のスパコンのみでソフトを開発し利用するシナリオは激減しており、むしろ多様な環境で同一のソフトが連続に動作し、それらを使い分ける事が日常である。また、初心者が通常のPCの環境から段階的にステップアップすることがスパコン利用層の増大により強く求められる。これらを実現する「みんなのスパコン」はTSUBAME1 の最も重大な特質の一つであり、多くの標準技術の採用により実現した。

(イ) **多プロセッサ・ファットノードのスパコン的アーキテクチャのx86 CPUによる実現：** しかしながら、スパコンはそれならではの利用価値がなければ意味がない。特に、高速性のみならず大容量の問題が解けることがスパコンの大きな魅力となる。このため、システム全体で多プロセッサ・高メモリ容量等であるだけでなく、内部でメモリを共有する個々のノード単位でも通常のPC クラスタ以上に多プロセッサ・高メモリであることが望ましい。過去には多くのスパコンが専用・高額なプロセッサおよび専用の筐体内メモリシステムを用いてそれらを実現していたが、TSUBAME1 では最新のx86プロセッサの共有メモリの新技術を用いることにより、それらと同様の構成を実現できた。これによりユーザにとってのメリットに加え、パーツ数の減少による信頼性等の向上などがもたらされた。更に業務用機関サーバに用いるレベルの電源・ファンの多重化や、ノードあたり数10にものぼる各種モニタリング用センサーおよび監視ネットワーク、更には当時としては高効率な冷却や省電力等の技術要員をベースに運用体制を構築することによって、一万以上のCPUコア・80テラフロップス級のスパコンの安価な構築が可能となり、その後その技術や設計手法は米国テキサス大学TACCのRangerやドイツJulichスパコンセンターのEuropaクラスタ、更には我が国では東大・筑波大・京大のT2K スパコンに引き継がれた。

(ウ) **密行列計算のためのアクセラレータ：** 既にTSUBAME1 の時点で、100テラフロップス近い性能を当時の技術で、コスト・電力・スペースなどの制約内に収めるのは困難であった。そこで、以前より松岡研究室で研究していたアクセラレータ技術を採用する検討する運びとなった。幾つかの候補を検討し、結局英国ClearSpeed 社が開発したSIMD型の密行列演算用のアクセラレータを導入した。これにより大型の密行列のBLASライブラリ利用時にはコマンドラインのスイッチの指定のみで倍近い性能をユーザが得ることが可能となった。当然Linpackの性能向上にも大幅に寄与したが、そのためには新たなハイブリッド型のアルゴリズムを研究開発する必要があった[3]。

(エ) **InfiniBandと大型スイッチによるファットツリー 型ネットワークの構成とI/Oネットワークの統合：** スパコンのもう一つ重要な要素は、ノード間を接続する高速なネットワークである。特に、単に一つのノードのバンド幅だけでなく、そのレーテンシが通常のネットワークと比較してマイクロ秒単位と著しく低いことや、全体が通信する際のバイセクションバンド幅が高いことが求められる。当然、しばしば相反する性質として同時に低コストや高信頼性が求められる。TSUBAME1 のネットワークは288ポートの大型InfiniBand のスイッチが2 階層・8台（下位層6 台、上位層2 台）で構成された。これにより個々のノードから20Gbpsとスパコンレベルの高バンド幅が安価に実現できただけでなく、対称性の高いネットワークにより柔軟な運用が可能となった。例えば一台ノードが故障しても、他のノードに性能的影響を及ぼさず代わりのノードを用いることが可能であった。またレーテンシも5マイクロ秒程度と、専用のスパコンのネットワークと同等となった。

(オ) **高機能ストレージにおける高密度・高バンド幅・並列ファイルシステムによる高性能達成：** スパコンでしばしば忘れられがちなのはストレージ部分であるが、TSUBAME1 では100テラフロップス級のシミュレーションおよび処理能力により、サブペタバイト級の莫大なデータを扱う必要性が生じていた。これにより、ストレージも計算ノード全く同様に、高いスケーラビリティと並列性・高信頼性・低コスト・低消費電力化が求められ、旧来のエンタプライズ系のIT技術を基盤としたストレージシステムでは不十分であった。そこで、48台のHDD と強力なコントローラ兼ストレージサーバを4Uのシャシーに内蔵するSun x4500 "Thumper"（一台あたり24テラバイト）42台を中心としたストレージ構成とし、ストレージネットワークも計算ノード間のInfiniBandに直接接続する形で各Thumperと計算ノードとの間での超高速なデータのやりとりを可能とし、その上にLustre並列ファイルシステムを載せ、全体数百テラバイトの容量、10ギガバイト/秒以上のI/O 性能を可能とした。

(カ) **数百名の同時利用を前提とした分かりやすく公平でCapacityとCapabilityを両立させるバッチスケジューラ：**TSUBAMEは2000名のユーザがアカウントを持ち、同時に100名以上のユーザがジョブを走らせている。年間のジョブ数は百万件以上にも及び、その利用形態―ジョブのサイズ・個数・並列性・実行時間・QoS要求・I/O性能・等々―は本当に様々であるし、ユーザも初心者から多くのスパコンを使いこなすエキスパート、利用もプロジェクトも個人レベルの小さいものから多人数で多くの資源を要求するものまで様々である。これらの多岐にわたる要求を公平に満たすにはTSUBAME1と言えどとてもとても容量は足りない。そこで、定額制と従量制の両立・高額な支払いにより高QoS確保手段の提供、更には大規模ジョブの為の予約制の導入、などの種々の機能を、ユーザに分かりやすくかつ公平感のある形で実現するバッチスケジューラの開発を行なった。この開発は運用と連携し、ユーザのフィードバックを得て完成度を高めていった。

これら以外にも、種々の技術的や運用上の工夫があった。特に毎年ソフトウェア・ハードウェア等種々の見直しを行い、追加調達を行ったが、それはアプリのソフトウェアライセンスのように単に運用上不足していると判明したものだけでなく、②の将来に向けた運用実験のものも含まれていた。

TSUBAME 1の影と TSUBAME 2.0に向けての技術的進歩 4

TSUBAME1上のアプリケーションの実行性能や運用上で判明した諸問題で、やはり上記のような小手先の改良だけでは改善できない点も種々明確になってきた。のみならず、スパコンの年率180%という性能向上の必要性と、それを担保する諸技術の研究開発により、TSUBAME1の設計時には最良の選択でも、2010年のTSUBAME2の稼働時は必ずしもそうならない点もあらわになってきた。以下にそれ等の点を挙げる：

(ア) **性能向上に対する消費電力の問題：**TSUBAME 1.0はピーク時に1MW程度の電力を消費した。これは東工大大岡山キャンパスの全体の電力の10%以上にあたり、電気代は年額一億円を超える。2010年TSUBAME 2.0では初期の目標性能は1.8の4乗~=10倍強で、約1ペタフロップであった。しかしながら、その為には電力性能比も10倍にしなくてはならない。既にClearSpeedを用いていたTSUBAME1は時代のスパコンとして高効率であった；その証拠として二年後に構築され、アクセラレータを用いない東大T2Kが同程度の消費電力でLinpack性能が高々2割程度しか高くないことからもわかる。よって、TSUBAME1並みの国際的な競争力を維持するには、大幅な電力性能比の向上が必要となった。幸い、JST-CRESTの「情報システムの超低消費電力化を目指した技術革新と統合化技術」に「ULP-HPC: 次世代テクノロジのモデル化・最適化による超低消費電力ハイパフォーマンスコンピューティング(ULP-HPC)」という研究課題で応募し、採択されることによって目標達成のための①基礎研究を進めることが外部資金で可能となった。ULP-CRESTにおける電力性能比向上は10年で1000倍であり、4年半だと約24倍となる。

(イ) **アクセラレータの広範かつ柔軟な適用性および性能の問題：**関連するが、アクセラレータも無論万能ではなく、特にClearSpeedはライブラリ経由での密行列の演算には力を発揮したものの、その他の演算ではプログラミングが困難・メモリバンド幅不足・メモリ容量不足等の利用で、他の用途への応用がほとんどなされない状態であった。幸い、バンド幅や性能が高く、汎用性も高く、かつ業界標準品であるが故安価なGPUが台頭し始めており、松岡の研究室や他の研究室でもそのHPCへの利用法の基礎研究が既に行われていた。よって、2008年頃には次期TSUBAME用にGPUを用いるかの②実験運用が行えるかが焦点となっていった。幸い、2007年頃よりNVIDIA社およびMicrosoft社等との技術パートナーシップでGPU活用の研究プロジェクトおよび共同研究が開始され、2007年末にはプロトタイプの128GPUのクラスタ計算機を、2008年10月にはTSUBAMEに680機の最新のNVIDIA Tesla GPU(図3, 図4)を接続し、TSUBAME 1.2として種々の運用実験を行った。

図3 TSUBAME1で使用したNVIDIA Tesla s1070

図4 TSUBAME2で使用するNVIDIA Tesla M2050, Thinノード1台に付き3機搭載

(ウ) **ノード内メモリおよびネットワークバンド幅の不足：**また、GPUに限らず、システム全体としてのバンド幅が全体的にTSUBAME1 では不足していた。メモリバンド幅に関しては二つの要因があり、多チャンネルメモリを大量に搭載したことによるメモリバスのクロック低下と、用いたCPUの共有メモリのハードウェア的要因による全体の総合メモリバンド幅の制限（ノード全体で20GB/s）である。また、ネットワークに関しては、InfiniBandの性能が各チャンネル1GB/s 強あるが、他のノード内のアクティビティと重なると半分に低下する現象が見られた。次回に譲るが、これらのバンド幅の大幅な向上は、単に現在のアプリケーションの高速化だけでなく、今後のシステムのスケーラビリティ確保のために重要な要因である。よって、TSUBAME 2.0 の設計では、FLOPS 向上を超えるバンド幅の向上が最重要課題となり、それがノード計算機の新開発の一つの大きな事由となった。

(エ) **ネットワーク全体のバイセクションバンド幅の不足：**さらに、ネットワーク全体でもバイセクションバンド幅は、エンドポイントが13テラビット秒であるのに対し、2.8 テラビットと1/5 程度であり、FFT のような大規模な全対全通信においては大きなボトルネックとなっており、地球シミュレータ等と比較して陰解法系が弱点となっていた。よって、より全体で高バンド幅なネットワークを実現することが急務となった。このため、地球シミュレータ同様のフルバイセクションバンド幅を実現するネットワークが必要となったが、ノード数が1500 程度と地球の2 倍以上に達するため、技術的な困難さが問題となった。種々の検討の結果、集中型の大規模スイッチと小規模スイッチを複合的に用い、中央に集中したスイッチに対しては下位スイッチから全面的に光ファイバ接続を用いることで解決の目途を立てた（図5）。

(オ) **冷却効率の問題：**TSUBAME 1.0 は冷却列と暖気列の分離を省スペースで実現するなど、2006 年当時としては最新の冷却技術を実現し、PUE 値（冷房の効率を現わす値・1 が理論的に最高で、2だとマシンと冷却電力が同等。旧来のデータセンターでは2を超えるものも多かった）として約1.44 を達成していた。しかし、マシン電力1 に対し冷却電力が0.44 も必要なのは近年の冷却技術としては最善ではない。よってPUE を少しでも1 に近づける技術の検討が急務となった。そのため、水冷や密閉式の空冷と水冷のハイブリッド、自然大気冷却など、幾つもの方式を検討し、1.2 台のPUE の実現を目指した。

(カ) **マシンサイズおよび重量の問題：**TSUBAME 1 は全体で80ラック近くの大きさで、その設置にGSIC 情報棟の計算機室のほとんどの床面積を占めてしまっていた。これでは拡張性がないのみならず、上下二階に多くの配線や管理が跨り、非効率であった。そこで、性能密度を大幅に向上させ、より小さいマシン作りが重要な課題となった。これはマシンのコスト低減がメリットだけではない；上記のフルバイセクションネットワークを実現するためには、なるべく多くのノードを中央スイッチ近接に設置できなくてはならないが、その為にはマシンが小さくなくてはならない。幸い、GPUの効率的な装着、および内部バンド幅向上など、様々な技術要件からノードを新設計することになり、今までにないレベルの高密度なノード実装が大きな技術目標となった（図6）。結果として、一ラックあたり50テラフロップスという性能密度が実現でき、TSUBAME 2.0 の総合ラック数は60 程度とTSUBAME1 の約3/4 となった。

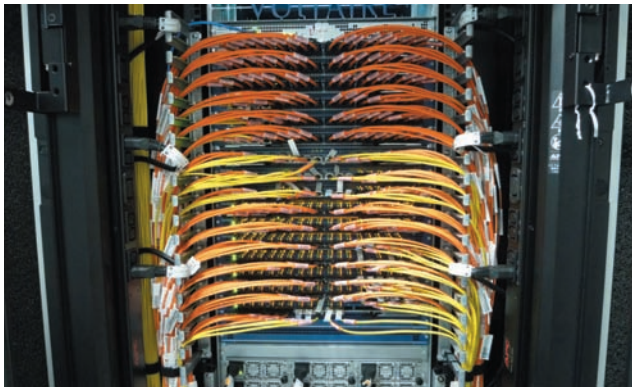
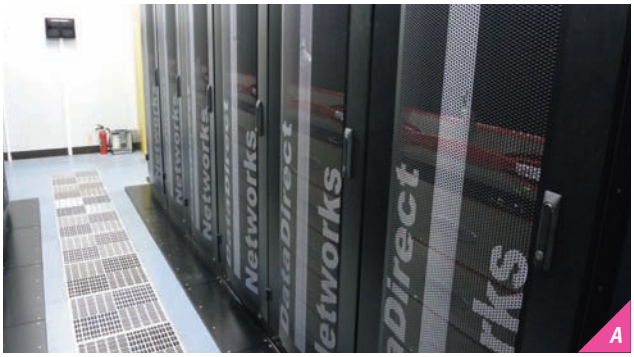


図5 下位スイッチからの光ファイバが集約されるInfiniBand コアスイッチ

図6 TSUBAME1 に比べ、サイズが1/4 となったTSUBAME2 計算ノード



A

B

A 図7 TSUBAME 2.0 のLustre用ストレージサーバ群

B 図8 ディスクエンクロージャ内部には60台の2TB SATA HDD

C 図9 GSIC 国際棟（TSUBAME2 のある情報棟とは別棟）にあるテープライブラリ

D 図10 TSUBAME2 計算ノードには60GB、または120GBのSSDを2機搭載

C

D

(キ) **ストレージの運用容量の不足：**ストレージはTSUBAME1で大幅な容量向上を果たしたものの、実際の運用容量は常に不足気味であった。これは様々な要因による。まずは三次記憶としてのテープシステムやMAID がなく、全て高性能ストレージであった為、バックアップも全て高額な高速ストレージに行う必要があった。また、ストレージ、特にLustre 並列ファイルシステムがsingle point of failure とならないよう多重化する必要性が生じたが、メタデータ管理の為に折角のThumper が消費されてしまう事態が生じた。同様のThumper の消費は他の理由でも起こり、またストライプ数が少ないRAID6 構成を余儀なくされたため、ユーザがLustre で高性能に使える領域は全体から見れば極小となってしまっていた。このため、2007 年にはさらにThumper を20 ユニット追加し、総合物理容量を1.6 ペタバイトに引き上げたが、それでもLustre の通常の運用領域はGaussian のスクラッチ用スペースを含めても200テラバイト程度の実容量であった。そこで、TSUBAME2 ではLustre を含むストレージの運用管理を効率化し、物理容量に対する実運用容量を大幅に増加させる設計として、同時に性能や信頼性を向上させることとした。特に、TSUBAME1 の2009 年以来の運用体制を発展させ、ストレージの管理サーバを専用化・多重化し、HDD のRAID 構成なども効率化し、かつ最大10ペタバイト以上に容量拡張できるテープ装置も別途導入した。（図7, 図8, 図9）

(ク) **ストレージのバンド幅の不足：**更に、後の目標性能である2-3ペタフロップス達成のためには、数百ギガバイト/秒のI/O速度が求められるが、その実現のためのストレージシステムは莫大なディスク数とストレージサーバが必要となる。幸いOak Ridge国立研究所のJaguarでのストレジワークロードの研究[4, 5]で、80~90%程度のI/Oワークロードはスクラッチおよびチェックポイントであるという結果が出、丁度速度・信頼性・価格等の面で競争力を持ち出したSSD を各計算ノードに装備し（図10）、超高速ローカルI/O 装置としてそれらのワークロードを担当させることにより、Lustre のバンド幅への要求を大幅に下げられるという結論に達した。これにより要求仕様に盛り込むと共に、これらを高信頼に用いるチェックポイントアルゴリズムの研究をスタートさせ、かつ有効な運用方法を技術的に検討することとした。実際、TSUBAME 2.0 に初期配備されるSSD の合算バンド幅は660ギガバイト/秒に達し、Jaguarの並列ファイルシステムのそれを大幅に上回る。

(ケ) **一部信頼性の向上の必要性：**TSUBAME はクラスタ計算機としては大規模ゆえにかなり高信頼を意識して設計された・実際TSUBAME の障害ログは全てつぶさにGSIC のTSUBAME のHP で公開されているし、メジャーな全システムのダウンは東京城南地区の珍しい大停電を含め4 年間に二回程しかなかった。しかしながら、single point of failure はストレージおよびバッチキューシステムに存在し、しばしばその不調は（マシン全体では無いものの）バッチキュークラス全体など、かなり大きな資源部分の障害を引き起こした。そこで、ストレージを含む多くの部分のハードレベルの多重化、およびサービスの多重化により、single point of failure の排除した設計に努めた。コストはかかるものの、多重性は常時運用時の性能アップも達成できるので、メリットが大きい。

TSUBAME 2.0 のいよいよなる始まり —後篇に向けて 5

以上、TSUBAME1 から 2.0 への進化の過程の概要を述べた。本稿執筆時点で、TSUBAME1 は 2.0 稼働準備の電力等確保のために段階的に縮退が進んでいる。順調に行けば、10 月末にはその歴史を終え、11 月初旬正式稼働の 2.0 の運用に直接引き継ぐ。多くの基本ソフトウェアスタックはTSUBAME1 から引き継がれるので、SGE ベースからPBS Proベースに変更となったバッチキューコマンドの多少のコマンドの違いや、キュー・ストレージ構成の多少の変更があるものの、ユーザはほとんど機能的な違いを感じないはずである。しかし、その性能をフルに発揮するアプリを走らせた場合、その30 倍の性能向上を体験することになるだろう。それは後編で紹介するポストペタからエクサへの入り口であり、GPUによるメニーコアやマルチスレッドやSSDによる超高速I/O, 更には超高速光ネットワーク、更には新たなハードやシステムの大規模化・高速化に対応する種々のソフトウェア上の新言語・新機能・新たなツール群など、スーパコンピューティングの時代的変革を感じるようになることであろう。後編では、それらの観点からTSUBAME 2.0 を評価し、かつベンチマークの性能でそれらを具体的に論じる予定である。乞うご期待されたい。

参考文献

[1] Satoshi Matsuoka, Toshio Endo, Naoya Maruyama, Hitoshi Sato, Shin'Ichiro Takizawa. "The Total Picture of TSUBAME 2.0", The TSUBAME E-Science Journal, Tokyo Tech.GSIC, Vol. 1, pp. 16-18, Sep. 2010

[2] Satoshi Matsuoka. Petascale Computing Algorithms and Applications --- Chapter 14 The Road to TSUBAME and Beyond, Chapman & Hall CRC Computational Science Series, pp.289-310, 2009.

[3] Toshio Endo and Satoshi Matsuoka. "Massive Supercomputing Coping with Heterogeneity of Modern Accelerators", In Proc. 22nd IEEE International Parallel & Distributed Processing Symposium (IPDPS 2008), The IEEE Press, Miami, FL, April 2008, pp.1-10, DOI: 10.1109/IPDPS.2008.4536251

[4] Weikuan Yu, Jeffrey S. Vetter, H. Sarp Oral. "Performance Characterization and Optimization of Parallel I/O on the Cray XT", In Proc. 22nd IEEE International Parallel & Distributed Processing Symposium (IPDPS 2008), The IEEE Press, Miami, FL, April 2008, pp.1-10, DOI: 10.1109/IPDPS.2008.4536277

[5] Julian Borrill, Leonid Oliker, John Shalf, Hongzhang Shan, Andrew Uselton. "HPC Global File System Performance Analysis Using A Scientic-Application Derived Benchmark", Parallel Computing, Volume 35 , Issue 6 (June 2009), pp. 358-373.

次世代気象モデルのフルGPU計算
―TSUBAME 2.0の3990GPUで145 TFlops―

下川辺 隆史*　青木 尊之**
* 東京工業大学　総合理工学研究科、　** 東京工業大学　学術国際情報センター

次期気象予報のために気象庁で開発されている非静力気象モデルASUCAをGPU上で効率良く走らせるには、
膨大なコード全体をGPUに移植する必要がある。CUDAによる書き換えと、さまざまな新しい計算手法の導入により、
TSUBAME 2.0の3990 GPUを用い145.0 TFlops（単精度計算）という高い実行性能を達成した。
HPCの主要アプリケーションである気象計算のプロダクション・コードに対し、
GPUスパコンの有用性を示すことができた意義は非常に大きい。

はじめに　1

気象予報は人々の日常生活や関連産業に大きな影響を与え、防災の観点からも非常に重要であることは言うまでもない。大気は地球規模のスケールで見れば非常に薄い層であり、鉛直方向の大気圧（勾配）と重力の釣り合いが近似的に良く成り立つ。これまで静力学平衡モデルの気象計算が行われてきたが、水蒸気の上昇気流による上空での雲形成などが重要であると認識されるようになり、大気の上下運動を考慮する3次元非静力学平衡モデルの気象計算が行われるようになってきた。

気象計算では観測データとそれまでの計算結果を4次元変分原理に基づいてデータ同化を行い初期値を作成する。気象はカオス的な現象であり、初期値から決定論的に時間発展させる長さには限界があり、度々初期値を入れ替えての再スタートが必要となる。

近年、ゲリラ豪雨のような突発的で局地的な豪雨が多く見られるようになった。これからのの気象予報では、このような現象を予報するために高解像計算を迅速に行うことが求められる。気象計算は細かい格子で3次元非静力学平衡モデルの大規模計算を行うようになりつつあり、HPC（High Performance Computing）の代表的なアプリケーションとなっている。

GPUによる気象計算　2

気象計算を高速に実行する（高い実行性能を得る）要求は非常に強い。米国大気研究センターを中心に開発されている次世代大気シミュレーションコード WRF（Weather Research and Forecasting）[1]は世界標準になりつつある研究用のオープンソースのコミュニティ・コードであり、執筆時点で世界最高速のスパコン上でも実行され50TFlopsを記録している[2]。

気象計算は力学過程と呼ばれる風速や気圧、湿度などの予報変数に対する流体力学的な運動の計算と、物理過程と呼ばれる水蒸気の凝縮や雲形成、降雨などの局所的な物理変化のモデリング（パラメタリゼーション）に対する計算に分けられる。力学過程では浮動小数点演算よりも圧倒的にメモリアクセスに時間がかかり、どの計算機でもピーク演算性能に対して十分高い実行性能は得られない。一方、物理過程はさまざまな経験的なモデリングやパラメータを多く含み、一部の計算は非常に負荷の高い浮動小数点演算を必要とする。

ここ数年、GPU（Graphics Processing Units）の持つ高い浮動小数点演算処理能力と高いメモリバンド幅が注目され、GPUを高性能アクセラレータとして汎用計算に使用するGPGPU（General-Purpose GPU）の研究が盛んに行われている。2006年にNVIDA社がGPGPU向けの総合開発環境であるCUDA [3] をリリースして以降、GPGPUのプログラミングが容易となり、HPC分野では数値流体力学をはじめとして、分子動力学、重力多体計算や高速フーリエ変換などでGPUを利用した研究が精力的に進められている。

気象計算の分野も例外ではない。高速化は従来型スパコンのCPU性能の向上に頼るだけでなく、いち早くGPU のような新しい高性能演算デバイスを利用する取り組みが開始されている。WRFのグループでは、計算負荷の高いモジュールのみを GPU に移植し高速化を図った[4, 5]。全体の計算は従来通り CPU 上で実行し、雲物理過程の一部のモジュールだけを GPU 化した。しかし、このようなコードの一部 GPU化ではアプリケーション全体はGPU上で高速実行されない上、CPU-GPU間の通信が頻繁に発生し、これがボトルネックとなるために GPU の持つ本来の性能を十分に発揮できない。移植されたモジュール単体は20倍の高速化に成功しているものの、アプリケーション全体では30%の速度向上に留まっている[4]。物理過程は小さなモジュールの集合体であり、頻繁なモデルの変更が行われる。一方、力学過程はコード全体に関係する予報変数の時間積分を行うため、その一部だけGPU化することは殆ど意味がない。

東京工業大学 学術国際情報センターでは、680基のGPUを搭載したスパコン TSUBAME 1.2 の後継機として、2010年11月に4200基を超えるGPUを搭載し日本初のペタスケールのスパコンTSUBAME 2.0 が稼働を開始する。TSUBAME 2.0 は、演算性能の大部分を GPU が担うこととなり、GPUから高い実行性能を引き出すアプリケーションの開発が重要となる。本稿では、気象計算のプロダクション・コードの全体をGPU化し、TSUBAME 2.0 上で実行させるためのプロセスと、その実行性能について紹介する。

3 次世代気象シミュレーションコード ASUCA

ASUCA（Asuca is a System based on a Unified Concept for Atmosphere）は気象庁が次期の気象予報のための現業コードとして開発を進めている次世代高分解能局地モデル[6]である。ASUCAのうち、現在までに開発を終えている全ての部分をGPU上へ実装することを試みた。

力学過程のGPU化はアプリケーション全体の枠組みに大きく影響するため、GPU版のASUCAの開発の方向性を決める上で重要なステップとなる。ASUCAは一般座標系 を採用し、力学過程の方程式系はフラックス形式の完全圧縮・非静力学方程式系である。ASUCAでは鉛直方向の音波関連項のみを陰的に扱う HEVI（Horizontally explicit-Vertically implicit）法を採用している。風速などに比べて伝播速度の速い音波、重力波に関係する項はサブステップを用いて小さい時間刻みで計算し、移流計算や物理過程など、その他の項は大きい時間刻みで計算する[7]。大きい時間刻みでは3段階 Runge-Kutta 法、小さい時間刻みでは2段階 Runge-Kutta 法を用い、これらはWRFと殆ど同じ計算内容である。雲物理過程としては現在のところ水蒸気、雲水、雨を取り扱うKesslerタイプのwarm-rainスキームが導入されている。

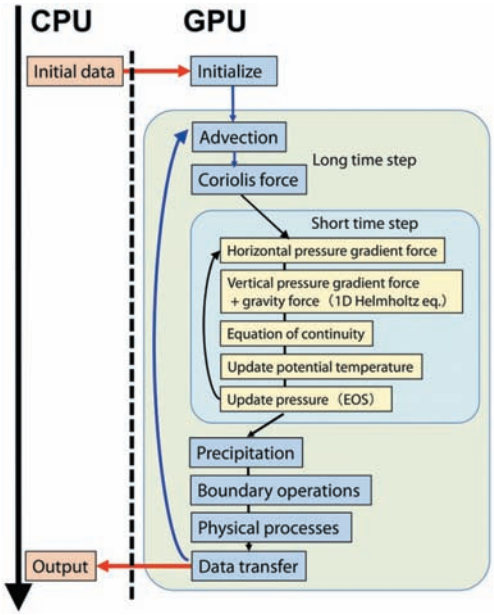


図1 ASUCA の時間発展計算の流れ。
時間ループ内の全計算はGPU上で行われ、
入出力のみCPU上で行う。

4 単一 GPU への実装と計算性能

TSUBAME 2.0に導入されたNVIDIA社製GPUであるTesla M2050（Fermi アーキテクチャ）を用いて実行することを目指し、GPUコンピューティング用の統合開発環境である CUDAを用いてGPUコードの開発を行った。最終的なASUCAの複数 GPUへの実装を説明する前に、単一GPU計算における最適化手法について述べる。

ASUCAの実行の流れは図1のようになっている。初期値データをCPUで読み込み、データをGPU上の global メモリへ転送する。全ての予報変数をGPUのglobalメモリ上に確保し、時間発展ループの内側の全ての計算をGPUで実行する。計算結果である予報値を出力するときのみ、最小限の必要なデータをGPUのglobalメモリからCPUのメインメモリへ転送する。

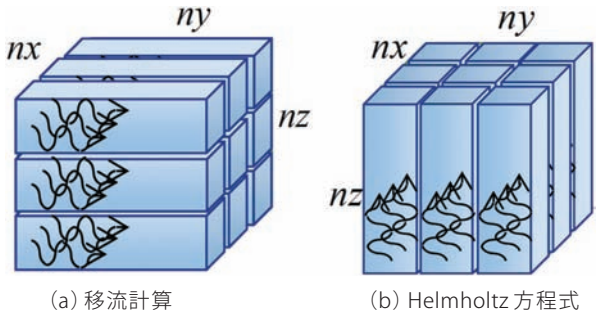
4-1 単一 GPUでの最適化手法

GPUの高い実行性能を引き出すためにASUCAのGPUコード開発で導入される重要な最適化手法について、(a) 移流計算の実装と(b) 楕円型のHelmholtz方程式の計算に対する実装について焦点を当て、以下に説明する。

(a) 移流計算の実装

3次元の移流計算を GPU 上で効率的に行うためには、global メモリへのアクセスをできる限り少なくすることが必要である。CUDA の block 内で共有される shared メモリをキャッシュとして利用することで計算効率を向上させる。

計算領域の格子点数を nx×ny×nzとする。CUDA の blockを(64, 4, 1)、grid を(nx/64, nz/4, 1)と割り当てCUDAのgridを物理空間のxz平面に平行に配置する。CUDA の y方向が物理空間の z方向となる。CUDAのblock内の各 threadを(x, z) 平面に割り当て、y方向にマーチングすることで ny 個の格子点に対し j = 0 から j = ny－1 まで順に計算を行う(図2-(a))。blockサイズは(64, 4, 1)と最適化し実行性能を高めている。



(a) 移流計算　(b) Helmholtz 方程式

図2 CUDAブロック内で実行される
スレッド・マーチング法

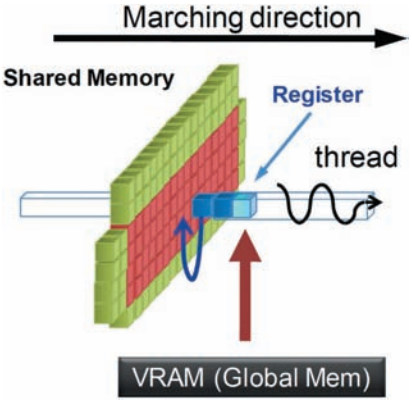


図3 レジスタの利用によるglobalメモリアクセスの抑制

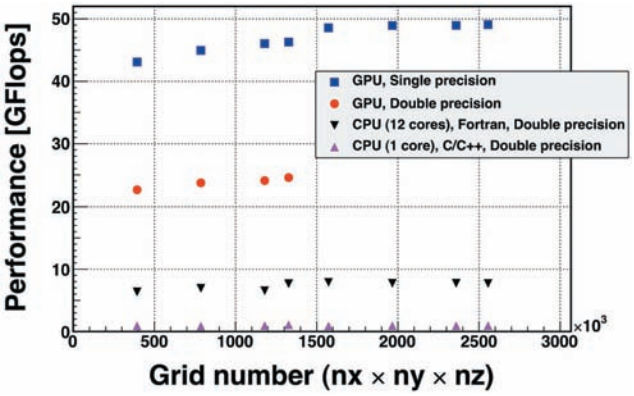


図4 ASUCA の 1GPUでの単精度、倍精度計算における計算性能と CPU（1コアと12コア）での倍精度計算における計算性能。

ASUCA の移流計算では流束制限関数を導入した風上差分法とLax-Wendroff 法を混合したスキームが用いられている。この計算はそれぞれの空間方向 4 点を参照する。各 blockはそのサイズよりも両方向に3大きい(64+3)×(4+3)の配列をsharedメモリ上に確保する。物理空間のxz平面のデータは計算に必要な時にglobalメモリから一旦このsharedメモリへコピーし、block内のthreadで共有して利用する。図3の外周領域には block の境界に位置するthread がアクセスする隣接データを格納する。sharedメモリを利用することでblock内のthread がglobalメモリへ重複してアクセスすることを回避できる。一方、y方向にはマーチングの方向なのでthread 間でデータを共有する必要がなく、該当するthreadの一時変数（register）に格納して利用する。j+1番目の計算でsharedメモリで共有するデータは j番目の計算時に既にregisterへ格納されているため、実際にはj+1番目の計算開始時にsharedメモリへはglobalメモリからでなくregisterからデータを移動し再利用することで計算効率を高めている[8]。

(b) 1次元 Helmholtz 方程式による気圧計算の実装

移流計算と異なり気圧に対する楕円型の1次元 Helmholtz方程式を解く計算では、3重対角行列を解くために鉛直方向に逐次計算を行う。CUDA のgrid を物理空間の xy平面上に取り、各 thread はブロック内の(x, y) 平面に割り当て、z方向に一往復マーチングすることで計算が完了する（図2-(b)）。

4-2 単一 GPUによる計算性能

気象庁において開発されてきた ASUCA はFortran言語で書かれているが、GPU上に実装するにあたり予報変数の配列の次元の順序を変更するため、その検証目的も含めて C/C++言語に書き換え、さらにCUDAでGPU化を実装した。CPUのC/C++コードはGPU との計算性能を比較するためにも用いる。また、GPU コードの浮動小数点演算数を直接測定することは困難であるため、ASUCA では CPU コードに対してPAPI(Performance API)[9]を用い実測した浮動小数点演算数を元に、GPU の実行時間から実行性能を評価している。

単一GPUを用い、格子数のx、z方向をそれぞれ nx＝256、nz＝48と固定し、y方向の格子数を変化させて単精度、倍精度でそれぞれASUCAを実行した。GPUはTSUBAME 2.0 に搭載されたNVIDIA Tesla M2050を使用し、その実行性能を図4に示す。また、GPUコードの元となったC/C++言語で書かれたASUCAをIntel社製 Xeon X5670（Westmere-EP）2.93GHz 6-core のうち1コア用いた時の計算性能と、Fortran で書かれたオリジナルのASUCA（Intel ifort でコンパイル）をXeon X5670 2.93GHz 6-core x 2の12コアで計算した時の計算性能についても合わせて示している。256×208×48 格子の計算では単精度で 49.1 GFlops の性能を示し、これは倍精度で TSUBAME 2.0 の1ノードにあるCPU 12コア用いたオリジナルのFortran コードによる計算の約6倍の性能を達成している。また、GPUの単精度と倍精度の実行性能の違いがメモリ転送量に比例した2倍程度になっている。

マルチGPU計算 5

TSUBAME 2.0 に搭載されるTesla M2050 は 1 GPU当たり 3GByte のメモリしか持たない、ASUCAは 1 GPU に対して 256×208×48の計算格子サイズまでしか計算することができない（単精度の場合）。これ以上の計算格子サイズを計算するためには複数のGPUを用いる必要がある。例えば気象庁の数値予報で用いている典型的な計算格子サイズは721×577×50である。

ASUCAでは鉛直方向の格子点数が最大で100程度であるため、複数GPU計算を行うために全体の計算領域をx、y方向に二次元領域分割し、それぞれの領域の計算を一つのGPUが受け持つ。隣接するGPU間での境界領域のデータ交換が必要になるが、GPUは他のGPUの global メモリ上のデータに直接アクセスすることができない。そこでGPU間のデータ転送はホストCPUのメモリを経由し、次の3段階で

構成される。(1)CUDAランタイムライブラリによるGPUからCPUへの転送、(2) MPIライブラリによるCPU間のデータ転送、(3)CUDAランタイムライブラリによるCPUからGPUへの転送を行う。

複数GPU計算では、アプリケーション全体の実行時間に対してGPU間のデータ交換に必要な通信時間が無視できない。通信時間はCPU計算の場合と同程度であるが、GPUはCPUと比較して格段に計算が速いため、相対的にアプリケーションの実行時間に占める通信時間の割合が大きくなる。このため、複数GPUを使うアプリケーションでは、通信のオーバーヘッドを隠蔽する工夫が必要になる。ASUCAでは、スケーラビリティを向上させるため、GPU 間のデータ交換に必要な通信をGPUの計算とオーバーラップさせる最適化手法を導入している[10]。

5-1 マルチGPUによる計算性能

TSUBAME 2.0 において複数GPUを用いた計算によるASUCAの性能について説明する。TSUBAMEの各ノードはIntel社製 Xeon X5670 (Westmere-EP)2.93GHz 6-core x 2、メインメモリ約50 Gbyteがあり、PCI-Express Bus 2.0 × 16にNVIDIA社製のGPU Tesla M2050が3個接続されている。各ノード間はQDR InfiniBand (4GB/sec) 2本で接続されている。各 GPUは単精度計算では 256×208×48格子、倍精度計算では256×108×48格子を担当する。これは1GPUあたりのメモリを最大限活用し同時実行するスレッド数を最大にし、コアレス・アクセスを最大にするサイズである。

図5に複数GPUで実行したASUCAの計算性能示す。3990 GPUを利用して14368×14284×48計算格子に対して行った計算では、145.0 TFlopsという極めて高い実行性能を達成した。また倍精度計算では3936 GPUを利用して10336×9988×48計算格子を計算し76.1 TFlops の実行性能を達成した。また、GPU数を増やしたときに問題サイズを変えて実行した場合、良い弱スケーリングが得られている。

図6に現在の数値予報で使用されている初期値データと境界値データを用いた台風の気象計算を行った例を示す。437GPU用い計算格子4792×4696×48(実際の格子間隔は水平500m)を単精度で計算した。

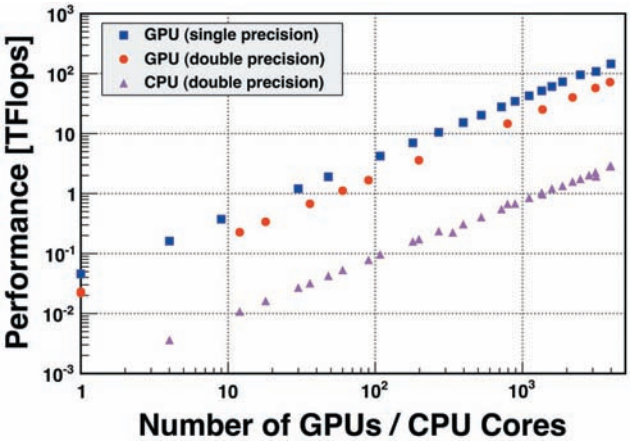


図5 TSUBAME 2.0 に搭載された
複数GPU計算と複数CPU計算の
実行性能 (弱スケーリング) の比較

図6 台風の気象計算をTSUBAME 2.0 の 437 GPU を用い
計算格子4792 × 4696 × 48 (水平500m格子) で単精度計算した例。雲を可視化している。

おわりに 6

次世代の気象予報を目指して開発している気象計算のプロダクション・コードをフル GPU 化し、東京工業大学の TSUBAME 2.0 スパコンの GPU で実行した。CPUにより実行した場合と比較すると、GPUによる計算では圧倒的な高速化を達成することができた。GPU を用いた計算は、マシンのコストと消費電力の面で大きなアドバンテージがあり、気象計算のような実用目的のコードに対する有効性を示すことができたことの意義は大きい。中国を始めとして、大量のGPU がスパコンに導入される時代が始まろうとしている。東京工業大学・学術国際情報センターのTSUBAME 2.0 はその先駆けであり、日本初のペタフロップスマシンである。そのような時代において、本稿で示したような大規模なGPU計算のアプリケーションがさらに開発されて行くことを期待する。

謝辞

ASUCAのオリジナルコードを提供していただきGPU版の開発に協力していただいた気象庁　室井ちあし氏、石田純一氏、河野耕平氏とTSUBAME 2.0 でASUCAを実行するにあたり協力していただいた東京工業大学　松岡聡教授、遠藤敏夫准教授、額田彰氏、丸山直也氏に深く感謝する。

本研究の一部は科学研究費補助金・基盤研究（B）課題番号19360043「多モーメント手法による多目的CFDコアの開発」、科学技術振興機構CREST「次世代テクノロジーのモデル化・最適化による低消費電力ハイパフォーマンス」および日本学術振興会（JSPS）グローバルCOE プログラム「計算世界観の深化と展開」（CompView）から支援を頂いた。記して謝意を表す。

参考文献

[1] W. C. Skamarock, J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, M. G. Duda, X. Y. Huang, W. Wang, and J. G. Powers, “A Description of the Advanced Research WRF Version 3,” National Center for Atmospheric Research, (2008)

[2] A. S. Bland, R. A. Kendall, D. B. Kothe, J. H. Rogers, and G. M. Shipman, “Jaguar: The world’s most powerful computer,” in 2009 CUG Meeting, pp. 1–7. (2009)

[3] “CUDA Programming Guide 3.2,” http://developer.download.nvidia.com/compute/cuda/3_2/toolkit/docs/CUDA_C_Programming_Guide.pdf, NVIDIA, (2010)

[4] J. Michalakes and M. Vachharajani, “GPU acceleration of numerical weather prediction.” in IPDPS. IEEE, pp. 1–7, (2008)

[5] J. C. Linford, J. Michalakes, M. Vachharajani, and A. Sandu, “Multi-core acceleration of chemical kinetics for simulation and prediction,” in SC ’09: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis. New York, NY, USA: ACM, pp. 1–11, (2009)

[6] J. Ishida, C. Muroi, K. Kawano, and Y. Kitamura, “Development of a new nonhydrostatic model “ASUCA” at JMA,” CAS/JSC WGNE Reserch Activities in Atomospheric and Oceanic Modelling, (2010)

[7] W. C. Skamarock and J. B. Klemp, “Efficiency and Accuracy of the Klemp-Wilhelmson Time-Splitting Technique,” Monthly Weather Review, vol. 122, pp. 2623–+, (1994)

[8] P. Micikevicius, “3D finite difference computation on GPUs using CUDA,” in GPGPU-2: Proceedings of 2nd Workshop on General Purpose Processing on Graphics Processing Units. New York, NY, USA: ACM, pp. 79–84, (2009)

[9] S. Browne, J. Dongarra, N. Garner, G. Ho, and P. Mucci, “A portable programming interface for performance evaluation on modern processors,” Int. J. High Perform. Comput. Appl., vol. 14, no. 3, pp. 189–204, (2000)

[10] T. Shimokawabe, T. Aoki, C. Muroi, J. Ishida, K. Kawano, T. Endo, A. Nukada, N. Maruyama, and S. Matsuoka, "An 80-Fold Speedup, 15.0 TFlops Full GPU Acceleration of Non-Hydrostatic Weather Model ASUCA Production Code", in SC ’10: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis. New York, NY, USA: ACM (2010) in press

MEGADOCKによる タンパク質間相互作用予測
～システム生物学への応用～

松崎 由理* 大上 雅史* 内古閑 伸之* 石田 貴士* 秋山 泰*
*東京工業大学 大学院情報理工学研究科 計算工学専攻

TSUBAME を用いて、タンパク質立体構造データに基づく剛体モデルのドッキング計算を高速化し、タンパク質間相互作用予測システム「MEGADOCK」を構築した。
これまでシステム生物学におけるネットワークレベルの大規模問題には十分活用されてこなかったタンパク質立体構造情報を、バイオインフォマティクスの手法によって活用する道を拓いた。
現在は百万ペア規模の予測問題への応用を目指して開発を進めている。

はじめに 1

生命は様々な分子の相互作用によって維持、発展している。我々はバイオインフォマティクスの手法によるシステム生物学の研究として、タンパク質間相互作用（Protein-Protein Interaction, PPI）の予測問題に取り組んでいる（図1）。例えばヒト細胞内では数万種類存在するといわれるタンパク質が相互にどのような制御関係にあるかを理解することは、病因の解明や薬剤の設計における重要な課題となっている。

従来は1対1のタンパク質間相互作用予測において、既に知られている相互作用の詳細な確認を行うのが計算機の役割と考えられていたが、本研究では数十から数百個のタンパク質群における相互作用の可能性を網羅的に高速に予測することを初めて可能とした。我々はタンパク質の形状相補性と静電相互作用のみに基づく単純化された評価モデルを提案し、フーリエ空間上での演算により計算量を大きく減じた上で、さらに大規模並列計算機上での効率的な並列計算が可能となるようにPPI 予測システム「MEGADOCK」のTSUBAME への実装を行った。

我々が提案する計算に基づく相互作用予測の手法は、数千から数万CPU コアが比較的自由に使える近年の計算機システムを前提とすれば、バイオインフォマティクスの基本的なスクリーニング手法になると期待される。

MEGADOCKのドッキング計算 2

MEGADOCK は、タンパク質のペアに対してそれぞれの立体構造情報を利用して剛体モデルによるタンパク質ドッキング計算を行い、その結果として得られる評価値に基づいて相互作用の有無を予測するシステムである。剛体モデルによるドッキングは、タンパク質の構造変化を考慮せず、主として表面形状の相補性に基づいた手法である。

MEGADOCK の中核をなす「ドッキング計算システム部」の処理は、形状の相補性に関する項Gと静電的相互作用の項Eの計算からなる。ここで、対象とするタンパク質ペアについて片方をレセプターR、もう片方をリガンドL と呼ぶことにする。それぞれのタンパク質を1 辺が1.2Å のボクセル空間上に表し、タンパク質の内部か表面かなどの種別によって、各ボクセル上に異なる数値を代入する。形状の相補性の項Gには、我々が提案したreal Pairwise Shape Complementarity（rPSC）スコアを用いる。 rPSC スコアは以下のように表される[1]。

$$G_R(l,m,n)=\begin{cases}\#\text{ of R atoms within (3.6Å + R atom } r_{vdW}) \\ -27 \text{ inside of the R}\end{cases}.$$

$$G_L(l,m,n)=\begin{cases}0 & \text{solvent accessible surface layer of the L}\\ 1 & \text{solvent excluding surface layer of the L}\\ 2 & \text{core of the L}\\ 0 & \text{open space}\end{cases}.$$

rPSCスコアは実数のみの表現で表面形状の相補性を的確に表現したものである(図2)。他の物理化学的相互作用を虚数部分に導入することにより、二つの異なる作用を一つの複素数で計算することができる。

MEGADOCKではリガンドを回転・平行移動させながら全空間におけるスコアの値を畳み込み和として計算する。リガンド回転角の刻み幅は通常15°とし、3,600通りの回転パターンで計算を行う。

静電的相互作用による項Eについてはボクセル$i(l,m,n)$に対する電界Φ_iを定義し計算する。アミノ酸残基ごとにCHARMM19[2] に基づいて原子に電荷を与え、ボクセルごとに電荷$q(l,m,n)$を決定し、静電的相互作用の項$E_R(l,m,n)$, $E_L(l,m,n)$を決める。以上を用いて、ドッキングスコアSを以下のように定義する。

$$R(l,m,n)=G_R(l,m,n)+iE_R(l,m,n).$$
$$L(l,m,n)=G_L(l,m,n)+iwE_L(l,m,n).$$
$$S(\alpha,\beta,\gamma)=\mathcal{R}\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N}R(l,m,n)L(l+\alpha,m+\beta,n+\gamma)\right].$$

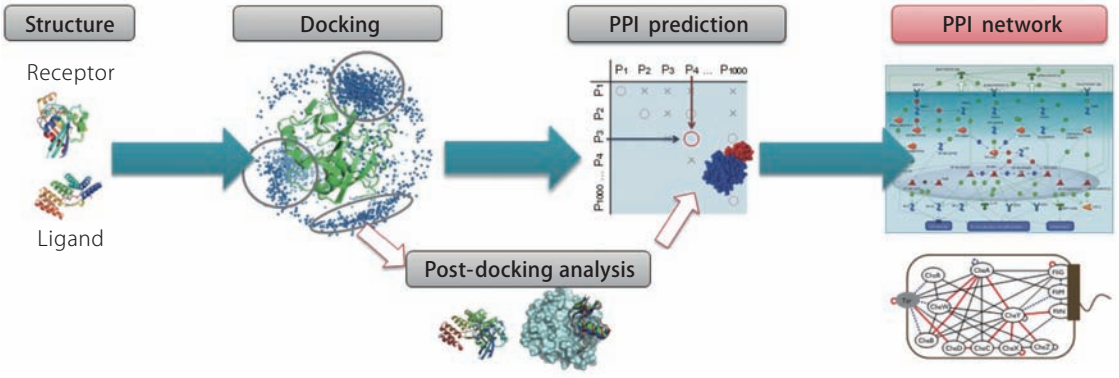


図1 PPI 予測によるネットワーク推定

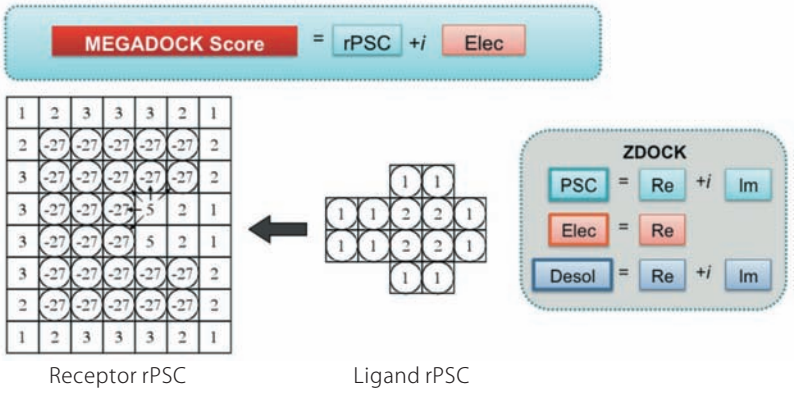


図2 rPSC とドッキングスコア

タンパク質間相互作用予測における形状相補性と物理化学的相互作用を考慮したスコアの計算量は、畳み込み和の直接計算では $O(N^6)$だが、離散フーリエ変換（DFT）と逆離散フーリエ変換（IFT）を用いて表現し、高速フーリエ変換（FFT）を行うことで $O(N^3 \log N)$ に削減することが可能となる[3]。FFT でのスコア Sは以下のように表される。

$$S(\alpha,\beta,\gamma) = \mathrm{IFT}[\mathrm{DFT}[\mathrm{R}(l,m,n)]^* \mathrm{DFT}[L(l,m,n)]].$$

rPSC の導入によるスコア関数の一本化によって、FFT の計算回数を減らし、計算の高速化を図っている。従来のドッキングツール（ZDOCK[4,5]）では図のような三つの要素を三つの複素数で計算しているのに対して、rPSC を用いたMEGADOCK では、ZDOCK の約 4 倍の計算速度向上を実現した。

3 大規模並列化

MEGADOCK はMPI ライブラリにより並列化されている。多数のレセプタータンパク質とリガンドタンパク質の間での網羅的 PPI 予測を行う場合、それぞれのペア毎の計算はほぼ独立であるため、様々なレベルでの並列化が可能である。レセプターとリガンドが入力されたとき、メモリ容量等を考慮してユーザが指定した方式により、複数のプロセッサ間でレセプターとリガンドを分配した並列計算を行うことができる。m 個のレセプターと n 個のリガンドを受け取ったプロセッサは、n 個のリガンドの一つずつを順に取り出し、指定された角度刻みごとにFFT 化を行い、そのたびに最内ループとして m 個のレセプターとの比較を行う。これは回転角ごとのFFT を無駄に繰り返さないためである。

I/O 能力が高いTSUBAME のようなシステムにおいて有効な手法として、網羅的計算においてレセプターまたはリガンドのFFTを予め全システム内で一度だけ計算し、様々な相手のタンパク質に対してディスクからの読み出しで畳み込み計算を直接始める機能をMEGADOCKには実装した(図3)。このFFTライブラリ化を用いることにより、TSUBAME 1.2

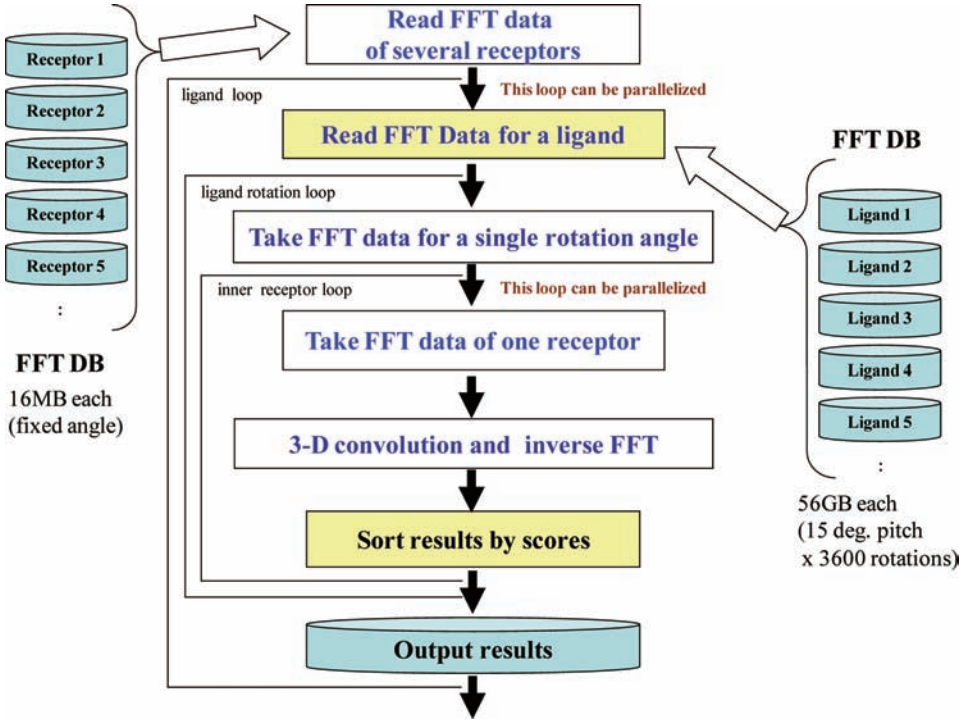


図3 FFTライブラリによる高速化

システムでは最大3倍程度の高速化を達成した。

MEGADOCK では、タンパク質を囲む立方体のサイズを必要最小限に抑えるため、2 のべき乗だけではなく、2、3、5 の三種類の素数の組み合わせを底としてFFT計算を行い、全体性能の向上を実現している。この時、FFT の底をどのように選ぶかにはトレードオフが存在し、様々な底を準備すれば立方体を最小限にできるが、その一方でライブラリ構成が複雑化する。

一方、GPU を用いてFFT計算の加速をする場合は、FFT ライブラリの読み込み機能はむしろ用いずに、FFT の底を自由に最適に選んでその場で毎回計算するほうが全体性能が高くなる。TSUBAME 2.0 での実装においては、このような点も考慮していく必要がある。

システム生物学への応用 4

まず、MEGADOCKの性能評価実験として、当分野で一般的に用いられるベンチマークデータにある44 組の複合体を対象として全対全の網羅的なPPI 予測を行い、精度を検証した。44×44 =1, 936通りの組合せに対してMEGADOCKによるドッキング計算を行い、得られた結果に対しクラスタリングなどのポストドッキング解析を行ってPPIを予測した。まず、ドッキング計算の結果得られた予測複合体構造の精度について確認を行った。その結果を図4(上)に示す。緑が天然構造であり、赤が予測構造を示すが、それらがほぼ一致しており、良い予測が行われていることが分かる。また、そのドッキング計算に基づいて行われたPPI予測の結果について、図4(下) に結果を示す。ここで対角線上の暖色は正しく予測できた複合体ペアを示している。このPPI予測の精度は、感度と特異度の調和平均であるF値によって評価した際、0.415という高い値が得られており、類似の研究と比較しても同等以上のものとなっている[1]。

このベンチマークの結果をふまえて、MEGADOCKの実用的な性能を示す例として、システム生物学における典型的な問題である細菌走化性系のシグナル伝達パスウェイの予測を試みた[3]。生物が外界からの刺激に応答して運動する性質を走性とよび、例えば大腸菌は栄養物質に対する化学走性(走化性) を示す。この系の分子間相互作用のほとんどは既知であるため、これらのPPI を「正解」と定義してMEGADOCK の評価を行った。公開データベースから収集した構造情報(タンパク質13種:構造89個) を用いて、MEGADOCK 2.1による89×89=7, 921通りのドッキングと相互作用予測を行った。結果を図5

に示す。この走化性系におけるPPI予測の精度はF値で0.436に達しており、ベンチマークデータを用いて行った性能評価の際に示された予測精度と同等のものとなっている。また、この過程で、CheYタンパク質とCheDタンパク質との間で図6に示すような、これまで知られていない相互作用の候補が発見されており、それらについても検討を行った[6]。最終的に相互作用の有無を確認するには実験による解析が必要となるが、CheY–CheD–CheC の三者の複合体形成過程を実験で解析し、予測された複合体構造と比較することができれば興味深い結果が得られると考えられる。

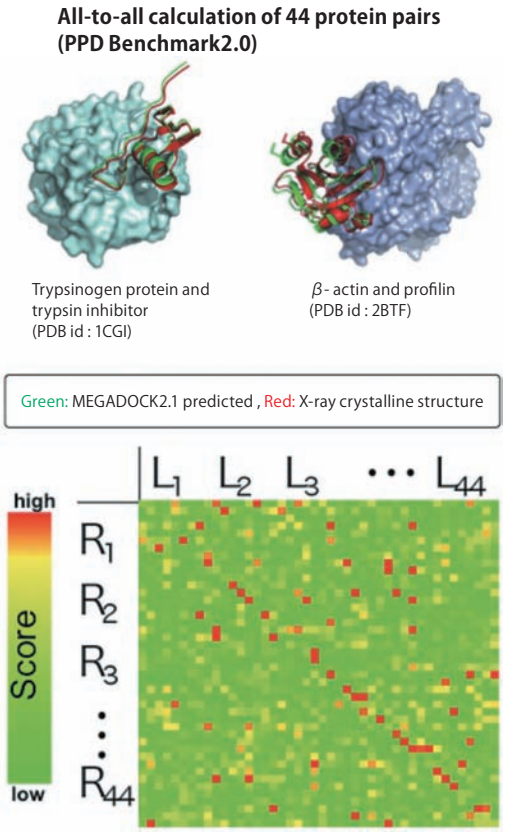
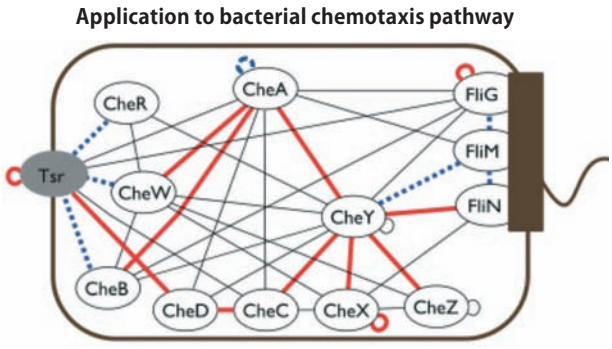


図4 PPI 予測結果： ベンチマークデータ



Red bold line : True Positive
Blue dashed line : False Negative
Black line : False Positive

図5 システム生物学への応用： 細菌走化性系

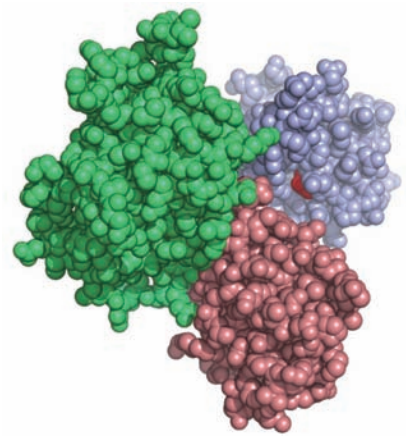


図6 予測された新たな複合体構造（CheY-CheD-CheC）

おわりに 5

MEGADOCK は、TSUBAME を利用した実験により、大規模並列計算機との相性が良く、システム生物学における重要な系の解析を従来よりも高速に実施可能であることが確認された。今後は、1,000 × 1,000（百万ペア）級の大規模計算を目標として、ガン細胞や微生物の細胞などを対象としたシステム生物学の研究に適用する予定である。現在は肺ガンと関連するヒトEGFR シグナル伝達系 を対象とした、500 × 500 規模のPPI 予測に取り組んでいる。

謝辞

本研究は、文部科学省 最先端・高性能汎用スーパーコンピュータの開発利用「次世代生命体統合シミュレーションソフトウェアの研究開発」、および科学研究費補助金（基盤研究（B）19300102）の支援を受けて行われたものである。

参考文献

[1] 大上雅史, 松崎由理, 松崎裕介, 佐藤智之, 秋山 泰, “MEGADOCK：立体構造情報からの網羅的タンパク質間相互作用予測とそのシステム生物学への応用”, 情報処理学会論文誌 数理モデル化と応用（TOM）, 3:91-106, 2010.

[2] Katchalski-Katzir E, Shariv I, Eisenstein M, et al., Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques., Proc Natl Acad Sci U S A, 89:2195-9, 1992.

[3] Chen R., and Weng Z., Docking unbound proteins using shape complementarity, desolvation, and electrostatics., PROTEINS, 47:281-294, 2002.

[4] Chen R., Li L., and Weng Z., ZDOCK: an initial-stage protein-docking algorithm., PROTEINS, 52:80-87, 2003.

[5] Brooks BR, Bruccoleri RE, Olafson BD, et al., CHARMM: A program for macromolecular energy, minimization, and dynamics calculations., J Comput Chem, 4:187-217, 1983.

[6] Matsuzaki Y., Matsuzaki Y., Sato T and Akiyama Y., *In silico* screening of protein-protein interactions with all-to-all rigid docking and clustering: an application to pathway analysis., J Bioinform Comput Biol, 7:991-1012, 2009.

TSUBAME
e-Science Journal

● TSUBAME e-Science Journal No. 2
2010 年 11 月 10 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：海馬 & キックアンドパンチ
編集：　TSUBAME e-Science Journal 編集室
青木尊之　渡邊寿雄　関嶋政和　ピパットポンサー・ティラポン　深山史子
住所：　〒 152-8550 東京都目黒区大岡山 2-12-1-E2-1
電話：　03-5734-2087　FAX：03-5734-3198
E-mail：　tsubame_j@sim.gsic.titech.ac.jp
URL：　http://www.gsic.titech.ac.jp/

GSIC
Global Scientific Information and Computing Center

TSUBAME

TSUBAME 共同利用サービス

『みんなのスパコン』TSUBAMEは、
当初は主に東工大学内の研究・教育のために利用されておりましたが、
平成21年7月よりTSUBAME 共同利用サービスを開始し、
学術・産業・社会へと広く貢献しております。

課題公募する利用区分とカテゴリー

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」のカテゴリーがあります。現在は随時申請を受け付けており、申請課題は厳正な審査の下、採択の可否を決定します。採択課題の利用期間は当該年度末までです。

TSUBAME 共同利用とは…

東工大学内のみならず、より多くの方にTSUBAMEサービスを提供

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・カテゴリー別の利用課金表を下に示しました。TSUBAME 2.0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（12CPU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。この計算機資源は、1000CPUコアを1日半とか、100GPUを3.75日といった利用も可能です。

利用区分	利用者	制度や利用規定等	カテゴリー	利用課金
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：100,000円
産業利用	民間企業を中心としたグループ	「先端研究施設共用促進事業」に基づく	成果公開	トライアルユース（無償利用） 1口：100,000円
			成果非公開	1口：400,000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：100,000円
			成果非公開	1口：400,000円

産業利用トライアルユース制度（先端研究施設共用促進事業）

共同利用サービスの「産業利用」は、東京工業大学学術国際情報センターが実施する文部科学省先端研究施設共用促進補助事業を兼ねております。その中のトライアルユース制度では、初めてTSUBAMEを利用する民間企業の方に限り、無償での利用（1利用期間は最長1年間、2回まで）が可能です。この制度でスパコンTSUBAMEの敷居を下げることで、より多くの方にスパコンの魅力を体験していただいております。

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　Tel. & Fax. 03-5734-2085
●e-mail ／ tsubame@gsic.titech.ac.jp
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