1 アクセラレータ技術の影と光
―ペタ～エクサの次世代 HPC の中心的な躍進技術へ

松岡　聡（東京工業大学）

はじめに―HPC（高性能科学技術計算）のコモディティ化とアクセラレーション

Cell[1]，GPU[2]，ClearSpeed[3]，MDGRAPE[4] などのアクセラレータ技術が昨今注目を浴びている．コンピュータシステムにおけるアクセラレーション自体は特に目新しいことではなく，マルチメディア処理やグラフィクス，I/O やネットワーク，暗号処理などを通常の CPU の外部の特殊ハードウェアや専用プロセッサで行うことは普通であった．しかしながら，昨今のアクセラレータへの注目は，それが汎用のプロセッサとして高度な科学技術計算や，それらに類した一般のアプリケーションにおける高負荷な処理―ビデオ圧縮・リアルタイム画像処理・ゲームにおける物理計算―などへ広く安価に適用できるところにある．それぞれのアクセラレータの内容は参考文献や本特集の他稿に詳しいが，ここでは過去より汎用アクセラレータの隆盛を繰り返してきた高性能計算（HPC）・スパコンに着目し，そこからアクセラレータの歴史・位置づけ・今後に関して述べてみよう．

いわゆるコモディティ技術による高性能 PC プロセッサや PC クラスタはスパコンに代表される HPC（高性能科学技術計算）の世界を革命的かつ根本的に一新した．現在では広く用いられている x86 系のプロセッサは，1980 年代初頭の 8086/8087 の登場時からすでに倍精度の浮動小数点演算をハードウェアでサポートしていた．これは，当時のトランジスタ規模，および前進の 8 ビット系の CPU が整数除算もサポートしていなかったことを考えれば画期的なことであった．初期は Cray に代表される当時のスパコンであるベクトル計算機と比較してその性能は微々たるものであったが，マルチメディアや小規模の科学技術計算のニーズに押されその性能は急速に向上し，1995 年に登場した Pentium Pro では，高度なパイプライン化により毎クロックごとの演算が可能となり，150MFlops を達成した．また，ネットワークやソフトウェアの進化により高性能 PC を汎用高速 LAN で結合した PC クラスタとしての並列計算機の構築も可能になり，その高いコストパフォーマンスによって 1994 年の最初の「Beowulf 型」の PC クラスタ（Wigraf）の登場から十数年，ハード・ソフトウェア技術の進歩によってそのサイズ・台数共に急激に増加し，裾野が広がるとともに，数万 CPU のマシンが構築されるようになった．我が国では大規模 PC クラスタである東工大 TSUBAME[5] が 2006 年 6 月に地球シミュレータを押さえ我が国トップのスパコンとなり，その地位を 2 年間維持した．また，2008 年 6 月の Top500 においてはついに IBM Cell と AMD Opteron の混合型のクラスタ計算機である米国 LANL/IBM のスパコン Roadrunner が，クラスタ計算機としてははじめて世界一となって，同時にはじめて性能値でペタ（10 の 15 乗）フロップスの「壁」を破った．

PC クラスタにより HPC はメインストリーム化・大衆化し従来はスパコンでしか実行可能でなかったアプリケーションが，ある程度の規模までならば PC ワークステーションなどで実行可能となり，従来のスパコンと比較して抜群なコストパフォーマンスの向上をもたらし，大規模なスパコンも同様に構築されるようになって，「デスクトップからスパコンまで」のソフトウェアや利用環境の共通化が果たされた．つまり，スパコンが PC テクノロジの世界的なエコシステムの一部として相互の技術転用を実現することが，HPC の世界を急速にメインストリーム化し，その発展における中心的な原動力となっている．これがユーザ層を広げるだけでなく，従来にはなかったより多くのアプリケーション分野―ビジネス分野からゲーム・エンタテインメントまで幅広く―への広がりをもたらした．たとえば，今や多くの映画製作で不可欠な実写と見まごう 3DCG が可能となったのは，数百台～数千台並列のレンダリング用 PC クラスタの構築が容易になったからである．TSUBAME においても，昨日パソコンで動いていたアプリユーザが，今日 1,000CPU 規模でアプリを動かしている例も枚挙にいとまがない．

しかしながら，2008 年現在においても，IBM, Cray, NEC, 富士通など各社から，専用設計のスパコンが売ら

れているのも事実であり，日米において後継機種の研究開発が続いている．コモディティのコストパフォーマンスからすれば，それら「恐竜」はとっくに絶滅しても良いはずだが，専用スパコンが存在し得るのは，それらが以下のようなPC-HPC技術の欠点のニッチを埋めているからである．

(1) **高性能・高効率なベクトル処理**：x86プロセッサの性能向上は著しく，SSEやAltiVec, AVXなどSIMDベクトル命令も備わっているが，単一スレッドの処理能力では，多くの場合これらの専用設計のプロセッサにはかなわない．これはベクトル並列性が専用プロセッサの8～16に比べてたかだか2～4程度しかないこと，メモリアクセスのバンド幅やランダムアクセス性が弱いこと，などに起因する．過去にはCPUクロックの持続的な上昇で単一スレッドの性能が高まり，専用設計に急速に追いついたが，消費電力や発熱の問題で2005年あたりを境に急速に止まり，今ではそれが困難になった．無論，通常はOpenMPやMPIなどで中～粗粒度の並列性を高めることによって専用プロセッサの性能を大幅に上回ることが多々あるが，それにも限界がある．たとえば計算中の主要アルゴリズムなどで，中～粗粒度レベルの逐次計算の必要性や，あるいはグローバルメモリアクセスのレーテンシに起因するボトルネックが発生すると，アムダールの法則によって性能が上げられず，数倍の性能差が本質的に生じてしまう．

(2) **電力・設置・メンテナンス効率の問題**：PCクラスタによっては効率が出にくいアプリケーションの性能を並列性でカバーしても，それによる電力・スペース・メンテナンス，最終的にはコストの大幅な増加が問題となってくる．多くの大規模スパコンの電力消費量は数百KWから数MWに達し，次世代ではさらに最大で数十MWに達すると目されている．小規模なマシンでも，電力は言うにおよばず，冷却・騒音・重量などの設置性や，サイズと反比例して低下するシステムの故障率が大きな問題となる．つまり，あるアプリケーションの実行に際して，同じ性能が得られるのならば，なるべく計算機としてはコンポーネンツの数や物理的なサイズの面で小さく，かつ効率が良いことが好ましい．それらの点で，専用設計のスパコンが有利になる場合においてはニッチとして存在し得る．

上記のような汎用コモディティプロセッサの欠点を大幅に是正し，専用設計のスパコン並み，あるいはそれを上回る性能を発揮するのが汎用技術をベースとした現代のアクセラレータである．今後のペタスケールから2019年ごろのエクサ（10の18乗）へのトップエンドスパコンの性能向上から，ボトムエンドのPCレベルでの通常のCPUでは効率の悪いビデオ編集などのアプリケーションの加速まで，多くの期待がなされており，研究開発が盛んに行われている．ちなみに，前述のTSUBAME, Roadrunnerとも，Top500においてはじめてアクセラレータを搭載した2台であることは，今後のHPCのトレンドを占う上で重要である．しかし，先にも述べたように，HPCにおいても汎用処理用のアクセラレータは過去より存在し，必ずしも成功したとは言い難い．それらと比較して，現代のアクセラレータはどのように違い，なぜ成功を期待されているのか，次に論じることにしよう．

HPCにおけるアクセラレータの影と光

HPC向けのアクセラレータ技術は古くから存在し，古くはメインフレームやミニコンの「科学技術演算用オプション」としても売られていた．近代スパコンでもThinking Machine社のCM-5や富士通AP1000などに採用されていたWeitek社製のベクトルアクセラレータなど，いくつもの試みがあった．しかしながら，それらは広く受け入れられることなく，場合によってはメーカ自身を含み消滅してしまった．これにより，HPC業界全体で，CPU自身の高速化がメインストリームになり，アクセラレータはかなり穿った目で見られてきたことも事実である．

しかしながら，現在はHPCにおいてアクセラレータが注目技術として脚光を浴びており，同じような理由で通常のPCや携帯などにおけるコンピューティングでも，グラフィクスなど特定用途以外の一般アプリへの適用が着目されている．これはなぜなのであろうか;つまり，過去の暗黒時代とどのように違うのであろうか．これらの技術的な点のみならず，HPCやアクセラレータを取り巻く社会環境までつまびらかにすることにより，今後のアクセラレータの成功を強く確信することが可能となる．

まずは，過去と異なりアクセラレータがPCベースのコモディティのエコシステムの一部として開発・設計がなされていることにある．MDGRAPE, ClearSpeedなども，そのホストとなるプラットフォームはPCであり，PCI-Expressなどの高速I/Oに接続し，主にPC上でプログラムを開発しそのHPCアプリケーションを加速するようになっている．CellやGPUはさらにそもそもの生い立ちがPCや組込み系のコモディティ環境のグラフィクスやマルチメディアの加速が目的であり，当然PCやゲーム機器を中心としたハードウェアや開発環境が安価に存在する．このようにPCエコシステムの一部とな

ることにより，その低価格により容易に一般利用が可能となり，結果としてアクセラレータのHPCアプリケーション開発や利用技術の発展が加速し，普及につながるという，自らのエコシステム維持が今回は可能でありそうなことがアクセラレータが再び脚光を浴びている大きな理由である．

次に，過去には通常のCPUがすぐにアクセラレータに性能的に追いついてしまい，その存在理由を希薄にしていたが，現在では前章の(1)で述べた通り，現代のCPUはクロック上昇による性能向上が2005年ごろに終焉し，その後マルチコア化により性能向上を維持しているがout-of-order実行などのために多くのトランジスタを演算以外の部分に費やしており，また過去との命令体系の互換性の維持のために，そのトランジスタ数の増加に比例した性能向上を得られなくなっている．また，多くのHPCアプリで重要なメモリバンド幅の不足もさまざまな理由で深刻となっているが，レガシーを多く含む一般的なアプリの万遍ない速度向上にはキャッシュの増量以外の決定打がなく，さらに演算回路が利用できるチップ面積を圧迫している．現状で，250mm～300mm程度のチップ面積において線幅45nmで作られるメインストリームのx86系のCPUの場合，SIMD演算をフルに駆使しても，倍精度では50GFlops程度以下，単精度では100GFlops程度以下である．

これらに対し，現代のアクセラレータはアプリケーションの適用範囲を絞り，かつレガシーに縛られないことによって，ベクトル・SIMD・マルチスレッド演算処理よりはるかに高い演算密度を得ている．Cell/B.E.は単精度で256GFlops，倍精度演算が改良されたPowerXCell 8iでは102.8GFlopsのピーク性能である．ClearSpeedも初期よりボードあたり倍精度80～96GFlopsに達している．MDGRAPEは単精度200GFlopsの性能によって，理研の初の分子動力学用ペタフロップスシステム「MDGRAPE-3」の中心的な演算装置となった．GPUのピーク演算性能はさらに高く，2008年末の最新のNVIDIAのTesla T10 GPU，AMDのFireStream GPUとも単精度では1 TFlop以上を可能としており，倍精度の性能も急速に向上しつつある．

また，メモリバンド幅もCell, GPUはCPUと比較して高い．Cell/B.E.は2006年にXDRメモリを用い25.6GB/sを実現し，当時のPCと比較して約4倍のメモリバンド幅を可能とした．PCがこのレベルまで追いついたのは2年以上後の2008年末のIntel Nehalemにおいてであり，しかもトップエンドの高性能モデルに限られる．GPUはさらにメモリバンド幅が高く，Graphics用のGDDRメモリのpoint-to-point接続を利用して，少ないチャネル数で高いバンド幅を実現しており，その性能は100GB/sを超え，NEC SX-9などの最新の専用ベクトル計算機に匹敵する．

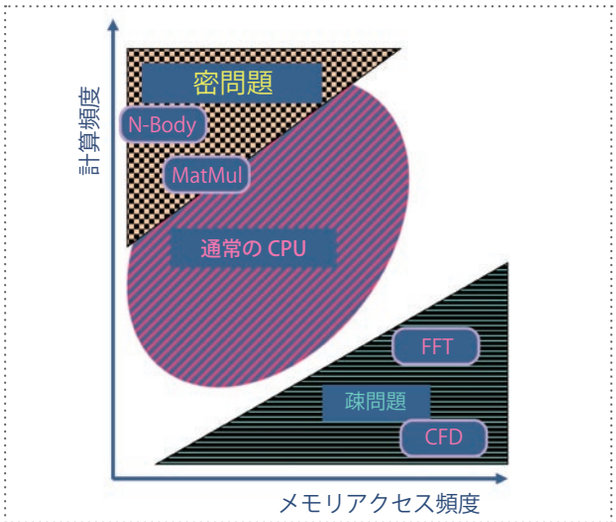

図-1　CPUとアクセラレータにおける計算頻度とメモリアクセス頻度の関係

以上をまとめると，図-1のようになる．ここではX軸は要求される単位時間あたりの計算回数，Y軸は要求される単位時間あたりのメモリアクセスの回数を示す．この中でそれぞれのアプリケーションはこのグラフのどこかにプロットされる．HPCアプリケーションは当然高性能であることが求められるので，グラフ上で右方向・上方向に位置するが，その中でもN-BodyやBLASに代表される密結合系の問題の計算カーネルと，流体力学(CFD)やFFTに代表される，高いメモリバンド幅を要求するものに大別され，前者は左上に，後者は右下に位置する．この中で，特に右下は旧来型のベクトル計算機が得意とするものであった．一方，通常のCPUが得意とする通常の汎用アプリケーションは，左下から右上の(加速が必要ない)エリアに位置する．今回取り上げるアクセラレータは，多くの並列演算ユニットを密にチップに内包し，すべて左上の密結合系の問題の計算カーネルを得意とする．さらに，Cell, GPUは，右下の高メモリバンド幅系のHPCアプリケーションも同時に得意とする．一方MDGRAPE, ClearSpeedはこの領域は不得意だが，その代わり性能に対する消費電力が相対的に低い，などの有利さがある．全体的にHPCで重要な領域がカバーされ，先の(1)，(2)の専用スパコンの存在を意味あるものにたらしめていた領域までも汎用技術でカバーするのが，現代のアクセラレータの過去との大きな違いなのである．

このように，汎用技術に基づき，PCやゲームなどのエコシステムを活用して大幅なコストダウンを果たし，広い普及を可能にしただけでなく，その性能面においてCPUが不得意であった多くのHPCのアプリケーショ

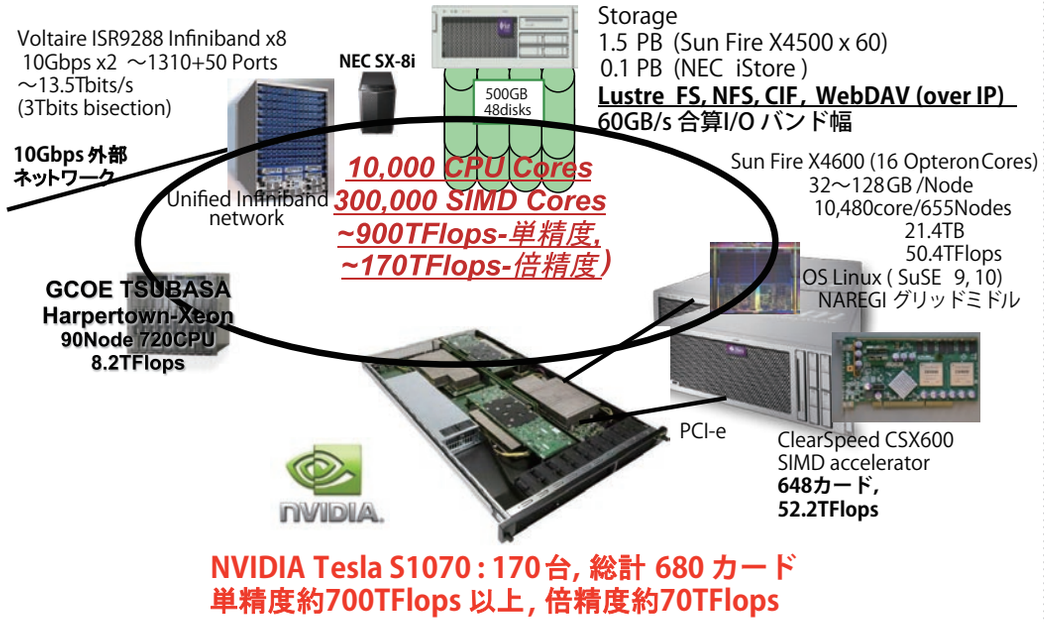

図 -2　TSUBAME1.2 におけるアクセラレータの追加

ン領域をカバーし，HPC の将来に対して大いなる可能性を示している．しかし，真の普及のためには，越えなければいけない大きなハードルがある．それがアクセラレータに適したアルゴリズムと，その開発を支えるソフトウェア技術や環境の進歩である．

アクセラレータ技術の問題点と今後の展望：アルゴリズムとソフトウェア

2008 年 10 月に，東工大において TSUBAME をホストマシンとして，「高速フーリエ変換演算加速装置」と銘打って図 -2, 3 のように 170 台の NVIDIA Tesla S1070 が配備された．従来のスパコンとしての TSUBAME の 10,480 コアの AMD Opteron CPU，従来からの合計 648 枚の ClearSpeed アクセラレータに加え，合算した性能は倍精度演算は約 170TFlops，単精度性能は約 900TFlops 近くに達し，その Linpack 性能は 77.48TFlops に達した．これは，Opteron のみの性能である，倍精度のピーク性能約 50TFlops，Linpack 38.18TFlops を大きく上回る．しかも，それらの設置面積は大変少ないもので（図 -3），かつ消費電力も TSUBAME 全体の 1/8 ～ 1/10 程度である．

しかしながら，アクセラレータで性能を達成するのは容易ではなく，x86 の通常の CPU（Opteron ＋ 一部 Intel Xeon），ClearSpeed，Tesla を複合した異機種環境において Linpack を効率よく動作させるには，負荷分散において我々の過去の異機種分散環境における高性能 Linpack の研究 [6] をさらに発展させた緻密なアルゴリズム設計を必要とした．高メモリバンド幅の 3 次元 FFT カーネルにおいても，Tesla GPU は TSUBAME の 1 ノード・16CPU（Opteron 2.4Ghz）における最適化ライブラリの性能（20GFlops）の 7 倍の 140GFlops の性能を示す [7]．しかしながら，この性能もベクトル計算機用の FFT のアルゴリズムである Multi-Row FFT をベースとしながら，並列化や転置の際のメモリレイアウトや順番の工夫など，細心の注意を払って得られた結果である．このように，アクセラレータはその性質上性能がピーキーであり，最高性能を得るためのアルゴリズムの開発負荷が大きい．その負荷をなるべく減らすには，並列記述を容易にするプログラム言語，性能プロファイリングおよびモデリング技術，並列化をサポートするコンパイラ技術，ランタイム時の自動チューニングを伴う自動最適化技術，並列デバッグ技術，さらにはアクセラレータにチューニングされた種々のライブラリ群など，種々の研究と，実用的なシステムを構築するための開発が必要である．

近年では，これらの課題を解決するために，すでにいくつかの研究開発が行われている．GPU においては，プログラム言語は初期は Brook や PeakStream などストリーム型の言語が中心であったが，NVIDIA 社のマルチスレッド型の言語の CUDA の成功に伴い，アルゴリズムやアプリケーションの開発が大幅に進みつつある．さらに，CUDA をベースに，AMD Firestream や Intel

図-3 TSUBAME への NVIDIA Tesla S1070 の追加．それぞれの S1070 には Tesla 10p プロセッサカードが 4 枚内蔵されている．TSUBAME の計算ノード間に S1070 を挿入し，PCI-Express の拡張ケーブルで計算ノードに接続する．挿入後もほとんど目立たず，アクセラレータの高い計算密度を示している（写真は筆者）．

Larrabee など，他のアクセラレータでもポータビリティが確保されるために，OpenGL の標準化を行っている Khoros グループが中心となって標準言語 OpenCL の規格制定と処理系開発が進んでいる．

我々も，東工大・学術国際情報センターを中心として，他の大学や企業と HPC における研究開発プロジェクトを進めている．JST の戦略的創造研究推進事業（CREST）における「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」では，今後 10 年間で電力性能比を 1000 倍向上させることを目標に，複数の大学の研究グループが合同で研究を行っているが，アクセラレータによる電力性能比の向上が目標達成に大いに寄与する．ここでは割愛するが，その中でさまざまなソフトウェア研究が行われており，また流体アプリケーションにおける加速と省電力化は，本特集の青木氏らの稿に詳しい．また，Microsoft Research との共同研究では，東工大の秋山泰教授と合同で，全対全のたんぱく質間のドッキングを GPU で大幅に加速する研究を行っているが，長時間の大規模計算を実現するために，GPU 計算の高信頼化アルゴリズムの開発を進めている．

しかしながら，汎用アクセラレータの HPC，さらには一般アプリケーションにおけるメインストリーム化のための研究開発は緒についたばかりである．本特集が参考になって，HPC のメインストリーム化，それによりさらに一般のアプリケーションに今までなかった質的な違いがもたらされれば大変に幸いである．

参考文献

1) Cell Broadband Engine Technology and Systems, IBM Systems Journal, 51-5 (May 2007).
2) Owens, J. D., Houston, M., Luebke, D., Green, S., Stone, J. E. and Phillips, J. C. : GPU Computing, Proc. IEEE, 96-5, pp.879-899 (May 2008).
3) ClearSpeed Technology Inc. : ClearSpeed Whitepaper : CSX Processor Architecture, http://www.clearspeed.com/docs/resources/ClearSpeed_Architecture_Whitepaper_Feb07v2.pdf (Feb. 2007).
4) Taiji, M. : MDGRAPE-3 chip : A 165 Gflops Application Specific LSI for Molecular Dynamics Simulations, Proc. Hot Chips 16, IEEE Computer Society Press (CD-ROM) (2004).
5) Matsuoka, S. : Petascale Computing Algorithms and Applications --- Chapter 14 The Road to TSUBAME and Beyond, Chapman & Hall CRC Computational Science Series, pp.289-310 (2008).
6) Endo, T. and Matsuoka, S. : Massive Supercomputing Coping with Heterogeneity of Modern Accelerators, IEEE International Parallel & Distributed Processing Symposium (IPDPS 2008), the IEEE Press (Apr. 2008).
7) Nukada, A., Ogata, Y., Endo, T. and Matsuoka, S. : Bandwidth Intensive 3-D FFT kernel for GPUs using CUDA, Proc. ACM/IEEE Supercomputing 2008 (SC2008), Austin, Texas, the IEEE Press (Nov. 2008).

（平成 21 年 1 月 16 日受付）

松岡　聡（正会員）
matsu@is.titech.ac.jp

1986 年東京大学理学部情報科学科卒業．東大・東工大教員を経て 2001 年東京工業大学学術国際情報センター教授．博士（理学，東京大学大学院）．高性能システム，並列処理，グリッド，クラスタ計算機，HPC の省電力化や加速，大規模データ処理，等の研究に従事．NAREGI プロジェクト・情報爆発特定科研のサブリーダー，2006 年我が国最高性能のスパコン TSUBAME を構築．1999 年本会坂井記念賞，2006 年学術振興会賞受賞．2009 年 ACM/IEEE Supercomputing '09 の論文委員長．