Vol.2011-HPC-130 No.21
2011/7/27

演算加速装置に基づく超並列クラスタ HA-PACS による大規模計算科学

朴 泰祐[†1] 佐藤三久[†1] 塙 敏博[†1]
児玉祐悦[†1] 高橋大介[†1] 建部修見[†1]
多田野寛人[†1] 藏増嘉伸[†1]
吉川耕司[†1] 庄司光男[†1]

筑波大学計算科学研究センターでは 2012 年 1 月の運用開始を目指し，大規模 GPU クラスタ HA-PACS の導入を進めている．HA-PACS は標準的技術による大規模 GPU クラスタ部分に加え，ノード間接続及び GPU 間接続に独自開発の専用相互結合機構 TCA を実装し，次世代のアクセラレータ間結合の要素技術開発を行う．HA-PACS は素粒子・宇宙物理・生命科学等の，極めて大量の演算を要求する大規模並列アプリケーションをターゲットとする．数百台〜千台規模の GPU を定常的に利用するような大規模並列実行によりこれらのアプリケーションを実行し，サイエンスの新しい分野を切り拓くことを目指す．本稿では HA-PACS 計画の概要と，TCA を実現する特別相互結合網 PEARL，そして HA-PACS によって推進が期待される計算科学の例について紹介する．

1. はじめに

近年，GPU (Graphics Processing Unit) の高い演算性能とメモリバンド幅に着目し，これを様々な HPC アプリケーションに適用する GPGPU (General Purpose GPU) 計算が盛んに行われている．また，GPU を搭載する高性能計算サーバをノードとする GPU クラスタの構築も増加の一途を辿り，2010 年 11 月の TOP500 リスト[1] では GPU クラスタによって中国 Chinese Academy of Sciences の Tienha-1A が 1 位にランクされ，日本でも東京工業大学の TSUBAME2.0 が国内第 1 位（世界第 4 位）にランクされている．これらに関しては GPU の持つ極めて高い潜在的演算能力だけでなく，TOP500 リストを決める Linpack ベンチマークを GPU によって加速する方法が確立しつつあることも大きな要因である．しかしながら，大規模 GPU クラスタで数百 TFLOPS の性能を定常的に利用するような大規模アプリケーションについてはまだ例が十分でない．TSUBAME2.0 では気象コード ASUCA の GPU 化[2] やその他様々な計算科学チャレンジが行われており，米国 Oak Ridge National Lab. でもライフサイエンス，気象等の大規模アプリケーションが準備されているが[3]，それらの種類と定常的な実行という点ではまだ不十分である．

筑波大学計算科学研究センターでは，CP-PACS[4]，FIRST[5]，PACS-CS[6] 等の準汎用または特定分野専用の大規模並列システムを開発・運用してきた．これらのシステムは対象とする明確なアプリケーションイメージを持ち，ミッションクリティカル的な導入と運用が行われてきた．我々は PACS-CS に続く次世代の大規模並列システムのターゲットアプリケーションとして，素粒子物理学，宇宙物理学，ライフサイエンスを中心に，限られた電力とスペースで大きな計算科学的成果を挙げるべく，システムの検討を進めてきた．その結果，ピーク演算性能 1PFLOPS クラスの，アクセラレータに基づく大規模クラスタの構築を目指すこととなった．本システムは HA-PACS (Highly Accelerated Parallel Advanced system for Computational Sciences) と名付けられる．[⋆1]

2. HA-PACS の概要

2.1 ベースクラスタ部

現時点で最も有望なアクセラレータ技術は，性能・価格・電力・技術の安定等のあらゆる面で，GPGPU であることは疑いない．問題は，計算ノードの構成，特に汎用 CPU 部と GPU 部をどのようにバランスさせ，大規模並列処理を効率的に実現するかである．このためには，計算ノード間のネットワーク構成ももちろん重要である．現時点で最高性能を持つ GPU は NVIDIA 社の Tesla 2090（コア数を拡張した Fermi アーキテクチャ）[7] で，1 台当たり 665GFLOPS の理論ピーク性能を持つ．我々は各計算ノードに複数の GPU カードを搭載し，最新の CPU と組み合わせることにより，GPU だけに依存しない様々なハイブリッド計算を実現する．

GPU クラスタの最大の問題の一つは，GPU が持つ潜在的演算性能の高さと，GPU と

†1 筑波大学 計算科学研究センター
Center for Computational Sciences, University of Tsukuba

⋆1 本稿公開時点で，HA-PACS のベースクラスタ部は調達作業中であり，最終的な仕様及び性能は確定していない．本稿は我々が求める理想的な形としての一般論として記述している．

Vol.2011-HPC-130 No.21
2011/7/27

CPU をつなぐインタフェースである PCIe (PCI Express バス) の性能との大きなギャップである．500GFLOPS を越える理論ピーク性能に対し，現行の PCIe は Gen2 (Generation 2) x 16 レーンが最大であり，理論ピークバンド幅は 80Gbps (8b/10b 変換により 8GByte/sec) と極めて貧弱である．また，現在の CPU は PCIe のレーン数に限界があり，Gen2 PCIe の総レーン数は 32 程度に留まっている[⋆1]．このため，ノード当たりの GPU 数を増強する場合，本来 Gen2 x16 レーンを要求する先進的 GPU に対し，同性能の PCIe を複数 GPU で共有する等，各 GPU に対して十分な I/O バンド幅を与えることができないのが実状である．あるいは，CPU 及びメモリから PCIe へのパスを提供するチップセット上でボトルネックが生じている場合もある．

これに加え，GPU を備えた多数のノードを結合する相互結合網における通信では，複数ノード上の GPU 間での通信は数回のデータコピー（例：GPU メモリから CPU メモリ⇒ CPU メモリからネットワークデバイス経由で別のノードの CPU ⇒ CPU メモリから GPU メモリ，の 3 ステップ）が必要であり，特に比較的サイズの小さいデータの転送ではレイテンシが性能ボトルネックとなる．一部には NVIDIA 社と Mellanox 社の協力により，通信用にピンダウンされたメモリへの GPU からの直接データ転送技術 (GPUDirect[8)]) 等も開発されているが，データのコピー回数が減少するだけで，基本的に GPU・CPU 間のデータコピーとネットワーク転送が別フェーズとして実現されることに変わりはない．

GPU クラスタにおける並列化の例として，素粒子物理学における QCD 計算の GPU 化の検討では，大規模並列化における GPU 間通信レイテンシが大きなボトルネックになることが報告されている[9)]．**図 1** は，Lattice size が 16^4 の場合と 32^4 の場合について，1 ステップ当たりの計算時間（通信時間隠蔽済み）と全体処理における通信時間の内訳を示している．右側の縦軸は 1 ノード 1GPU の場合に対する並列処理効率を示している．これは相互結合網が GbEthernet であるため極端な例ではあるが，GPU を 1 台使った場合の性能向上が，4 ノードで既に維持できなくなっており，通信性能の重要性を示している．

大規模 GPU クラスタにおける実効計算性能を向上させるには以下の要素が不可欠である．

(1) 複数 GPU をノードに搭載する場合，CPU から GPU への PCIe のバンド幅を出来る限り確保し，GPU が CPU 及び他の GPU と通信する際のボトルネックを解消する．

(2) ノード間相互結合網に出来る限り高いバンド幅と小さいレイテンシを実現し，通信時間隠蔽プログラミングを行った場合の性能を最大限に引き出す．

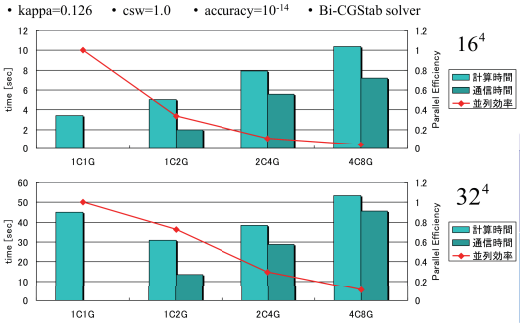

図 1 QCD 計算の GPU 並列化における性能限界の例．ノード当たり GPU 数 (G) とノード数 (C) の組み合わせを表す．CPU: Core i7@2.67GHz, GPU: GeForce GTX285．(広島大学：尾崎祐介・石川健一両氏提供)

(1) に関しては，次世代 PCIe 技術である Gen3 (Generation 3) の利用が有望である．現時点で，PCIe Gen3 を CPU からの直接接続として提供可能な CPU としては Intel SandyBridge-EP,EX シリーズが予定されている．同 CPU ファミリは最大で 40 レーンの PCIe Gen3 を提供可能であり，さらに従来のような PCIe チップセットを用いず，CPU から直接 PCIe をドライブできる．従って，ノード当たり 2 ソケットの SandyBridge-EP,EX を搭載すれば最大 80 レーンの PCIe Gen3 が利用できる．これを最大限に利用すれば，ノード当たり 4 台の GPU を，GPU 側の PCIe バンド幅に合わせて損失なく結合できる．[⋆2]

(2) に関しては，大規模クラスタを実現するための相互結合網として，Infiniband を想定する．ノード当たり 4 台の GPU を考えると，最大 1PFLOPS 程度の構成は 300〜400 ノードに相当する．これは現在の Infiniband のスイッチ規模として全く問題ないサイズである．2012 年 1 月時点のテクノロジとしては現行の QDR に加え FDR 規格も利用可能の見込みであるが，Infiniband HCA を接続する PCIe のバンド幅の有効利用を考えると FDR よりも QDR を数レール実装するのが効果的である．全体の性能バランスを考慮し，HA-PACS では Infiniband QDR x 2 レール相当の，8GByte/sec 程度を考える．

以上が HA-PACS の基本部分で，これを「HA-PACS ベースクラスタ」と呼ぶ．

⋆1 正確には CPU から PCIe へのブリッジとなる周辺チップの問題であるが，CPU からそれらのチップへの接続バスである HyperTransport (AMD) あるいは QPI (Intel) のバンド幅制約によりレーン数上限が制約されている．

⋆2 GPU そのものが PCIe Gen3 に対応するのは 2012 年以降の予定で，現時点では GPU 側は PCIe Gen2 で結合せざるを得ない．

2.2 TCA: アクセラレータ間結合機構

我々は HA-PACS を CPU, GPU, SAN 等の既製の HPC 技術の集合としての GPU クラスタとして実現するだけでなく，次世代のアクセラレータ技術の要素技術として，ノード内及びノード間のアクセラレータ (GPU) 間を直接結合し，演算加速装置を用いた並列プラットフォームにおけるバンド幅/レイテンシ問題の解決を目指す．このため，HA-PACS ベースクラスタ部に加え，アクセラレータ間の直接結合を実現する結合機構を新規に研究開発する．この機構を用いることにより，ノード内 GPU だけでなくノード間に跨がる GPU 間の直接結合を実現する．この機構を「密結合並列演算加速機構：TCA (Tightly Coupled Accelerators)」と呼ぶ．

HA-PACS はベースクラスタに加え，TCA を用いた小規模クラスタの集合からなる HA-PACS/TCA 部を持つ．TCA については次節に詳細を述べるが，基本的なハードウェア技術としては PCIe を用いる．現在研究開発中の TCA はそのネットワークのみで数百ノードのクラスタを構築することはできず，結合ノード数に限界がある．このため，HA-PACS/TCA では 8 台〜16 台程度のノードを TCA によって結合し，特に高バンド幅・低レイテンシな通信を要求するアルゴリズムに対してこれを適用する．全ノードはこの他にベースクラスタと同じ Infiniband 網でも結合されるため，大規模 GPU アプリケーションを階層的ネットワークで構築すれば，高速な局所通信と大規模な一般通信を効率的に組み合わせることが可能となると考えられる．HA-PACS/TCA の概念的構成図を**図 2** に示す．詳細については次節を参照されたい．

3. HA-PACS/TCA と PEARL

3.1 PCI Express による通信リンク PEARL とコミュニケータ PEACH

我々はこれまでに，PCIe リンクを直接ノード間通信に用いる PEARL (PCI Express Adaptive and Reliable Link) を提案してきた[10]．

PCIe は，PC にデバイスを接続するためのシリアルインタフェース規格[11] であり，今日では Ethernet, InfiniBand などのネットワークインタフェース，Serial ATA などのストレージコントローラや，GPU など，ほぼ全ての I/O デバイスが PCIe インタフェース経由で接続されている．各レーンの転送レートは，現在主流の Gen 2 規格における 2.5GHz, 5GHz に加えて，最新の Gen 3 規格では 8GHz までサポートされる (それぞれ実効レートは，2Gbps, 4Gbps, 約 8Gbps となる)．さらに，複数のレーンを束ねてバンド幅を拡張することができ (レーン数を “x4” のように表記する)，x1, x2, x4, x8, x16, x32 が可能である．

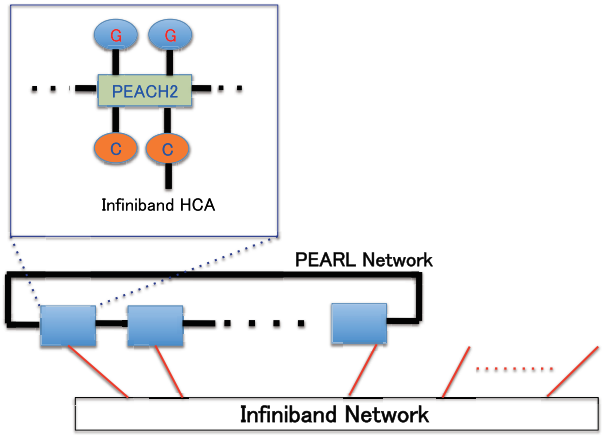

図 2 HA-PACS/TCA の概念図

PCIe における操作は，主に CPU とデバイスとの間でのメモリリード・ライトであるが，実際は，CPU 側に当たる Root Complex (RC) と，デバイス側に当たる複数の EndPoint (EP) と，それぞれの間で双方向のパケット通信を行っているに過ぎない．そこで，我々はこの PCIe リンク上の高速パケット通信をノード間接続に拡張した PEARL を提案し，さらに PEARL を実現するための一種のルータとして，PEACH (PCI Express Adaptive Communication Hub) チップを開発した[12]．

隣接ノード同士を PCIe で接続するためには，一方は RC，もう一方は EP がペアの関係になる必要がある．しかし，ノード CPU は必ず RC であるため，PC の PCIe インタフェース同士を直接接続することはできない．そこで，各ノードには PEACH チップを搭載した PCIe 準拠の PEACH ボードを装着し，その間を PCIe 外部接続ケーブル[13] を用いて接続する．このとき，ケーブルの両端のポートが，RC と EP の対になるように，PEACH ボードならびに PEACH チップではポートの属性をスイッチで切り替えることにより，柔軟に接続を行うことができる．

PEACH チップは，**図 3** に示すように，PCIe Gen2 x4 レーンを計 4 ポート，ルネサスエレクトロニクス社 M32R プロセッサ SMP 4 コア，DMA コントローラ，PCIe パケットバッファ用の SRAM 512KB を内蔵する．

4 個の PCIe ポートのうち 1 つはノード CPU への接続に使われ，残りの 3 ポートで隣

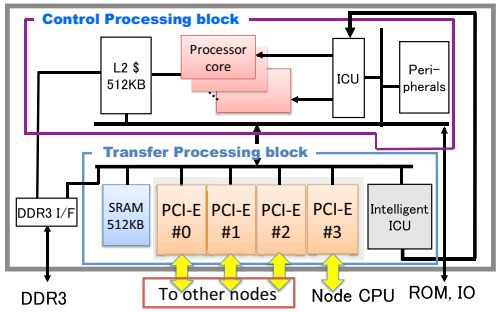

図 3 PEACH チップの構成[12)]

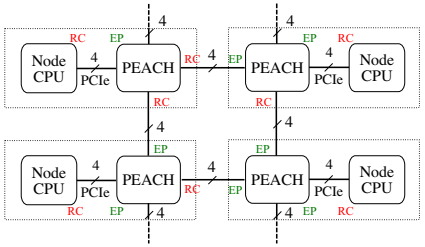

図 4 PEARL によるクラスタ構成例

接するノードの PEACH に接続される．各ポートは PCIe Gen2 x4 であることから最大 20Gbps の転送速度を持ち，理論ピークバンド幅は 2GB/s となる．

内蔵 M32R プロセッサにより，PCIe パケットの中継処理や DMA 起動設定，トランザクション層の制御や，ノード全体の監視やリンク状態の監視などを行う．

図 4 に PEARL を用いたクラスタの構成例を示す．先に示したように，RC と EP が必ず対になるように接続することで，それぞれの PCIe リンクが構成される．

3.2 PEARL によるアクセラレータ直接結合

2 節でも述べたように，通常の GPU クラスタにおいて，複数ノード上にある GPU 間で通信を行うには，少なくとも 3 ステップのデータコピーを伴う．例えば，ノード A 上の GPU A からノード B 上の GPU B に通信を行うには，以下のデータコピーが必要となる．

(1) GPU A のメモリから PCI Express 経由でノード A のメモリにコピー
(2) ノード A のメモリからネットワーク経由でノード B の CPU メモリにコピー
(3) ノード B のメモリから PCI Express 経由で GPU B のメモリにコピー

ここで，ネットワークを PEARL に置き換えることで，PCI Express のプロトコルのまま，ホストメモリにデータをコピーすることなくノード A 上の GPU A からノード B 上の GPU B へ通信することが可能になる．前節で述べたように，PEARL の各リンクは PCIe そのものであるため，物理的にはリンクの先に GPU デバイスを直接接続することができる．**図 5** に，PEARL リンクに GPU を接続した例を示す．これにより，PEARL を用いて GPU の直接結合が可能となり，複数 GPU での通信レイテンシを大幅に削減することが可能になる．一方，**図 6** のような構成も考えられる．この例では，CPU と GPU 間の接続は従来通りで，別の PCIe スロットに PEACH ボードを接続することになる．この場合にも，GPU と PEACH の間は PCIe スイッチを介して直接 PCIe プロトコルで通信可能である．

このように，PEARL をベースに，アクセラレータ間の PCIe 通信プロトコルを直接制御することにより，ノード内のアクセラレータ間はもちろん，PEARL ネットワークを介して接続される複数ノード上のアクセラレータ同士でも直接通信を実現することが理論的に可能である．このようなアクセラレータ間直接通信は GPU クラスタにおける並列処理において性能上極めて重要である．これは従来のような中間メディア（Ethernet や Infiniband 等）を介したノード間通信では基本的に不可能であるが，PEARL は PCIe を直接通信メディアとして利用するため，これが可能になっている．このように，PEARL ネットワークを用い，ノード間に跨がるアクセラレータ間の直接通信を実現し，並列 GPU アプリケーションの性能を大幅に向上させる機構が TCA である．

3.3 PEACH2 に向けて

これまでに開発してきた PEACH チップでは，PCIe 1 ポート当たり Gen2 x4 であり，現在の標準的な GPU の持つインタフェース Gen2 x16 レーンと比較すると高々1/4 のバンド幅しか持たないため，GPU 直接結合の利点を活かすことができない．また，PEACH では組込みプロセッサによる柔軟な処理を優先して設計したため，DMA 転送の起動設定や PCIe パケット中継など機能の一部は，PEACH に内蔵された M32R プロセッサ上の割込みハンドラの動作に頼っている．しかし，内蔵プロセッサの性能は限られており，割込みハンドラが動作する際のオーバヘッドや，プロセッサから PCIe 制御領域へのアクセスレイテンシが大きいことなどから，PEACH 転送性能の低下を招く．これらのことから，我々は現在，HA-PACS/TCA に向けて新たに PEACH2 を開発している．

PEACH2 の開発に当たって，近い将来 GPU の PCIe インタフェースが PCIe Gen3 x8 になるという予測の元に，PCIe Gen 3 x8 を 4 ポート搭載するチップを検討した．近年，

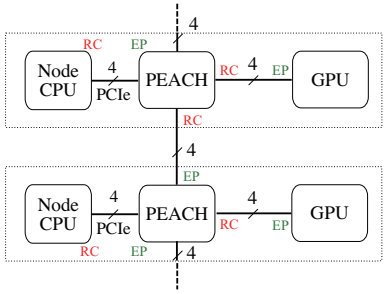

図 5 PEARL に GPU を接続した例 (1)

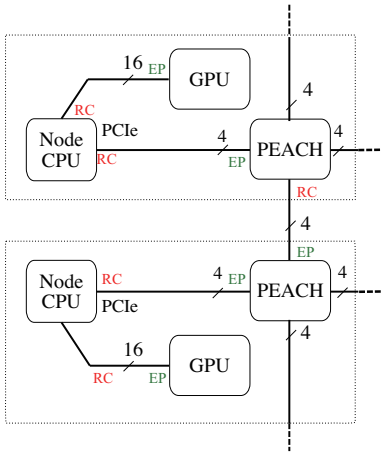

図 6 PEARL に GPU を接続した例 (2)

FPGA デバイスに PCIe のハード IP を内蔵した製品が増えてきており，Altera 社の Stratix V, Xilinx 社の Virtex 7 においては，PCIe Gen3 x8 レーンを 4 ポート搭載可能な製品もアナウンスされている．しかし，今回の HA-PACS/TCA の開発計画の時間的制約により，本研究では，既存の FPGA の中から，Altera 社の PCIe Gen 2 x8 レーン 4 ポート搭載可能な FPGA である Stratix IV GX[14] を用いて PEACH2 を実装することとした．

PEACH2 には，PCIe Gen2 x8 レーンを 4 ポート搭載し，1 ポートはホスト CPU と接続する．さらに，図 5 に示す構成 1 の場合は，1 ポートは GPU との接続に使用し，残りの 2 ポートで隣接ノードと接続する．図 6 に示す構成 2 の場合には，3 ポートを用いて隣接ノードと接続することができる．

PEACH2 では FPGA を用いるため，FPGA 内部の回路を柔軟に変更できるという利点がある．従って，PCIe ポートにおける RC と EP の切替については，それぞれ個別に FPGA のコンフィグレーションデータを用意することで対応する．DMA コントローラには，Scatter/Gather 機能，Chained DMA 機能を備えたものを搭載し，高速な DMA を可能にする．パケットバッファとしては，FPGA 内蔵のエンベデッドメモリ，および外付けの DDR3 SDRAM を用いる予定である．

また，PEACH2 には経路表書き換えなどの操作のために管理用プロセッサを搭載する．PEACH では処理によっては割込みハンドラを動作させるため高性能が必要であったが，PEACH2 ではパケット転送に関わる処理は全てハードウェアによって処理を行う．従って，PEACH2 ではプロセッサには性能が必要でないため FPGA 内蔵向けの小規模なプロセッサで十分対応できる．

PEARL を用いたアクセラレータ直接結合は，特に構成 (1) の場合にはリングトポロジに制約されるため，ノード数 n に対してレイテンシが $n/2$ となり，大規模な結合網には向かない．従って，前節で述べたように，HA-PACS/TCA では，8 から 16 ノード程度の小規模なノード集合 (グループ) を構成し，この上でアクセラレータ間直接通信を実現する．

4. HA-PACS によるサイエンス

HA-PACS ではベースクラスタ部及び HA-PACS/TCA 部において，数百 TFLOPS〜1PFLOPS 規模の演算性能を要求する超大型並列 GPU 演算を実現する．ここでは，現在計画されている 3 つのサイエンスについて概要を紹介する．

4.1 強い相互作用が織り成す物質形態の QCD による統一的解明

物理点での格子 QCD シミュレーションが可能となった現在，次の展開として目指すべきものは，強い相互作用が織り成す物質の様々な物質形態を QCD に基づいて統一的に理解・解明することである．HA-PACS を用いて行う研究課題として，以下の 2 つのテーマを設定する．

(1) 強い相互作用におけるマルチスケールの物理：
階層性を持つ物理系 (クォーク・グルーオン⇒陽子・中性子⇒原子核) をクォーク・グルーオンの力学を記述する QCD のみによって究明する．これは第一原理計算に基づく原子核構造研究という新たな研究分野の開拓である．現在は原子核構造論において最も基本となる 4He 原子核までしか構成できていないが，今後段階的に質量数を増やし魔法数の導出を目指す．また，中性子過剰核を構成してその諸性質解明に向けての先鞭をつける．

(2) 有限温度・有限密度下における QCD の相構造解析と状態方程式の決定：
高密度 QCD は宇宙に存在する星の内部で実現していると考えられているが，その相構造解析は殆ど進んでいない．問題は，格子 QCD を用いて有限密度シミュレーションを行う場合，化学ポテンシャルの導入が符号問題を引き起こしてしまうことである．われわれはこのアルゴリズム的問題を解決し，幅広い密度領域における相構造解析を行う．特に，具体的目標として中性子星内部 (核密度の 5 倍〜10 倍程度) の相構造解明と状態方程式の決定を目指す．

Vol.2011-HPC-130 No.21
2011/7/27

4.2 ブラックホールと輻射流体シミュレーション

HA-PACS で実現可能な宇宙物理学分野の数値シミュレーションとして，以下の 2 つを計画している．

(1) 球状星団の銀河中心ブラックホールのフルスケールシミュレーション：
球状星団や銀河中心ブラックホールのシミュレーションは，一般的に知られる N 体シミュレーションとは異なり恒星の軌道を極めて高精度に解く必要がある衝突系の自己重力多体シミュレーションである．その為，従来の数値シミュレーションでは，GRAPE などの重力多体専用計算機を利用しても，恒星ひとつひとつを N 体計算の粒子で分解してシミュレーションすることは不可能であった．HA-PACS では，GPU を用いた重力計算の高速化だけではなく，ホスト計算機と GPU 間の高いバンド幅や GPU 間の高速データ転送などにより，世界で初めて球状星団や銀河中心ブラックホールのフルスケールシミュレーションを行い，球状星団の起源や銀河中心部での巨大ブラックホールの合体・成長の過程を調べることが可能になると考えられる．

(2) 銀河形成の輻射流体シミュレーション：
流体力学計算と共に輻射輸送計算を GPU を用いて加速することを目指す．これまでの銀河形成のシミュレーションでは，銀河外や銀河内の恒星からの輻射によるガスの加熱の効果が重要であるにもかかわらず，輻射輸送計算の計算コストが極めて大きいという理由で，無視されるか現象論的な取り扱いに終始していた．輻射輸送計算は輻射源からでる光の吸収・再放射を ray-tracing 法で計算する．その為，計算の並列性が高く，計算に必要なデータ量に対して演算量が極めて大きい compute-intensive な計算であることから GPU を使った高速化に極めて適している．HA-PACS では，GPU で輻射輸送計算を高速化することによって，未だに観測が及ばない宇宙の暗黒時代における天体形成・宇宙の再電離などの重要な未解決問題に迫ることが出来ると考えられる．

4.3 QM/MM 混合計算による生命科学計算

古典分動力学法 (MD) を用いた場合，現在の計算機では典型的なタンパク質酵素 (5 万原子) を数百 ns シミュレーションすると数ヶ月かかるので計算対象はこの領域に制限されている．しかしながら多くのタンパク質酵素の機構発現は数十μ s のタイムスケールであり，次世代計算機によりこのシミュレーションが可能となると，生命科学研究において非常に大きな意味を持つ．また，現行の時間スケールでは十分なサンプリングが行いえないため，初期構造は高分解能 X 線結晶構造が必須となっている．しかしながら次世代計算機で比較的低分解能の初期構造からでも計算が始められるようになれば，低解像度の構造しか得られていない多くの重要なタンパク質が計算対象となる（**図 7**）．

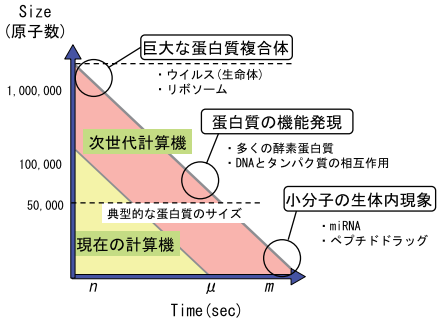

図 7 計算機性能と生命科学シミュレーションの進化

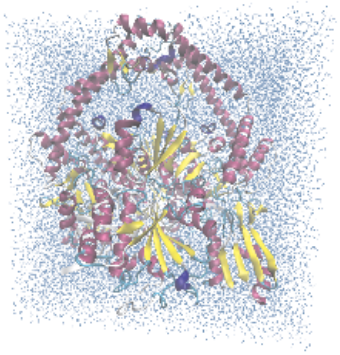

図 8 トポイソメラーゼ II の全体構造

一方で，系のサイズを大幅に大きくすることも可能であり，これにより巨大なタンパク質複合体が計算対象となってくる．これはウイルスやリボソームをまるごと取り扱えるレベル (〜数百万原子系) に到達する．一方，サイズを小さくし，長時間シミュレーションを行う事を考えれば，miRNA やペプチドドラックなど生体内で重要な機能を果たす比較的小さいサイズの生体高分子を数 ms シミュレーションできるようになると考えられる．これは生体内現象を再現できる時間スケールであり，構造予測や機能予測を計算のみで行うこと (in silico) が可能になると期待される．

量子古典混合法 (QM/MM) 法でも同様に，取り扱えるシミュレーション時間を大幅に大きくできる．これにより広い空間を探索できるようになり，従来困難であった正確な熱力学的統計量 (自由エネルギー等) の計算が可能になると期待される．生体分子については統計量を議論する事は非常に重要である．このように次世代計算機により生命科学で扱える問題が大きく広がるため，今後の実証が多いに期待される．

これらのため，現在以下の研究を進めている．(1) と (3) の研究は創薬や病気に直結し，(2) の研究は環境問題に深く関わっているため，非常に注目されている．

(1) トポイソメラーゼ II(TOPOII) による DNA 再結合の反応機構解明：
これは非常に大きな系であり（全体で 20 万原子）DNA を切断・結合させる事で DNA のトポロジーを変えている（**図 8**）．現在低解像度の結晶構造しか得られていないが，

重要な部分の構造を計算により求める事で，反応機構の詳細を初めて議論できると考えている．

(2) 一酸化窒素還元酵素 (NOR) の反応機構解明：
これは自然界における窒素サイクルの重要な 1 ステップを担っている．QM/MM 法を用いた高精度計算によって反応機構の解明に取り組んでいる．

(3) プリオンタンパク質 (PrP) の立体構造予測：
の PrP は 130 残基程度の比較的小さいポリペプチドであるが，立体構造が分かっていないため，構造モデルの検証について研究を行っている．

5. おわりに

現在，HA-PACS はベースクラスタ部の調達を推進中で，この部分の最終的な構成は 2011 年 9 月に確定する予定である．基本仕様は 2 節に示した通りであるが，部分的な変更はあり得る．ベースクラスタを用い，大規模 GPU アプリケーションの基本的なコーディングと性能評価を行う．HA-PACS/TCA 部は PEACH2 チップの開発とテストボード開発を 2011 年度中に完了し，次年度からは TCA を用いた本格的なアプリケーション開発を進める予定である．

謝辞 本プロジェクトの一部は文部科学省特別経費「エクサスケール計算技術開拓による先端学際計算科学教育研究拠点の充実」事業による．

参考文献

1) Dongarra, J., Meuer, H., Stromaier, E. and Simon, H.: TOP500 List, (online), available from ⟨http://www.top500.org/⟩.
2) Shimokawabe, T., Aoki, T., Muroi, C., Ishida, J., Kawano, K., Endo, T., Nukada, A., Maruyama, N. and Matsuoka, S.: An80-foldspeedup,15.0 TFlops full GPU acceleration of non-hydrostatic weather model ASUCA production code, *Pro. SC10*, pp.1–11 (2010).
3) Vetter, J.: Toward Exascale Computational Science with Heterogeneous Processing, 情報処理学会ハイパフォーマンスコンピューティングと計算科学シンポジウム論文集 (2010).
4) Boku, T., Itakura, K., Nakamura, H. and Nakazawa, K.: CP-PACS: A massively parallel processor for large scale scientific calculations, *Proc. ICS1997*, pp.108–115 (1997).
5) 朴　泰祐，梅村雅之，佐藤三久，高橋大介，中本泰史，須佐　元，森　正夫：FIRST - 第一世代天体の起源解明のための専用・汎用計算機融合型クラスタ，情報処理学会 HPC 研究会報告，2005-HPC-103 (2005).
6) Boku, T., Sato, M., Ukawa, A., Takahashi, D., Sumimoto, S., Kumon, K., Moriyama, T. and Shimizu, M.: PACS-CS: A large-scale bandwidth-aware PC cluster for scientific computations, *Proc. CCGrid2006* (2006).
7) NVIDIA: Tesla M2090 Announcement, (online), available from ⟨http://www.nvidia.com⟩.
8) NVIDIA and Mellanox: NVIDIA GPUDirect Technology - Accelerating GPU-based Systems, (online), available from ⟨http://www.mellanox.com/pdf/whitepapers/TBGPUDirect.pdf⟩.
9) 尾崎祐介，石川健一：GPU クラスタによる格子 QCD 計算，日本物理学会第 65 回年次大会一般講演 (2010).
10) Hanawa, T., Boku, T., Miura, S., Okamoto, T., Sato, M. and Arimoto, K.: Low-Power and High-Performance Communication Mechanism for Dependable Embedded Systems, *Proceedings of 2008 International Workshop on Innovative Architecture for Future Generation Processors and Systems*, pp.67–73 (2008).
11) PCI-SIG: *PCI Express Base Specification, Rev. 3.0* (2010).
12) Otani, S., Kondo, H., Nonomura, I., Uemura, M., Hayakawa, Y., Oshita, T., Kaneko, S., Asahina, K., Arimoto, K., Miura, S., Hanawa, T., Boku, T. and Sato, M.: An 80Gbps Dependable Communication SoC with PCI Express I/F and 8 CPUs, *2011 IEEE International Solid-State Circuits Conference*, pp. 266–267 (2011).
13) PCI-SIG: *PCI Express External Cabling Specification, Rev. 1.0* (2007).
14) Altera: *Stratix IV Device Handbook*. http://www.altera.co.jp/literature/lit-stratix-iv.jsp.