国際科学技術共同研究推進事業
（戦略的国際共同研究プログラム）

（研究領域「情報通信技術」）

研究課題名「ポストペタスケール・コンピューティングのための

フレームワークとプログラミング」

平成24年度実施報告書

佐藤三久

（筑波大学・計算科学研究センター　センター長・教授）

1. 研究実施内容

1－1. 研究実施の概要　公開

現在、最先端の計算科学に用いられる高性能計算システムの性能はペタフロップス(1 秒間に 1,000 兆回の演算能力)に達し、ポストペタスケールシステムとして、エクサスケールのシステムに向かおうとしている。我々の目標は、ペタスケールコンピューティングを超えてエクサスケールコンピューティングに到達する、超最先端の高性能コンピューティングへの道を開拓するべく、ソフトウェア技術、プログラミングモデルおよび言語を確立することである。

初年次においては、プログラミングモデルと言語の基本設計(タスク 1)、大規模システムのための実行時システム技術(タスク2)、アクセラレータ(演算加速機構)技術(タスク3)、並列アルゴリズムとライブラリ・フレームワーク(タスク4)など、ポストペタスケール・コンピューティングに必要な技術要素について、それぞれタスクに分けて、検討を進めた。個々のタスクにおける成果(deliverables)・見込みを示し、中間評価を受けた。

これを受けて、2年次の後半からは、初年度の各タスクで検討された設計を基に、並列プログラミング言語拡張仕様XcalableMP(XMP)とワークフロー言語YMLの統合やメニーコア向けの新しい通信レイヤを持つ実行時システムの研究を進めた(タスク 5)。本プロジェクトでは、ポストペタスケール・コンピューティングのための並列プログラミング基盤として、XMP を用いて、他のグループで研究開発されるコンポーネントへのインターフェースを与えるとともに、上位のプログラミングパラダイムとして、ワークフロー言語である YML を用いる。この XMP と YML の統合環境 FP2C (Framework for Post-Petascale Computing)の開発・評価を YML の開発を担当するベルサイユ大学グループと INRIA サクレグループと協力し、行った。

通常の大規模 PC クラスタの他、理研 AICS グループとの協力の下に、大規模環境での有効性を検証するために京コンピュータへの移植作業も進めた。理研 AICS のグループにおいては、その準備として、京コンピュータにおける XMP のプログラムの性能評価と最適化について検討した。特に、科学技術計算の典型的な通信パターンであるステンシル型の通信について、添え字式の冗長な変換の除去、実行時ルーチンのメモリコピーのマルチスレッド化を行い、良好な結果が期待できることが分かった。

また、これからのエクサスケールに向けたシステムのノードはメニーコアプロセッサになると思われる。東大グループは、メニーコア向けの低レベル通信ライブラリ DCFA(Direct Communication Facility for Accelerator) を設計・実装した。これを、INRIA ボルドー・グループで開発を行っている MPI 通信ライブラリ NewMadeline との統合について検討を進めている。XMP プログラムの通信レイヤは MPI であり、これにより、これから登場するメニーコアシステムにおいて XMP プログラムの高速化が期待される。

さらに東工大グループでは、昨年度の成果であるチェックポイント型の耐故障性インターフェースである FTI と INRIA で開発されている メッセージロンギング MPI ライブラリ Hydee を統合した。また、FTI と INRIA で提案されている障害予測機構の統合を行った。この統合型耐障害性システムでは、システムのログ情報から事前に障害を予測し、予測された障害の直前にチェックポイントをとることにより、復旧後の再計算の時間を大幅に短縮できる。東大グループにおいても、並列プロセスグルーピングによる効率的な耐故障機能の研究を進めている INRIA グループと協力し、効率の良いチェックポイント取得のための並列プロセスグルーピング手法を提案するなど、耐故障機能についての研究を進めている。

GPU などの演算加速機構に関しては、StarPU を利用した実行時システムと言語サポート、ライブラリについて研究を進めた(タスク 6)。StarPU は、INRIA ボルドー・グループが研究開発を進めている GPU と CPU に計算を自動的に振り分ける実行時システムである。筑波大グループではこれまで、XMP を拡張して GPU に対応させた

拡張仕様 XMP-dev を開発しているが、INRIA ボルドー・グループとの共同作業により、XMP-dev のプログラムの記述をほとんど変えることなく GPU だけでなくマルチコア CPU を含めた全リソースに計算を分配する処理系 XMP-dev/StarPU を開発した。また、東工大グループでは、FFM(Fast Multi-pole Method)などの具体的なカーネルアプリケーションにおいて、StarPU を含む GPU と CPU の負荷分散の方法について研究を進めている。その結果、StarPU が提供する単純なスケジューリングでは不適当なケースがあり、StarPU の開発グループにフィードバックすることにより、演算加速機構に関する実行時技術の高度化が期待される。

ポストペタスケール・コンピューティングに必要な大規模データ管理技術については、筑波大グループで研究開発を行ったマルチマスター型分散メタデータサーバ HGMDS と、INIRA レンヌのグループで研究開発を行ったロックフリー並列ストレージシステム BlobSeer をベースとして、両者を統合した広域ファイルシステム BlobSeer-WAN/HGMDS のプロトタイプ実装・評価を行っている。

数値計算ライブラリの研究については、本プロジェクトで提案している FP2C による階層的なプログラミングモデルを用いる階層的並列構造を持つ固有値解法などのハイブリッドアルゴリズムやポストペタスケール・コンピューティングに必要とされる GPU を活用したアルゴリズム、自動チューニング技術、アプリケーション・フレームワークに関する研究開発を行ってきた。これらについては、プログラミングモデルとしてどのように統合していくのかをフランス側と議論しつつ、インターフェースについて議論・検討を行ってきた(タスク7)。

筑波大グループでは、ポストペタスケールを想定した大規模計算環境において性能を発揮する分散型のアルゴリズム開発を行っている。これは階層型の固有値解法のアルゴリズムで、当該年度においては効率的に固有値を探索するための推定アルゴリズムと個々の線形計算カーネルの最適化を行った。線形計算カーネルとして、CNRS IRIT グループの開発した疎行列直接解法ライブラリ MUMPS を検討している。このアルゴリズムを YML と XMP の FP2C での実装を、ベルサイユ大学のグループとともに行っている。フランス側のグループでは同様な階層アルゴリズムのマルチリスタートの Arnoldi 法について開発を行っている。また、開発されたアルゴリズムの評価ベンチマークのために素粒子のQCD計算やフラグメント分子軌道法で現れる行列を収集し、実アプリケーションで現れる行列データの整備を行った。

理研 AICS グループでは、CNRS IRIT グループとともに、XMP から数値計算ライブラリを呼び出すためのインターフェースの設計を行った。このインターフェースは並列ライブラリの呼び出しに必要な分散を XMP で記述し、ライブラリに引き渡すもので、これにより、FP2C のプログラミングにおいて、この数値計算ライブラリのタスク(タスク 7)の成果を全体に統合することができる。

東京大学グループでは、悪条件問題向け並列前処理手法に関する研究開発および有限要素法等の不規則データ構造を有する手法の GPU 上への実装，疎行列反復解法であるリスタート付き GMRES の、リスタート周期に関する自動チューニング(AT)方式の研究開発を行った。特に、ポストペタスケールのシステムにおいては、システムの不確定要素が多くなり、実際に試行を行って各種のパラメータを決定する自動チューニング技術が重要になる。当該年度においてはフランス側が開発した GMRES(m)法のリスタート周期調整の AT 技術である、階層メモリの大きさを考慮して最大リスタート周期を決める方式や日本側で開発した最大／最小比率(MM 比)による方式を検討した。この他、有限要素法について GPU を有効に利用するために INRIA ボルドーチームとともに StarPU を適用する技術についても検討した。

ポストペタスケール・コンピューティングにおいては、特定のドメインのアプリケーションを対象に高レベルの記述を行い、それを最適な並列コードに変換することも有効な方法の一つであり、アプリケーション・フレームワークと呼ばれる。京大グループでは、ループボディとループ構造を分離した記述法によるアプリケーション・フレーム

ワークに関する研究開発を行い、フレームワークによる記述を変換する場合の並列化技法の選択と性能の関係を、線形ソルバー、FDTD 法、および PIC シミュレータを対象に、適切な技法選択を行うための改良を行った。その選択には自動チューニング的な技法を取り入れるなどの工夫も行っている。

フランス側とのスケジュールの調整により、以上のタスクについて、2012 年(平成 24 年)末までとし、当初計画を1年(10 か月)延長して取り組んだ。

2. 研究実施体制 公開

2－1. 日本側の研究実施体制

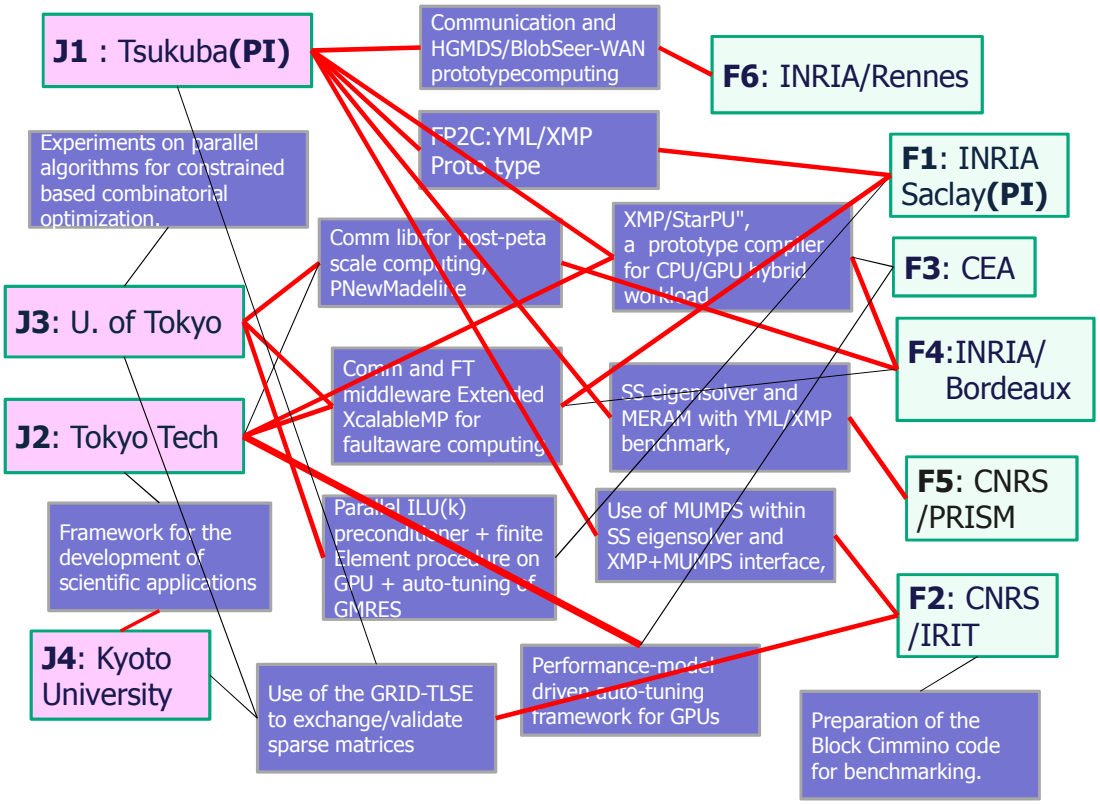
研究代表者/ 主な共同研究者	氏名	所属	所属部署	役職
研究代表者	佐藤三久	筑波大学	計算科学研究センター	センター長・教授
主な共同研究者	朴　泰祐	筑波大学	計算科学研究センター	教授
主な共同研究者	櫻井鉄也	筑波大学	システム情報系	教授
主な共同研究者	建部修見	筑波大学	計算科学研究センター	准教授
主な共同研究者	中島研吾	東京大学	情報基盤センター	教授
主な共同研究者	石川　裕	東京大学	情報基盤センター	センター長・教授
主な共同研究者	松岡　聡	東京工業大学	学術国際情報センター	教授
主な共同研究者	中島　浩	京都大学	学術情報メディアセンター	センター長・教授
主な共同研究者	村井　均	理化学研究所	計算科学研究機構	研究員

2－2. 相手側の研究実施体制

研究代表者/ 主な共同研究者	氏名	所属	所属部署	役職
研究代表者	Serge　Petiton	INRIA-SACLAY		教授
主な共同研究者	Michel　Daydé	CNRS　IRIT, Toulouse		教授
主な共同研究者	Christophe Calvin	CEA-DEN　SACLAY		上級研究員
主な共同研究者	Raymond Namyst	INRIA　Bordeaux		教授
主な共同研究者	Gabriel Antoniu	INRIA　Rennes		上級研究員
主な共同研究者	Nahid Emad	CNRS PRISM, ベルサイユ大学		教授

2－3. 両国の研究実施体制

以下に、日本の研究グループ(J)とフランスの研究グループ(F)と研究テーマの関係を示す。赤線が、具体的な共同研究、細線が間接的な協力関係を示す。



3. 原著論文発表　公開

3－1. 原著論文発表

① 発行済論文数

	うち、相手側チームとの共著（※）
国内誌　2 件	（　0 件）
国際誌 17 件	（　3 件）
計　19 件	（　3 件）

※本共同研究の相手側チーム研究者との共著に限る

1. T. Nomizu, D. Takahashi, J. Lee, T. Boku, M. Sato, "Implementation of XcalableMP Device Acceleration Extention with OpenCL", Proc. of PLC2012 (with IPDPS2012), Shanghai, CD-ROM.

＊ 2. T. Odajima, T. Boku, T. Hanawa, J. Lee, M. Sato, "GPU/CPU Work-Sharing with Parallel Language XcalableMP-dev for Parallelized Accelerated Computing", Proc. of P2S2-2012 (with ICPP2012), Pittsburgh, CD-ROM.
YML と XMP を統合したモデルに加えて、GPU を階層的にサポートするための機構をフランスの INRIA ボルドー・グループが開発した StarPU と XMP の GPU 拡張 XMP-dev でサポートすることを示し、本プロジェクトの中心的なコンセプトを述べた論文。

3. Shinnosuke Yokota, Tetsuya Sakurai, "A projection method for nonlinear eigenvalue problems using contour integrals", JSIAM Letters, Vol.5, pp.41-44, 2013.

4. Akira Imakura, Tetsuya Sakurai, Kohsuke Sumiyoshi and Hideo Matsufuru, “A Parameter Optimization Technique for a Weighted Jacobi-Type Preconditioner”, JSIAM Letters, Vol.4, pp.41-44, 2012.

5. Michihiro Naito, Hiroto Tadano and Tetsuya Sakurai, "A modified Block IDR(s) method for computing high accuracy solutions", JSIAM Letters, Vol.4, pp. 25-28, 2012.

6. Masae Hayashi and Kengo Nakajima, “OpenMP/MPI Hybrid Parallel ILU(k) Preconditioner for FEM Based on Extended Hierarchical Interface Decomposition for Multicore Clusters”, Proceedings of 10th International Meeting on High-Performance Computing for Computational Science (VECPAR 2012) , 2012

7. Kengo Nakajima, "OpenMP/MPI Hybrid Parallel Multigrid Method on Fujitsu FX10 Supercomputer System", IEEE Proceedings of 2012 International Conference on Cluster Computing Workshops, 199-206, IEEE Digital Library: 10.1109/ClusterW.2012.35, 2012

8. Takahiro Katagiri, Pierre-Yves Aquilanti and Serge Petiton, "A Smart Tuning Strategy for Restart Frequency of GMRES(m) with Hierarchical Cache Sizes", Proceedings of iWAPT2012, 2012

9. Satoshi Ohshima, Masae Hayashi, Takahiro Katagiri, Kengo Nakajima, “Implementation and Evaluation of 3D Finite Element Method Application for CUDA”, Proceedings of 10th International Meeting on High-Performance Computing for Computational Science (VECPAR 2012) , 2012

10. Min Si and Yutaka Ishikawa, "Design of Direct Communication Facility for Manycore-based Accelerators," Proceedings of CASS2012 in conjunction with the 20th International Parallel and Distributed Processing Symposium (IPDPS12), 2012
11. 河合直聡，岩下武史，中島浩,”ブロック化赤−黒順序付け法による並列マルチグリッドポアソンソルバ”，情報処理学会論文誌：コンピューティングシステム, Vol. 5, No. 3, pp. 1-10, 2012.
12. Yohei Miyake, Hideyuki Usui, Hirotsugu Kojima, and Hiroshi Nakashima, “Plasma Particle Simulations on Stray Photoelectron Current Flows Around a Spacecraft”, J. Geophysics Research, Vol. 117, No. A09210, pp. 1-13, 2012.
13. 南武志，岩下武史，中島浩,”冗長な計算を伴わない 3 次元 FDTD 法の時空間タイリング”，情報処理学会論文誌：コンピューティングシステム, Vol. 6, No. 1, pp. 56-65, 2013
14. Irina Demeshko, Satoshi Matsuoka, Naoya Maruyama, Hirofumi Tomita. “Ultra-high Resolution Atmospheric Global Circulation Model NICAM on Graphics Processing Unit”, In Proc. of the 2012 International Conference on Parallel and Distributed Processing Techniques and Applications (PDTPA’12), Jul. 2012.
15. Irina Demeshko, Satoshi Matsuoka, Naoya Maruyama and Hirofumi Tomita. “Multi-GPU implementation of the NICAM atmospheric model”, In Proc. of Tenth International Workshop on Algorithms, Models and Tools for Parallel Computing on Heterogeneous Platforms (HeteroPar’2012) in conjunction with EuroPar’2012, Aug. 2012.
16. L. Bautista Gomez, B. Nicolae, N. Maruyama, F. Cappello, S. Matsuoka. “Scalable Reed-Solomon-based Reliable Local Storage for HPC Applications on IaaS Clouds”, In Proc. of International European Conference on Parallel and Distributed Computing (EuroPar 2012), Aug. 2012.
17. Leonardo Bautista Gomez, Thomas Ropars, Naoya Maruyama, Franck Cappello, Satoshi Matsuoka. “Hierarchical Clustering Strategies for Fault Tolerance in Large Scale HPC Systems”, In Proc. of IEEE Cluster 2012, IEEE Press, Sep. 2012.
18. Kento Sato, Adam Moody, Kathryn Mohror, Todd Gamblin, Bronis R.de Supinski, Naoya Maruyama, Satoshi Matsuoka. “Design and Modeling of a Non-blocking Checkpointing System”, In Proc. of 2012 ACM/IEEE International Conference for High Performance, Networking, Storage, and Analysis (SC’12), Salt Lake City, IEEE Press, Nov. 2012., pp.1-10.
19. Akira Nukada, Kento Sato and Satoshi Matsuoka. “Scalable Multi-GPU 3-D FFT for TSUBAME 2.0 Supercomputer”, In Proc. of 2012 ACM/IEEE International Conference for High Performance, Networking, Storage, and Analysis (SC’12), Salt Lake City, IEEE Press, Nov. 2012.

② 未発行論文数

	うち、相手側チームとの共著（※）
国内誌　0 件	（　0 件）
国際誌　7 件	（　2 件）
計　7 件	（　2 件）

※本共同研究の相手国チーム研究者との共著に限る

1. Tetsuya Sakurai, Yasunori Futamura and Hiroto Tadano, "Efficient parameter estimation and implementation of a contour integral-based eigensolver", J. Alg. Comput. Tech., (accepted)
2. Masae Hayashi and Kengo Nakajima, “OpenMP/MPI Hybrid Parallel ILU(k) Preconditioner for FEM Based on Extended Hierarchical Interface Decomposition for Multicore Clusters” , Selected Papers of 10th International Meeting on High-Performance Computing for Computational Science (VECPAR 2012) , Lecture Notes in Computer Science, 7851, Springer, 2013 (in press)
3. Takahiro Katagiri, Pierre-Yves Aquilanti and Serge Petiton, "A Smart Tuning Strategy for Restart Frequency of GMRES(m) with Hierarchical Cache Sizes", Selected Papers of 10th International Meeting on High-Performance Computing for Computational Science (VECPAR 2012) , Lecture Notes in Computer Science, 7851, pp.314-328, Springer, 2013 (in press)
4. Satoshi Ohshima, Masae Hayashi, Takahiro Katagiri, Kengo Nakajima, “Implementation and Evaluation of 3D Finite Element Method Application for CUDA”, Selected Papers of 10th International Meeting on High-Performance Computing for Computational Science (VECPAR 2012) , Springer LNCS 7851, pp.140-148, 2013 (in press)
5. Min Si, Yutaka Ishikawa and Masamichi Takagi, "Direct MPI Library for Intel Xeon Phi co-processors," Proceedings of CASS2013 in conjunction with the 21st International Parallel and Distributed Processing Symposium (IPDPS13), 2013 (in press).
6. M. S. Bouguerra, Anna Gainaru, Franck Cappello, Leonardo Bautista Gomez, Naoya Maruyama and Satoshi Matsuoka, “Improving the Computing Efficiency of HPC systems using a Combination of Proactive and Preventive Checkpointing”, Proceedings of IEEE IPDPS 2013, Boston, MA, the IEEE Press, May 2013, pp.1-10. (to appear)
7. Abdelhalim Amer, Naoya Maruyama, Miquel Pericas, Kenjiro Taura, Rio Yokota, and Satoshi Matsuoka, “ Bulk-Synchronous and Data-Driven Execution Models on Multi-Core Architectures: Case study of the FMM”, In Proc. of International Supercomputing Conference (ISC’13), Jun. 2013, (to appear).