

4. 超高速計算システム研究部門

1. メンバ

教授 佐藤 三久、朴 泰祐

准教授 建部 修見、高橋 大介、塙 敏博

助教 多田野 寛人

2. 概要

本研究グループは、高性能計算システムアーキテクチャ、省電力システムアーキテクチャ、並列数値処理の高速化研究、広域分散環境におけるデータ共有を中心とするグリッド計算技術等の研究を行っている。研究分野は計算機アーキテクチャ研究グループとグリッド研究グループに分かれるが、常に両グループ間で共同研究を行っているため、以下の報告では両者を区別せずまとめて記述する。

3. 研究成果

【省電力／高性能／ディペンダブル並列システムに関する研究】

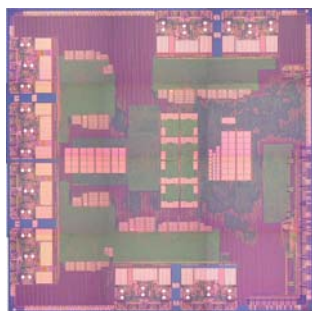
(1) クラウド技術を用いた高信頼並列ソフトウェアのテスト環境 D-Cloud の開発(佐藤)

クラウドコンピューティングシステムを用いての計算資源の管理をベースとして、テストシステムの構築、フォルトインジェクションによるハードウェア故障エミュレーション、様々なテストケース・シナリオの自動化機能を提供するクラウドコンピューティングシステム D-Cloud を提案し、そのプロトタイプを開発した。D-Cloud を用いることにより短期間で効率的なプログラムテストが可能となる。なお、JST-CREST「省電力でディペンダブルな組み込み並列システム向け計算プラットフォーム」プロジェクトは、中間評価を受けた(下記(2), (3)も同様)。

(5) 高性能・省電力・耐故障並列システムリンクに関する研究(朴, 塙)

ルネサステクノロジ社との共同研究により、PCI-Express gen.2 テクノロジーに基づく並列システム向け相互結合網リンク PEARL の開発を行い、これを実現する通信スイッチチップのプロトタイプである PEACH チップの実装を行った。PEACH は4レーンの PCI-Express gen.2 リンクを4ポートと、これらの間の通信を制御する4コア

の組み込みプロセッサを持つ通信チップである。PC クラスタ等の PCI-Express バスに接続することにより、高性能で耐故障性を持つ通信リンクを省電力で実現できる。過去3年間に渡り設計と開発を行い、PEACH チップの実装に成功した。完成した PEACH チップの物理特性と基本動作について評価・確認を行い、基本コンセプトに基づく応用に問題がないことを確認した。



PEACH チップフロアプラン PEACH チップを搭載した PCI-Express テストボード

(6) Ethernet トランキンクに基づくクラスタ間結合ネットワーク(朴, 埴)

対価格性能比の高い通信ネットワークである Gigabit Ethernet を束ね、クラスタ内ネットワークとして用いる RI2N の開発を継続して行った。今年度の成果は、各経路におけるトランキンク本数が一定でない形状においても、各リンクに対して最適な通信量配分を行うアルゴリズムを組み込み、さらに時間方向における通信の粗密がある場合でも、通信モニタ機能により常に最適化された通信バランスを保つように改良した。様々なトポロジや通信パターンにおいて、従来の RI2N よりも最大で 40%程度性能が向上することを確認した。

【次世代並列処理言語 XcalableMP の研究開発】(佐藤, 朴)

E-Science プロジェクト「並列プログラミング言語に関する研究開発」において、分散メモリ構成を基本とする大規模並列処理システムにおける並列 HPC アプリケーションのため、並列プログラミング言語 XcalableMP (XMP) の開発を行っている。XMP は OpenMP のようなディレクティブベースの言語で、グローバルビューでの並列化が簡易に行うことができる。これに加えて、PGAS 的な概念に基づき、ノードを中心としたローカルビューをプログラミングをサポートとしている。モデルで提供される動作を明確に定義することによって、性能チューニングが可能なフレームワークを提供する。当該年度は、SC09 において、HPCC ベンチマーク Class2 に結果を提出し、ファイナリストにのこり、Honorable Mention を受けた。

【仮想マシン環境における省電力化技術の研究とデータセンター電力ベンチマーク】(佐藤)

NEDO グリーン IT「エネルギー利用最適化データセンタ基盤技術の研究開発／データセンタのモデル設計と総合評価」において、データセンター電力ベンチマークの検討を行うとともに、現在、データセンターの運用で注目されている仮想マシン環境において、省電力化を行うための研究を進めた。

【大規模広域分散ファイルシステム及びグリッド／クラウド技術に関する研究】

(1) 大規模環境における広域ファイルシステムの評価(建部)

オープンソースで研究開発を進めている Gfarm ファイルシステムの大規模広域環境における性能評価を行った。国内 13 拠点 (はこだて未来大, 東北大, 筑波大, 国情研, 東大, 東工大, 早稲田, 慶応, 京大, 神戸大, 広大, 九大, 九工大) の計 239 ノードの PC クラスタ群を利用し, 147 TByte の広域ファイルシステムを構築した。メタデータ操作の性能測定では, 11 拠点のクライアントからの同時アクセスで毎秒 3,570 操作の性能を実現した。同一共有データの複数拠点, 複数クライアントからの読出しでは, 7 拠点の計 56 クライアントの同時読み出しまでクライアント数を変えて計測したが, 読込み性能はクライアント数に対しスケールし, 全体で毎秒 5,133MByte の読込み性能を達成した。メタデータサーバの性能ボトルネックとはなっておらず, さらにクライアント数が増加した場合も引き続き性能がスケールすることが見込まれる。本性能評価結果については, New Generation Computing 誌に投稿し, 採録が決定した。

(7) JLDG におけるグループ内広域ファイル共有の実現(建部, 佐藤)

素粒子物理研究グループとの共同研究により, 国内の素粒子物理学研究者における大規模データ共有基盤 JLDG の構築, 運用を行っている。JLDG は筑波大を中心とし, KEK, 金沢大, 京大, 阪大, 広島大で構成している Gfarm 広域ファイルシステムをベースに構築, 運用が行われている。これまでの運用では, スパコンで計算した公開データの共有を行っていたが, 本年度はそれぞれの研究グループの日々のデータの拠点間での共有について運用を開始した。研究グループ内におけるデータ共有を実現するためには, 組織とは無関係な研究グループによるアクセス制御の実現が必須となる。そのため, 仮想組織管理サーバが管理する研究グループにおけるアクセス制御が可能となるよう, 広域ファイルシステムの開発を行った。これにより, 公開データの共有だけではなく日々の研究データの共有も可能となった。

(8) T2K グリッド連携, 9 大学グリッド連携(建部, 佐藤, 朴)

筑波大, 東大, 京大の三大学における T2K システム間の連携を図るため, 筑波大 CCS 認証局の運用を行っている。グリッド環境としては, 単一認証におけるログイン, ファイル転送, 広域分散ファイルシステムの運用を行っている。本年度は単一認証を利用した遠隔からのジョブ起動について検討を行った。現在, GRAM2,

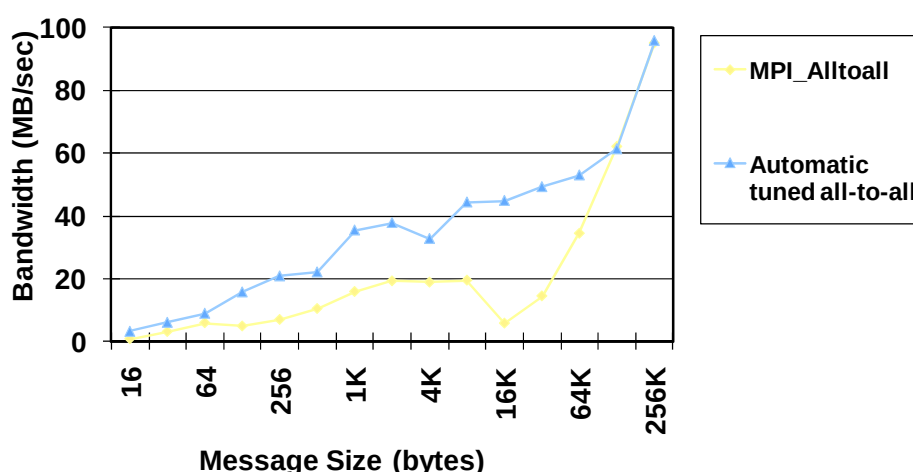
GRAM4 および OGSA-BES というジョブ起動の規格があるが、現在の T2K 筑波スーパーコンピュータシステムの運用を変えないで導入可能な規格は GRAM2 であり、GRAM2 について遠隔からジョブ起動が可能となるようシステム開発を行った。これにより、標準のグリッドミドルウェアを用いた T2K 筑波システムに対するジョブ起動が可能となった。また、9 大学のスパコンセンターにおけるグリッド連携に参加し、T2K 筑波システムの計算ノードへのジョブ起動を可能にするとともに、9 大学のアカウントを発行するためのグリッドパックにおいて、T2K 筑波システムのアカウント作成フローに組み込み、自動的にアカウント設定が可能となるよう国情研のサーバとの連携システムの開発を行った。これにより、グリッドパックにより申請されたアカウントは通常のアカウント作成フローでグリッド連携での利用も可能となった。

【高性能並列数値計算に関する研究】

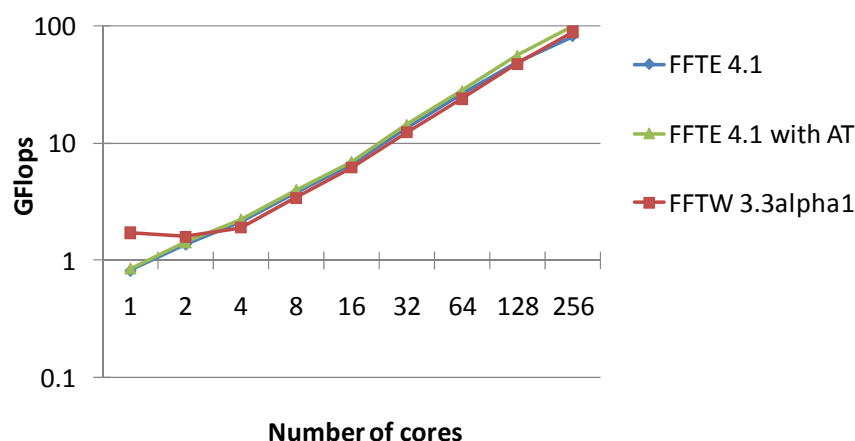
(1) 並列高速フーリエ変換(FFT)の自動チューニング手法に関する研究(高橋)

科学技術計算で広く用いられている並列 FFT の性能を改善するために、自動チューニング手法の研究を行った。性能に関わるパラメータとしては、全対全通信方式、基底、ブロックサイズがあるが、それぞれに対して最適な性能パラメータを探索することで、現在ソースコードを公開している並列 FFT ライブラリである FFTE の性能を向上させることができた。

T2K筑波システムにおける全対全通信の性能
(64ノード, 1024コア, 1024MPIプロセス)



T2K筑波システムにおける並列三次元FFTの性能 ($N_1 \times N_2 \times N_3 / P = 2^{24}$)

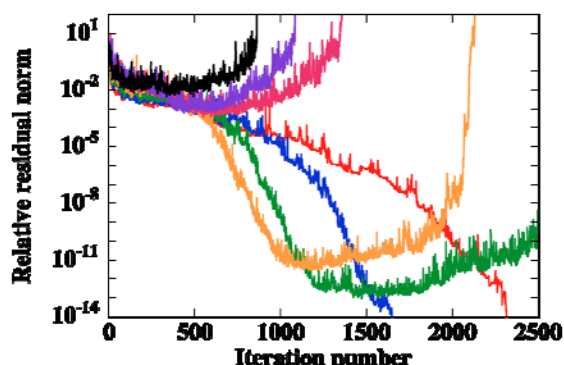


(9) T2K 筑波システムによる円周率 2 兆 5769 億 8037 万桁計算(高橋)

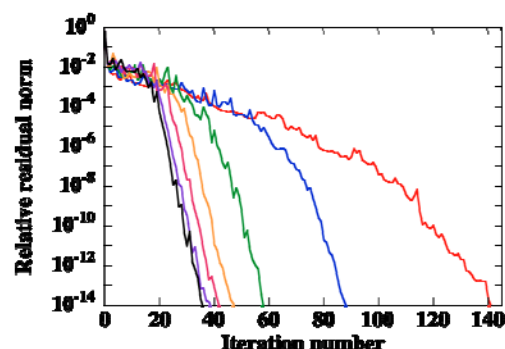
平成 20 年 6 月に導入された T2K 筑波システムにおいて, 640 ノードを用いて円周率 2 兆 5769 億 8037 万桁を計算した. 主計算が 29 時間 5 分, 検証計算が 44 時間 30 分で完了した. その間一度もシステムダウンはなく, T2K 筑波システムの高い信頼性を実証することができた.

(10) Block Krylov 部分空間反復法に関する研究(多田野)

複数本の右辺ベクトルをもつ連立一次方程式 $AX = B$ を高速・高精度で解くための Block Krylov 部分空間反復法の研究を行った. 従来の解法では, 右辺ベクトル数 L を増やすと反復回数は減少するが, 解の精度が悪化していた. 新たに開発した Block BiCGGR 法により, 反復回数の減少と高精度近似解生成の両立を達成した. また, 格子 QCD 計算で現れる連立一次方程式に対しては, Jacobi 前処理付き Block BiCGGR 法が有効であることを明らかにした.



(a) 前処理無し.



(b) Jacobi 前処理付き.

格子 QCD における Block BiCGGR 法の相対残差履歴. ■: $L = 1$, ■: $L = 2$, ■: $L = 4$,

■: $L = 6$, ■: $L = 8$, ■: $L = 10$, ■: $L = 12$.

4. 研究業績

< 学術雑誌論文 >

1. 今田 貴之, 佐藤 三久, 木村 英明, 堀田 義彦: 分散型 Web サーバでの負荷変動を考慮した省電力化のためのノード状態制御, 情報処理学会論文誌コンピューティングシステム, Vol.2, No.2, pp.75-88
2. 米元 大我, 埴 敏博, 三浦 信一, 朴 泰祐, 佐藤 三久, "トラフィック量に適應する非対称マルチリンク Ethernet トランッキング", 情報処理学会論文誌コンピューティングシステム, Vol.3, No.1, pp.25-37
3. Barbara M. Chapman, Jesus Labarta, Vivek Sarkar, Mitsuhsa Sato: Programmability Issues. IJHPCA 23(4): 328-331 (2009)
4. Jun-Ichi Iwata, Daisuke Takahashi, Atsushi Oshiyama, Taisuke Boku, Kenji Shiraishi, Susumu Okada and Kazuhiro Yabana: A massively-parallel electronic-structure calculations based on real-space density functional theory, Journal of Computational Physics, Vol. 229, No. 6, pp. 2339-2363
5. 佐藤佳州, 高橋大介: モンテカルロ木探索によるコンピュータ将棋, 情報処理学会論文誌, Vol. 50, No. 11, pp. 2740—2751
6. T. Sakurai, J. Asakura, H. Tadano, and T. Ikegami. Error Analysis for a Matrix Pencil of Hankel Matrices with perturbed Complex Moments. JSIAM Letters, Vol. 1, pp. 76--79, 2009.

7. Asakura, T. Sakurai, H. Tadano, T. Ikegami, and K. Kimura. A Numerical Method for Nonlinear Eigenvalue Problems Using Contour Integral. JSIAM Letters, Vol. 1, pp. 52--55, 2009.
8. H. Tadano, T. Sakurai, and Y. Kuramashi. Block BiCGGR: A New Block Krylov Subspace Method for Computing High Accuracy Solutions. JSIAM Letters, Vol. 1, pp. 44--47, 2009.
9. 多田野寛人, 櫻井鉄也. 周回積分法に対する Block Krylov 部分空間反復法の適用と分子軌道計算への応用. 情報処理学会論文誌 コンピューティングシステム, Vol. 2, No. 2, pp. 10--18, 2009.

< 査読付き国際会議発表 >

1. Ryo Kanbayashi, Mitsuhsa Sato: A Distributed Architecture of Sensing Web for Sharing Open Sensor Nodes. GPC 2009: 340-352
2. Miwako Tsuji, Mitsuhsa Sato: Performance Evaluation of OpenMP and MPI Hybrid Programs on a Large Scale Multi-core Multi-socket Cluster, T2K Open Supercomputer. ICPP Workshops 2009: 206-213
3. Shin'ichi Miura, Toshihiro Hanawa, Taiga Yonemoto, Taisuke Boku, Mitsuhsa Sato: RI2N/DRV: Multi-link ethernet for high-bandwidth and fault-tolerant network on PC clusters. CAC2009 (in IPDPS 2009): 1-7
4. Toshihiro Hanawa, Mitsuhsa Sato, Jinpil Lee, Takayuki Imada, Hideaki Kimura, Taisuke Boku: Evaluation of Multicore Processors for Embedded Systems by Parallel Benchmark Program Using OpenMP. IWOMP 2009: 15-27
5. Jinpil Lee and Mitsuhsa Sato: Reliable Software Distributed Shared Memory using Page Migration, The Fifteenth International Conference on Parallel and Distributed Systems (ICPADS'09), pp.456-463
6. Taiga Yonemoto, Shin'ichi Miura, Toshihiro Hanawa, Taisuke Boku, and Mitsuhsa Sato: Flexible Multi-link Ethernet Binding System for PC Clusters with Asymmetric Topology, The Fifteenth International Conference on Parallel and Distributed Systems (ICPADS'09), pp.49-56
7. Takayuki Imada, Mitsuhsa Sato and Hideaki Kimura: Power and QoS Performance Characteristics of Virtualized Servers, in Proceedings of Energy Efficient Grids, Clouds and Clusters Workshop(E2GC2) in the 10th IEEE GRID 2009, pp.232-240
8. Daisuke Takahashi: Automatic Tuning for Parallel 3-D FFT with 2-D Decomposition, 2010 SIAM Conference on Parallel Processing for Scientific Computing, Grand Hyatt Seattle, Seattle, Washington, USA, February 25, 2010.

9. T. Sakurai, J. Asakura, H. Tadano, T. Ikegami and K. Kimura. A Method for Finding Zeros of Polynomial Equations Using a Contour Integral Based Eigensolver. Proc. Symbolic Numeric Computation 2009, pp. 143--147, 2009.

<招待講演>

1. Daisuke Takahashi: Parallel Implementation of Multiple-Precision Arithmetic and 2.576 Trillion Digits of Pi Calculation on a Massively Parallel Cluster of Multi-Core Processors, Workshop on Ultra Performance and Dependable Acceleration Systems (held in conjunction with PDCAT'09), Gakushi-kaikan, Hiroshima University, Higashi-Hiroshima, December 11, 2009.

<ポスター・シンポジウム・研究会>

1. Masahiro Tanaka, Osamu Tatebe, "Impact of Gfarm, a Wide-area Distributed File System, upon Astronomical Data Analysis and Virtual Observatory", Poster, Astronomical Data Analysis Software and System XIX (ADASS2009), 2009
2. 米元大我, 塙敏博, 三浦信一, 朴泰祐, 佐藤三久:トラフィック量に適応する非対称マルチリンク Ethernet トランッキング 第21回コンピュータシステム・シンポジウム(ComSys 2009)論文集, pp. 21-30, 2009
3. 辻 美和子, 佐藤 三久:大規模 SMP クラスタにおける OpenMP/MPI ハイブリッドプログラムの性能評価 先進的計算基盤システムシンポジウム(SACIS 2009)論文集, pp.265-272, 2009
4. 緑川博子, 斉藤和広, 佐藤三久, 朴泰祐:クラスタをメモリ資源として利用するための MPI に基づいた高速大容量仮想メモリ 電子情報通信学会技術研究報告. CPSY, コンピュータシステム 109(168), 37-42, 2009
5. 坂西隆之, 小泉仁志, 神林亮, 佐藤三久:プログラムテスト環境を提供するクラウドコンピューティングシステムの検討 情報処理学会研究報告, 2009-OS-112(19), pp. 1-8, 2009
6. 米元大我, 三浦信一, 塙敏博, 朴泰祐, 佐藤三久:非対称な形状に適応する高バンド幅 multi-link Ethernet 情報処理学会研究報告, 情報処理学会研究報告, 2009-OS-112(19), pp. 1-9, 2009
7. 今田貴之, 佐藤三久, 木村英明:仮想計算機環境における省電力化を目的としたサーバ資源制御 情報処理学会研究報告, 2009-HPC-121(36), pp. 1-8, 2009
8. 李珍泌, 朴泰祐, 佐藤三久:分散メモリ向け並列言語 XcalableMP コンパイラの試作と評価 情報処理学会研究報告, 2009-H PC-121(6), pp. 1-10, 2009