

IV-1 計算機アーキテクチャ分野

1. メンバ

教授 佐藤 三久
教授 朴 泰祐
准教授 建部 修見
准教授 高橋 大介
准教授 埴 敏弘
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

【1】高性能省電力計算システムアーキテクチャに関する研究

（1）省電力ディペンダブルサーバに関する研究(佐藤)

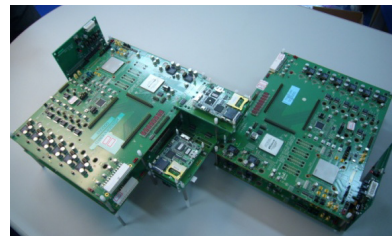
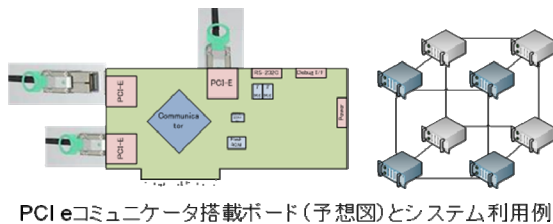
分散 web サーバの省電力化及び高信頼 DSM の開発を行った。分散 web サーバ省電力化においては、複数台の web サーバに対するアクセス頻度や負荷に応じ、適宜サーバの電力制御(DVFS 等)を行い、全体の電力性能最適化を行う。高信頼 DSM では、相互結合網で結合されたノード間で、リモートページングと同じ技術を用いたメモリページのコピーを行い、他のノード上にチェックポイントデータを作成する。これを自動的に行うことにより、HDD を持たない組み込み機器等の省電力並列システム上で信頼性の高い DSM の構築を行う。これらの基本システムの設計・実装を行った。

（2）高性能・耐故障・省電力並列システムリンクに関する研究(朴, 埴)

ルネサステクノロジとの共同研究により、PCI-e gen2 の仕様をベースに、並列ノード間の短距離結合網の開発を行った。同リンクは PEARL と呼ばれ、PCI-e gen2 の multi-lane (最大4)と、拡張された PCI-e 通信プロトコルを用いることにより、リンク故障対応と省電力制御を可能とする。PEARL の実装として、各並列ノードに PEACH と呼ばれる通信リンク制御チップ(ルータ)を設け、これらの間を相互に結合する。今年度は PEACH の試作版として、PEARL の PHY 部分とトランスポートレイヤ制御部分を FPGA 化したテストボードを実装し、この上で基本的な通信アルゴリズム、耐故障アルゴリズムの研究開発を行った。

PEARL を用いたノード間は、通信レーン故障に対応し、さらに現在開発中の4チャンネルの PEACH を用い、複数経路を持つ相互結合網を構築すれば、小規模システムにおけるマルチパス(複数経路)が実現でき、う回路

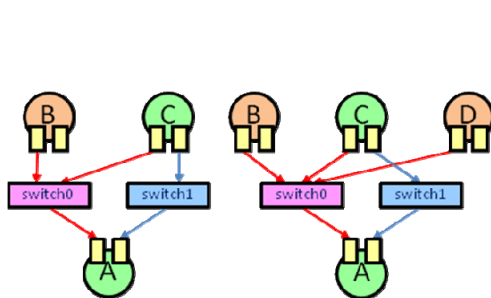
を利用した通信性の確保も実現可能である。さらに、1つのノード上の PEARL に slave device を接続することにより、ノード故障時にそのノードに結合されている PEACH 間結合でメッシュやキューブ系ネットワークが構築可能である。



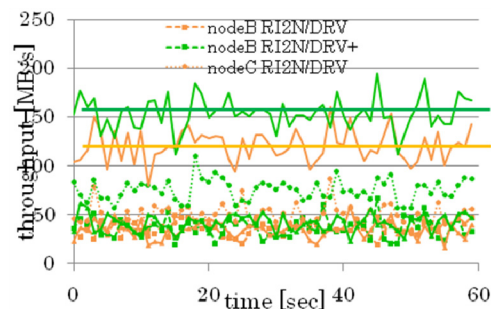
PCI-e に基づく PEARL リンクの試作 FPGA ボード

(3) HPC クラスタ向け耐故障・高性能コモディティネットワーク(朴, 塙)

研究室や小規模 HPC 向け PC クラスタでは、高価な SAN ではなく GbE が多用されている。これらのシステムでは、相互結合網の費用が安価ではあるが、絶対的バンド幅の不足が問題となり、また GbE の持つ性能的不安定性や一時故障を含む故障率の高さが問題となる。複数の GbE リンクを束ねてこれを解決する RI2N (Redundant Interconnection with Inexpensive Network) に対し、完全対称型でなく一部に非対称性を含むマルチリンク GbE 結合に関する改良を行った。改良型 RI2N では、相互結合網の一部が単一リンクで構成され、残りが複数リンクの場合でも、その非対称性を検出し、動的なパケット振り分けに基づく最適化を行った。結果として、従来の RI2N に比べ、対称結合の場合は全く性能は変わらないのに対し、非対称結合の場合は最大で 30% 以上の性能を向上するシステムを開発した。



非対称マルチリンク結合の例



右側のネットワーク構成におけるパケット分布とリンク毎バンド幅

【2】並列処理言語, 並列処理システム性能評価に関する研究

(1) 次世代向け超並列処理言語の研究開発(佐藤, 朴)

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、XcalableMP (XMP) と名付けられた言語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あるが、実装系では共有メモリではなく MPI におけるメッセージパッシングの呼び出しを想定する。しかし、PGAS 的な概念に基づき、並列ノード集合としてはグローバルビューに基づく変数インデキシング、MPI 的な通信においてはローカルビューに基づくノード内メモリ参照の両方を許し、基本的にはディレクティブに依存する並列化を支援しつつ、性能チューニングを行う上級者向け、ライブラリ開発者向けには MPI 通信が間接的

に見えるような形でのプログラム最適化手段を提供する. XMP は分散メモリに基づく通信コスト削減と高いプログラマビリティを基本とし, 様々なレベルでの並列化支援を行う. 本年度は言語のプロトタイプ開発を行った.

```
#pragma xmp template T[10]
#pragma xmp distributed T[block]
```

```
int array[10][10];
#pragma xmp aligned array[i][*] to T[i]
```

```
main(){
  int i, j, res;
  res = 0;
  #pragma xmp loop on T[i] reduction(+:res)
  for(i = 0; i < 10; i++){
    for(j = 0; j < 10; j++){
      array[i][j] = func(i, j);
      res += array[i][j];
    }
  }
}
```

data distribution

add to the serial code :
incremental parallelization

work sharing and data
synchronization

```
int array[YMAX][XMAX];
```

```
main(int argc, char**argv){
  int i,j,res,temp_res, dx,llimit,ulimit,size,rank;
```

```
  MPI_Init(&argc, &argv);
  MPI_Comm_rank(MPI_COMM_WORLD, &rank);
  MPI_Comm_size(MPI_COMM_WORLD, &size);
  dx = YMAX/size;
  llimit = rank * dx;
  if(rank != (size - 1)) ulimit = llimit + dx;
  else ulimit = YMAX;
```

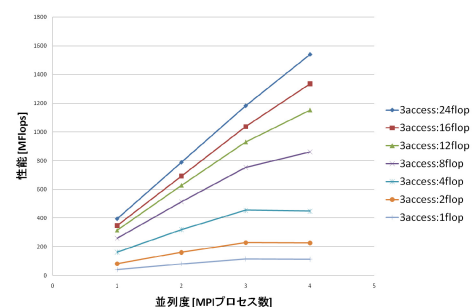
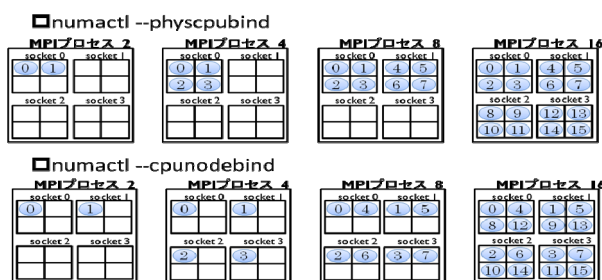
```
  temp_res = 0;
  for(i = llimit; i < ulimit; i++){
    for(j = 0; j < 10; j++){
      array[i][j] = func(i, j);
      temp_res += array[i][j];
    }
  }
```

```
  MPI_Allreduce(&temp_res, &res, 1, MPI_INT, MPI_SUM, MPI_COMM_WORLD);
  MPI_Finalize();
}
```

XMP プロトタイプで記述されたディレクティブベース言語(左)と MPI 関数を用いて変換された例(右)

(2) マルチコア・マルチソケット・マルチレール構成クラスタの性能解析(朴)

T2K-Tsukuba のような現在の主流 HPC クラスタは, マルチコア/マルチソケット構成によりノード内に十数程度までのコアを持つようになり, メモリ性能及びネットワーク性能を圧迫している. このようなシステムでは, マルチレール化によるノード間通信性能はある程度向上しているが, ノード内のメモリ性能がボトルネックとなる可能性が高く, ノード内コアの演算以外への有効利用手法を模索している. 基本的研究として, numactl によるスレッドとコアのマッピング及び byte/FLOP 値を人工的に設定した仮想ベンチマークでのメモリ性能が全体並列性能に与える影響を調べ, 通信処理の隠蔽にマルチコアを利用し, 必ずしも全コアを演算に割り当てない方法における性能解析を行った.

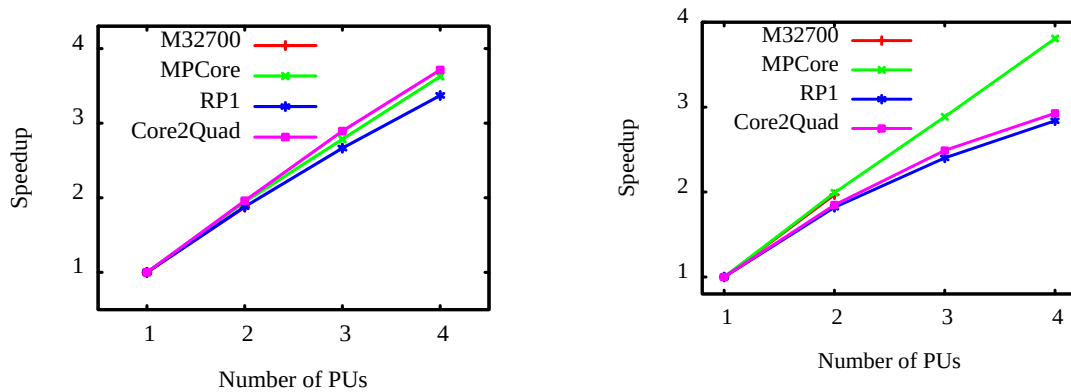


numactl による各種コアマッピング(左)と, コア当たりの byte/FLOP 値を変化させた場合の並列処理性能向上の saturation の様子(右)

(3) 組み込み向けマルチコアプロセッサの性能評価(佐藤, 塙)

近年の CPU のマルチコア化は組み込みプロセッサでも進んでおり, その性能評価と高性能システムへの適用は重要な課題である. 本研究ではルネサスの M32R, ARM/NEC エレクトロニクス社の MPCore, ルネサスの RP1, インテルの Core3 Duo 等の省電力マルチコアプロセッサを比較し, それらのスケーラビリティや並列適応性等を, 各種ベンチマークの本システムへの適用を行い, 性能を評価した. 結果として, 極端に memory

intensive なアプリケーションでなければ使用コア数と共に」性能も向上すること、ロック機構の違いが消費電力に以外に反映されることがわかった。

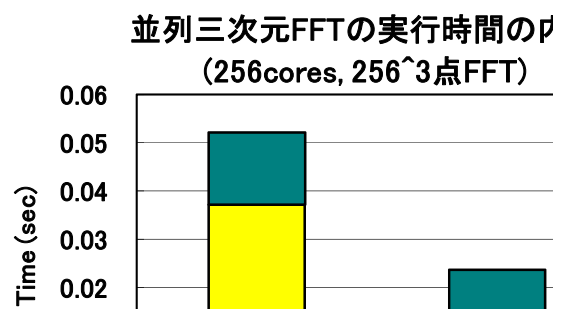
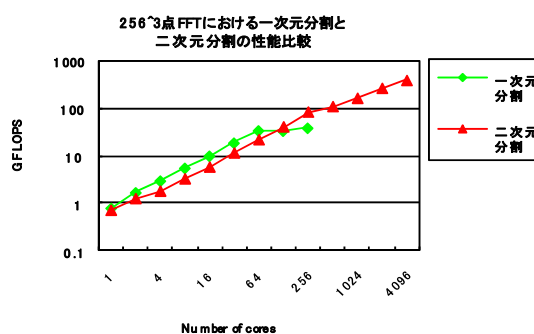


組み込み向けマルチコアプロセッサにおける Susan Smoothing(左)及び NAS-PB CG(右)のスケラビリティ

【3】大規模高性能並列システムにおける高速並列数値処理

(1) 次世代スーパーコンピュータに向けた高速フーリエ変換アルゴリズム(高橋)

文科省が推進している次世代スーパーコンピュータ(NGS)におけるナノ分野グランドチャレンジアプリケーションの一つである3D-RISMにおけるFFT性能の改善に関する研究を行った。同応用では、対象となるFFT問題の規模が小さく、strong scalingな問題となることがわかっている。本研究では並列三次元FFTを二次元分割化し、通信時間を削減することでMPIプロセスが多い場合にも対応するようにした。その結果、4096コアを用いるT2K-Tsukubaシステムにおいても、 $N=256^3$ 点FFTにおいて最大401GFLOPSを超える性能が得られた。

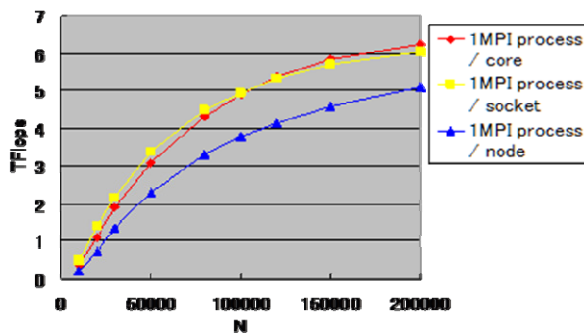


(2) T2K-Tsukuba における Linpack 性能チューニング(高橋)

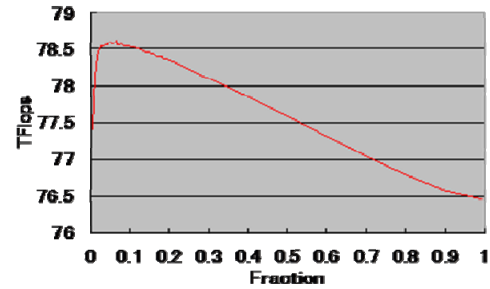
平成20年6月に導入されたT2K-Tsukubaシステムにおいて、625ノードを用いたLinpack性能評価を行った。大規模密行列の連立一次方程式の求解を通じ、fat-node構成とマルチレール構成を併せ持つT2K-Tsukubaのようなシステムの特徴を調べ、High Performance Linpackにおいて $N=1,508,000$ の場合に76.46TFLOPSの性能が得られた。これは、ピーク性能値の83.1%に当たり、IA-32ベースのシステムとしてほぼ

世界最高の実効効率を引き出している。

64ノード、1,024コアを用いた場合の
HPL実行方法の比較(NB=232)



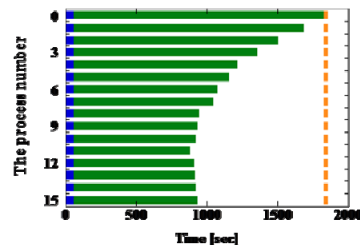
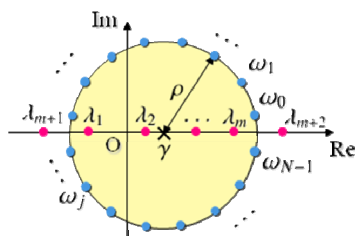
625ノード、10,000コアを用いたHPL実行にお
けるTFlops値の推移(N=1,508,000)



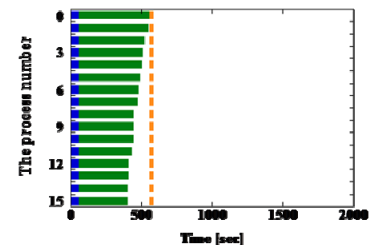
HPL のプロセスとコアのマッピングの違い(左)と, T2K-Tsukuba における実効性能(右)

(3) Block Krylov 部分空間法に関する研究(多田野)

Block Krylov 部分空間法の応用と高速化について, これを適用した周回積分を用いた固有値解法の高速度を行った. 一般化固有値問題 $Ax = \lambda Bx$ を互いに独立な連立一次方程式群 $(\omega_j B - A)Y_j = V, j = 0, 1, \dots, N-1$ に変換し, 小規模固有値問題に帰着させる. この求解において, 各固有値の値と ω の値に着目し, 円領域と接点の配置に注目し, 計算時間・通信時間を短縮させる. 結果として, 元の連立一次法定期の求会が大幅に高速化された.



(a) COCG法.



(b) Block COCG法.

■: データ転送, ■: 前処理, ■: 連立一次方程式の求解, ■: 固有対の計算.

円領域と接点の配置(左), T2K-Tsukuba におけるオリジナル処理(中), 改良後の時間短縮の様子(右)

【4】グリッド技術に関する研究

(1) Gfarm 広域ファイルシステムに関する研究(建部)

オープンソースベースで開発を継続している Gfarm 広域分散ファイルシステムに関する研究を継続している. 本年度は, 科研費「情報爆発」における研究開発用分散プラットフォームである InTrigger 上で全国 239 ノード, 147TByte に及ぶ実験環境を構築し, バンド幅測定等の実験を行った.

(2) JLDG におけるグループ内広域ファイル共有の実現(建部)

素粒子物理研究グループとの共同研究により, JLDG 内における分散ファイル共有環境を構築した. 研究は筑波大を中心とし, KEK, 金沢大, 京大, 阪大, 広島大により構築された SINET3 専用線で相互結合されたファイルシステム上で JLDG データベースの実運用を行った. 標準プロトコルである GridFTP による単一認証アクセスを実現したが, 拠点間ファイル複製作成は 50~70MB/ノードと落ち込んだ.



JLDG のための分散広域ファイルシステムと仮想組織管理サーバによる管理

(3) 国内大学間連携のためのグリッド環境構築(建部, 佐藤)

筑波大・東大・京大の3大学における T2K システム間の連携をはかるため, 筑波大 CCS 認証局の運用, 簡単な手順での証明書発行を実現し, 単一認証によるログイン, ファイル転送, 広域分散ファイルシステム運用を可能とした.

また, 9大学のスパコンセンター間を相互乗り入れする支援環境として, T2K-Tsukuba 上に「グリッドバック」の導入を行い, 相互のアカウント利用運用を開始した. さらに, T2K の3大学間連携のための NAREGI を始めとする各種環境の実装と運用を行った.

4. 研究業績

<学術雑誌論文>

1. 今田 貴之, 佐藤 三久, 堀田 義彦, 木村 英明: 分散型 Web サーバにおけるノード状態制御による省電力化の検討, 情報処理学会トランザクション ACS
2. 高橋 睦史, 佐藤 三久, 高橋 大介, 朴 泰祐, 宇川 彰, 中村 宏, 青木 秀貴, 澤本 英雄, 助川 直伸: 演算加速機構を持つオンチップメモリプロセッサの電力性能評価, 情報処理学会トランザクション ACS
3. Y. Nakajima, M. Sato, Y. Aida, T. Boku, F. Cappello: Integrating Computing Resources on Multiple Grid-Enabled Job Scheduling Systems Through a Grid RPC System. J. Grid Comput. 6(2): 141-157 (2008)
4. T. Sakurai, Y. Kodaki, H. Tadano, D. Takahashi, M. Sato, U. Nagashima: A parallel method for large sparse generalized eigenvalue problems using a GridRPC system. Future Generation Comp. Syst. 24(6): 613-619 (2008)
5. 横澤 拓弥, 高橋 大介, 朴 泰祐, 佐藤 三久, “行列積を用いた古典 Gram-Schmidt 直交化法の並列化”, 情報処理学会論文誌コンピューティングシステム, Vol.1, No.1, pp.61-72, 2008

6. 岡本 高幸, 三浦 信一, 朴 泰祐, 塙 敏博, 佐藤 三久, “ユーザ透過に利用可能な高性能・耐故障マルチリンク Ethernet 結合システム”, 情報処理学会論文誌コンピューティングシステム, Vol.1, No.1, pp.12-27, 2008
7. G. Antoniu, L. Cudennec, M. Ghareeb, O. Tatebe: “Building Hierarchical Grid Storage Using the GFARM Global File System and the JUXMEM Grid Data-Sharing Service”, Lecture Notes In Computer Science, Springer, Vol. 5168, pp.456-465, 2008

<国際会議論文>

1. H. Kimura, M. Sato, T. Imada and Y. Hotta: Runtime DVFS Control with instrumented code in Power-scalable Cluster System, Proc. 10th IEEE International Conference on Cluster Computing (CLUSTER 2008), pp.354-359
2. H. Midorikawa, M. Kurokawa, R. Himeno, M. Sato: DLM: A Distributed Large Memory System using Remote Memory Swapping over Cluster Nodes, Proc. 10th IEEE International Conference on Cluster Computing (CLUSTER 2008), pp.268-273
3. T. Imada, M. Sato, Y. Hotta, H. Kimura: Power management of distributed web savers by controlling server power state and traffic prediction for QoS. HP-PAC2008 (with IPDPS 2008): 1-8.
4. Y. Nakajima, Y. Aida, M. Sato, O. Tatebe: Performance Evaluation of Data Management Layer by Data Sharing Patterns for Grid RPC Applications. Euro-Par 2008: 554-564
5. J. Lee, M. Sato, T. Boku, “*OpenMPD: A Directive-Based Data Parallel Language Extensions for Distributed Memory Systems*”, First International Workshop on Parallel Programming Models and Systems Software for High-End Computing (P2S2), 2008.
6. R. Kanbayashi, M. Sato, A Distributed Architecture of Sensing Web forSharing Open Sensor Nodes, GPC 2009
7. S. miura, T. Okamoto, T. Boku, T. Hanawa, M. Sato, “RI2N: High-bandwidth and Fault-tolerant Network with Multi-link Ethernet for PC Clusters”, Proc. Of Cluster2008, CD-ROM, Tsukuba, 2008.
8. S. Miura, T. Boku, T. Okamoto, T. Hanawa, “A Dynamic Routing Control System for High-Performance PC Cluster with Multi-path Ethernet Connection”, Proc. of CAC2008, CD-ROM, Miami, 2008
9. S. Nishida, N. Katayama, I. Adachi, O. Tatebe, M. Sato, T. Boku, A. Ukawa: “High Performance Data Analysis for Particle Physics using the Gfarm file system”, Journal of Physics: Conference Series, Vol. 119, No. 062039, 2008.
10. D. Takahashi: A Volumetric 3-D FFT on Clusters of Multi-Core Processors, 2009 SIAM Conference on Computational Science and Engineering, Miami Hilton Downtown, Miami, Florida, USA, March 5, 2009
11. T. Yokozawa, D. Takahashi, T. Boku, M. Sato: Parallel Implementation of a Recursive Blocked Algorithm for Classical Gram-Schmidt Orthogonalization, Workshop on State-of-the-Art in Scientific and Parallel Computing (PARA 2008), NTNU, Trondheim, Norway, May 15, 2008.

12. Y. Ishikawa, H. Fujita, T. Maeda, M. Matsuda, M. Sugaya, M. Sato, T. Hanawa, S. Miura, T. Boku, Y. Kinebuchi, L. Sun, T. Nakajima, J. Nakazawa, and H. Tokuda, “Towards an Open Dependable Operating System,” Proc. of 12th IEEE International Symposium on Object / component / service-oriented Real-time distributed Computing (ISORC2009)
13. H. Tadano, T. Sakurai. A performance evaluation of block Krylov subspace methods for the contour integral method. International Symposium on Frontiers of Computational Science 2008, Nagoya, Japan, Nov. 2008
14. T. Sakurai, H. Tadano, An Inner/Outer Loop Free Parallel Method for Interior Eigenvalue Problems. 7th International Workshop on Accurate Solution of Eigenvalue Problems, Dubrovnik, Croatia, June. 2008.
15. J. Asakura, T. Sakurai, H. Tadano, T. Ikegami and K. Kimura. A Linearization Method for Polynomial Eigenvalue Problems Using a Contour Integral. 7th International Workshop on Accurate Solution of Eigenvalue Problems, Dubrovnik, Croatia, June. 2008.

<招待講演>

1. T. Boku, “Supercomputing Driven by Cluster Technology -It’s Light and Shadow-”, Supercomputing & KREONET/GLORIAD Symposium, Jeju, 2008 (Keynote Speech)
2. 朴 泰祐, “高性能計算環境の動向と今後の展望”, 大阪大学シグマ講演会(招待講演)
3. 建部修見: “クラウドコンピューティングを高度化するストレージ技術(招待講演)”, WebDB Forum 2008, 招待講演, 2008.
4. 高橋大介: 高速フーリエ変換の超並列計算に向けて, スーパーコンピューターワークショップ 2009, 自然科学研究機構岡崎コンファレンスセンター, 岡崎市, 2009 年 1 月 20 日