Vol.2010-ARC-192 No.14
Vol.2010-HPC-128 No.14
2010/12/17

クラウドコンピューティングの性能評価

池上 努[†1] 高野了成[†1] 田中良夫[†1]
中田秀基[†1] 関口智嗣[†1]

クラウド資源を活用した高性能計算 (HPC) の可能性を探るため、仮想化が計算機性能に及ぼす影響を、ディスク IO、 MPI 通信性能、計算性能のそれぞれについて調べた。ディスク IO 性能は準仮想化環境で実ハードウェアと遜色ない性能を示した。計算性能に対する仮想化のオーバヘッドはアプリケーションに依存するものの、5 ∼ 15 % 程度に収まった。一方、計算ノードを 16 個用意して MPI 通信性能を計測すると、最善でも実ハードウェアの半分程度に止まった。現状の仮想化環境における MPI 性能は細粒度並列のアプリケーションでは実用的とは言えず、今後の改善が期待される。

Performance evaluation of the cloud computing

Tsutomu Ikegami ,[†1] Ryousei Takano ,[†1]
Yoshio Tanaka ,[†1] Hidemoto Nakada [†1]
and Satoshi Sekiguchi [†1]

The feasibility of the cloud computing in the field of high performance computing was assessed by surveying the performance of a virtual machine. Namely, the penalties of the virtualization on the disk I/O, MPI communication, and computational performances were measured by comparing with those on the real machine. The disk I/O performance of a para-virtual machine was found to be comparable to the real one. The overhead of the virtualization on the computational performance was as low as 5 ∼ 15 %, depending on the type of applications. On the other hand, the MPI communication throughput measured at 16 nodes was a half of the real machines, at best. The MPI communication in the present virtual environment is not powerful enough for fine-grained parallel applications, where further developments are wanted.

1. はじめに

クラウドコンピューティングは、計算資源を抽象化して運用する手段として近年その利用が拡大している。中でも、計算機のコモディティ化の急速な進展を背景に、計算機そのものを仮想化する Infrastructure as a Service (IaaS) が実用的なサービスとして提供されるようになった。IaaS は既存の物理サーバをそのまま仮想化できる点で柔軟性が高く、また複数のサービスを一台のハードウェアに集約する際のセキュリティにも優れている。仮想化にはオーバヘッドを伴うが、CPU のマルチコア化が進んで性能が向上した結果、その負担は相対的に小さくなっている。また CPU 命令セットや周辺デバイスなどハードウェア面でも仮想化への対応が進められており、効率は改善される傾向にある。

仮想化技術は、サーバの集約化など計算機資源の有効利用や、実ハードウェアの故障に対する耐障害性の観点から利用されることが多い。一方、クラウドコンピューティングに対する高性能計算 (High Performance Computing, HPC) の高い需要を受け、HPC を意識した IaaS も現れてきた。2010 年 7 月から提供が開始された、Amazon Elastic Compute Cloud (EC2) の Cluster Instance がそれである[1)]。Cluster Instance は 8 個の 64-bit コアと 23 GB のメモリを搭載した仮想計算機で、恐らく一台の物理サーバをそのまま仮想化した構成を取っている。仮想化のオーバヘッドを考えると、HPC アプリケーションを実行するのに仮想計算機を用いるのは得策とは言えず、生のハードウェアをそのまま使用する方が効率が良い。しかし HPC アプリケーションの種類は多岐に亘り、これら全てをカバーする Software as a Service (SaaS) や Platform as a Service (PaaS) を設計するのは容易ではない。このため、HPC 向けの汎用クラウド資源として、資源提供側は IaaS を供給するのが自然である。

一方、HPC 向け計算資源の仮想化はユーザ側にはメリットがないかというと、そうとも限らない。パラメータサーベイのように、大量の独立なタスクを分散処理する場合、複数のサイトにまたがって計算資源を確保するケースがある。このとき、従来型のアプローチでは利用するサイトごとに HPC アプリケーションを整備していく。しかし HPC アプリケーションは往々にして実験的なコードを含んでおり、親切なインストーラがいつも利用できるとは限らない。この場合、計算機環境の調査と設定・バイナリの構築・テストデータを用い

†1 産業技術総合研究所
AIST

Vol.2010-ARC-192 No.14
Vol.2010-HPC-128 No.14
2010/12/17

たバイナリの検証といった一連の作業を、利用するサイト毎に繰り返していくことになる。これは特に巨大アプリケーションでは大変な作業で、利用サイトの追加を阻む大きな要因となっている[2),3)]。ここでもし各サイトで共通の仮想計算機を設定し、計算機環境の差異を仮想計算機のレベルで吸収してしまえば、アプリケーションの整備は一回で済む。この省力化が、ユーザ側における仮想化のメリットである。

HPC アプリケーションでは、CPU 処理能力だけでなくディスク IO や通信などの実効性能が重要な要素となる。本稿ではクラウドコンピューティングに用いられる仮想化技術に焦点を置き、仮想化がこうした基本性能にどのような影響を与えるか調べる。特にディスク IO 性能と MPI 通信性能を計測し、実ハードウェア環境と比較する。また実際の HPC アプリケーションを用いたベンチマークを実施し、現時点におけるクラウドコンピューティングの総合性能を評価すると共に、今後解決すべき技術的課題を明らかにする。

2. 計算機環境

実験には AIST Green Cloud (AGC) の 16 ノードを使用した。AGC の各ノードは、Quad Nehalem (E5540 2.53 GHz) 2 基、メモリ 48 GB、ハードディスク 2 台 (SAS 300 GB, Hardware RAID1) から構成され、ノード間は non-Blocking の 10 Gb Ethernet (Broadcom NetXtreme II 5771x) および InfiniBand (Mellanox Technologies MT26428) で接続される。OS は実ハードウェア (Bare Metal Machine, BMM)、仮想環境 (Virtual Machine, VM) 共に 64 bit 版の Linux CentOS 5.5 を使用した。ただし、OS 標準の 10Gb Ethernet ドライバは MPI 通信性能の面で問題があることがわかったため、これを最新版のドライバ (bnx2x 1.52.15) で置き換えた。HyperThreading は使用しなかった。

仮想環境は Xen 3.4.3 を用いて構築し、準仮想化 (Para-VM, PVM) と完全仮想化 (Hardware VM, HVM) の両方について計測した。HVM のネットワークインタフェースは、ネットワークインタフェースカード (NIC) を忠実にエミュレートするドライバではなく、PVM 同様これを迂回する準仮想化ドライバ (Para-Virtual Driver, PVD) を利用した。一方、Gb Ethernet NIC (e1000) をエミュレートする設定 (HVM-emu) も用意し、一部の計測を実施した。Domain-0, Domain-U 共に 8 個の CPU を割り当て、どちらも物理 CPU コアと仮想 CPU コアを一対一で対応付けるよう設定した。また Domain-0 には 1 GB のメモリを排他的に割り当てた。

計測に用いた計算機環境を **表 1** にまとめる。 VM は Domain-0 上の仮想ブリッジを介してネットワークに接続される。この仮想ブリッジとスイッチ、および BMM/Domain-0 のネットワークインタフェースの MTU 値は全て 9000 に揃えた。仮想計算機側は、PVM では MTU 値を 9000 に設定できたが、HVM, HVM-emu では大きな値を設定すると通信できなくなるため、1500 のままとした。

表 1 実験で用いた計算機環境
Table 1 Real/virtual computational environments employed in the experiment

	BMM	BMM-IB	PVM	HVM	HVM-emu	EC2
N_CPU	8	8	8	8	8	8
Memory (GB)	48	48	46	46	46	23
Network Interface	10GEth	IB	PVD	PVD	e1000	PVD
MTU	9000	–	9000	1500	1500	9000

比較のため、Amazon EC2 Cluster Instance 上でも実験を実施した。Cluster Instance には Quad Nehalem (X5570 2.93 GHz) 2 基とメモリ 23 GB が割当てられており、起動ディスクは Elastic Block Store (EBS) 上に置かれる。起動ディスクの他に 845 GB の一時ディスクが 2 基 (/dev/sdb, /dev/sdc) 用意される。HyperThreading は有効になっており、見掛け上 16 コア存在する。しかし予備的な計算の結果、ノードあたり 16 スレッド起動しても性能が上がらない (むしろ若干、劣化する) ことから、実験はノードあたりの最大スレッド数を 8 に制限して実施した。OS は Amazon 側で用意されている CentOS 5.4 をそのまま用いた。複数の Instance を同一の placement group 内に立ち上げると、相互に non-blocking の 10Gb Ethernet で接続される。仮想環境は完全仮想化で、ネットワークインタフェースには PVD が用いられているが、MTU は 9000 になっている。CPU は AGC より高性能で、またネットワーク回りもチューンされていると考えられることから、EC2 は AGC の HVM 環境よりも性能的に優れていると予想される。

3. ベンチマークソフト

ハードディスク IO 性能は iozone 3[4)] を用いて測定した。オプションは `-a -g 1g -e` を用いた。すなわち、レコードサイズを変えながら最大 1 GB のファイルについて読み書きのスループットを調べ、またその際、書き込み時間は flush まで含めて計測する。

MPI 通信性能は Intel MPI Benchmarks 3.2[5)] を用いて測定した。ノードあたりのメモリ搭載量が肥大化していることから、特にメッセージサイズの大きな所まで (∼ 1 GB) 調べた。ノード間の通信性能を測定するため、ベンチマークでは 16 ノードを用い、各ノードに 1 ランクずつ起動した。

Vol.2010-ARC-192 No.14
Vol.2010-HPC-128 No.14
2010/12/17

HPC アプリケーションでのベンチマーク測定には、NAS Parallel Benchmarks 3.3.1 (NPB)[6] および自作のアプリケーション Bloss を用いた。NPB は MPI と OpenMP のハイブリッド並列性能を測定する Multi-Zone 版を選択した。Multi-Zone 版には LU, SP, BT の 3 種類が含まれるが、LU はランク数の上限が 16 に制限されることから省いた。問題サイズは クラス C を選択した。クラス C は全体で 800 MB 程度のメモリを使用する。予備的な実験の結果より、MPI 1 ランクあたり 2 本の OpenMP スレッドを割当てた。したがって 16 ノードで最大 64 ランクとなる。

Bloss はブロック櫻井・杉浦法を用いた疎行列非線形固有値問題の内部固有値解法アプリケーションである[7)–9)]。プログラムは ∼ 10 GB のメモリを必要とする OpenMP 並列ジョブを MPI で束ねる構成をとっており、MPI 部分は比較的粗粒度の並列となっている。MPI 通信パタンが単純であること (∼ 1 GB の集団通信が主体)、また OpenMP 部分で大規模なメモリアクセスが発生することが特徴である。Bloss の実行では、主にメモリ要求量の観点から MPI 1 ランクあたり 4 本の OpenMP スレッドを割当てた。したがって 16 ノードで最大 32 ランクとなる。

コンパイラは gcc/gfortran 4.1.2 を用い、最適化オプションは`-O3 -fopenmp` を指定した。Bloss ではさらに Intel Math Kernel Library 11.1 を用いた。MPI は OpenMPI 1.4 を用い、通信路は 10Gb Ethernet を指定した。参考のため、BMM 上では通信路に InfiniBand を用いた計測も実施した。この時、MPI 集団通信の性能を向上させるためには、実行時オプションとして`--mca mpi_leave_pinned 0` を付加する必要があった。バイナリは BMM 上で一度だけ生成し、これを全環境で流用した。

4. 結果と考察

4.1 ハードディスク IO 性能

ハードディスクの write/read スループットをファイルサイズの関数として**図 1** にプロットした。測定は起動ディスクに対して行い、簡単のためレコードサイズ 64 kB のデータのみ示した。 Write 性能は flush の時間を含めてスループットを計算しているので、OS 側のディスクキャッシュの影響はない。AGC の RAID カードには 512 MB のキャッシュが搭載されており、BMM の write 性能はこれを反映している。PVM は BMM とほぼ同じ挙動を示し、仮想化の影響はほとんど見られない。一方、HVM の write 性能は BMM の半分程度で、同じ完全仮想化マシンである EC2 も同様の傾向を示す。なお、一般の SATA ディスクについて同様の測定を行うと、最大でも 90 MB/s となった。HVM の性能はこれを上

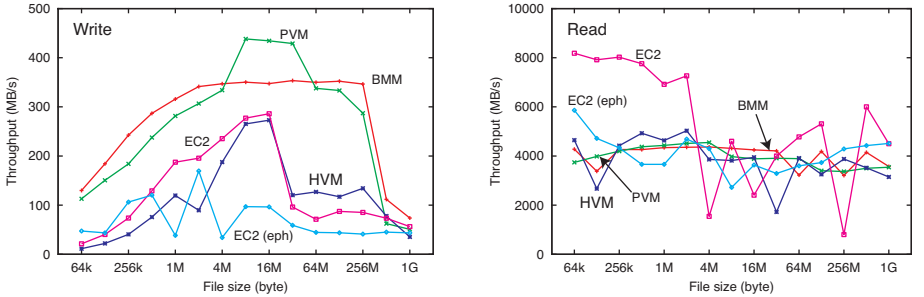

図 1 ハードディスクの write/read スループット
Fig. 1 Write/read throughput of hard disk.

回っており、一般的な用途ではまず問題ないと思われる。EC2 の一時ディスク (ephemeral disk, EC2-eph) の write 性能は起動ディスクのさらに半分程度であった。一時ディスクは大容量であるが、ファイルサイズの大きな所での書き込み性能は SATA ディスクの半分程度であり、その用途には注意が必要となる。

Read 性能は OS 側のディスクキャッシュの影響を排除するのが難しく、測定結果はディスクキャッシュ性能を反映したものとなっている。この場合、BMM と仮想環境で差はほとんど見られない。初回読み出し時の性能を反映したものではないことに、注意されたい。

4.2 MPI 通信性能

MPI 通信性能の測定では、メッセージサイズを最大 1 GB まで変えながら PingPong, Bcast, Reduce, Allreduce のそれぞれについて所要時間を計測し、スループットを算出した。結果を **図 2** に示す。 PingPong は 16 ノード中の 2 ノード間の通信、残りの集団通信は 16 ノード間の通信となる。BMM と比較して、PVM の通信性能はだいたい 1/3 程度である。HVM はネットワークインタフェースに準仮想化ドライバを使用しているが、その性能は PVM のさらに半分程度である。NIC までエミュレートする HVM-emu の性能はさらに低劣で、特に集団通信の性能は顕著に下がるため、測定はメッセージサイズ 16 MB で打ち切った。一方、EC2 は HVM 相当の完全仮想化マシンだが、その通信性能は PVM を凌駕しており、丁寧なチューニングが施されていることがうかがえる。

メッセージサイズ 1 GB 時のスループットを、BMM-IB の結果と併せて**表 2** にまとめる。 MPI 通信における仮想化のオーバヘッドは大きく、現状では非仮想化環境の半分以上の性能を出すことは難しい。InfiniBand のような専用ハードウェア使用時と比較すると、そ

Vol.2010-ARC-192 No.14
Vol.2010-HPC-128 No.14
2010/12/17

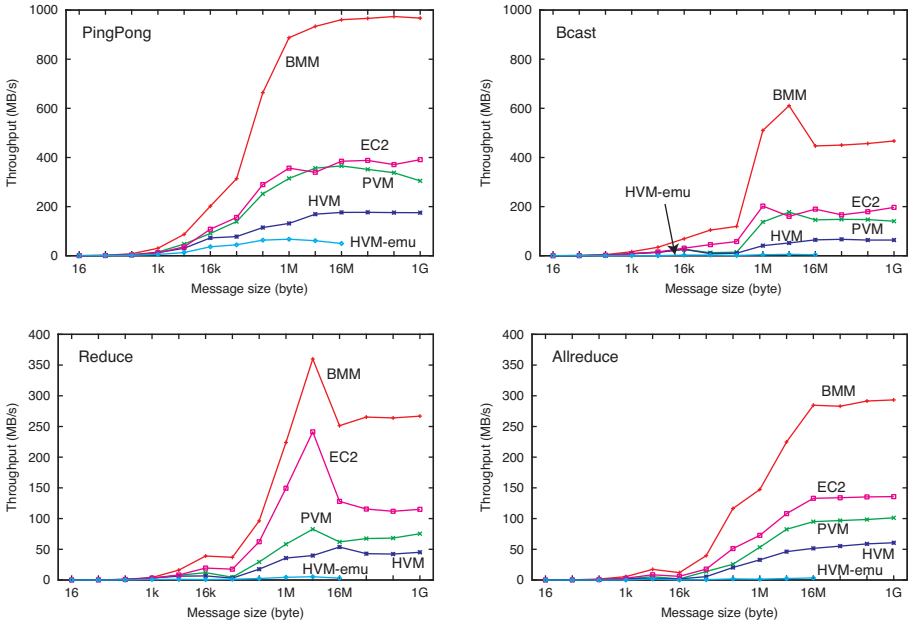

図 2 MPI 通信性能
Fig. 2 MPI data communication throughput.

表 2 1 GB 通信時の MPI スループット (MB/s)
Table 2 MPI communication throughput at 1 GB message size (MB/s)

VM type	PingPong	Bcast	Reduce	Allreduce
BMM	967.3	467.28	266.80	293.29
PVM	304.9	140.68	75.47	101.19
HVM	175.4	64.02	45.12	60.69
EC2	391.2	196.78	115.03	135.72
BMM-IB	2233.8	913.94	450.30	540.55

表 3 計算ノード単体の性能 (計算時間)
Table 3 A single node performance (wall clock time)

VM type	SP-MZ (sec)	BT-MZ (sec)	Bloss (min)
BMM	86.63	131.83	21.06
PVM	100.44	137.90	22.33
HVM	101.08	141.12	22.66
EC2	88.00	126.01	20.00

の力不足は一層顕著である。後で見るように、細粒度の通信を伴う並列アプリケーションでは、これは特に不利に働く。現在、PCI passthrough を使って高い通信性能を実現できないか、検討中である。

4.3 HPC ベンチマーク: 計算ノード単体の性能

まず始めに、仮想化がノード単体の性能に及ぼす影響について調べる。ノード上の CPU コアを全て使用するため、NPB では 4 ランク、Bloss では 2 ランクを立ち上げ、実行時間を測定した。結果を **表 3** にまとめる。 PVM と HVM はほぼ同じ性能を示し、BMM との比較から仮想化のオーバヘッドはアプリケーションに依存して 5 ～ 15 % 程度あることがわかる。事前に実施した小規模なテストでは性能差はほとんど見られず、オーバヘッドは主にメモリ管理回りに由来すると思われる。EC2 は AGC と比べて 基本 CPU 性能が良いことを考えあわせると、EC2 でも同様のオーバヘッドを生じていると思われる。10 % 程度のオーバヘッドは、クラウド上の豊富な計算資源を活用できること、またデプロイメントの手間の軽減を考えると、十分許容範囲と言える。特にノード間通信を必要としない分散アプリケーションでは、仮想化ベースのクラウドコンピューティングが現状で既に実用レベルと考えられる。

4.4 HPC ベンチマーク: 並列性能

NPB ではランク数を 1 から 64 まで増やしながら並列化効率を測定した。結果を **図 3** に示す。 並列化効率は 1 ランク実行時の計算時間に対して算出した。1 ランクあたり 2 CPU コアを割り当てているので、4 ランクまではノード内並列、8 ランク以降が複数ノード間の並列化となる。SP は静的負荷分散型の並列アプリケーションで、並列化効率は通信性能を素直に反映したものとなっている。一方 BT では、PVM, HVM は SP と同様に振る舞うが、BMM と EC2 は不可解な挙動を示し、通信性能の改善が効率向上につながっていない。BT は動的負荷分散を用いており、その計算時間は MPI-OpenMP 間の分割にも影響を受ける。詳細な理由は不明だが、BMM, EC2 では負荷分散に失敗しているものと思われ

Vol.2010-ARC-192 No.14
Vol.2010-HPC-128 No.14
2010/12/17

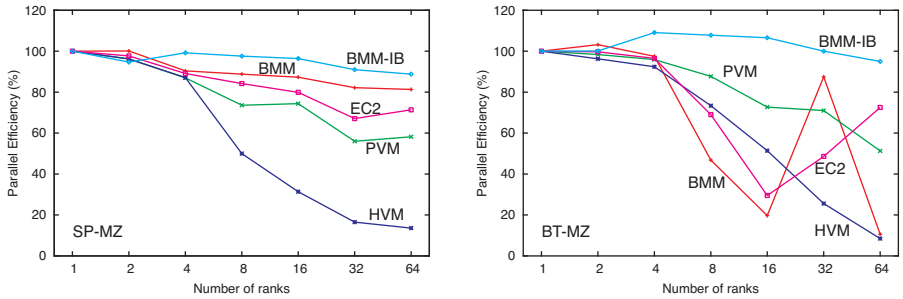

図 3 NPB の並列化効率
Fig. 3 Parallel efficiencies of NPB.

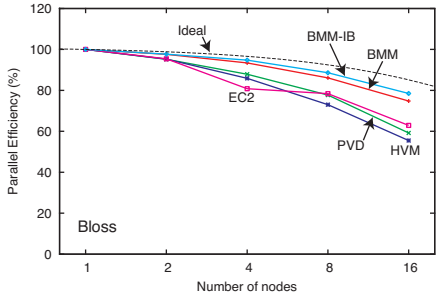

図 4 Bloss の並列化効率
Fig. 4 parallel efficiencies of Bloss.

る。通信路に InfiniBand を用いる BMM-IB はいずれのケースでも卓越した性能を示しており、仮想環境から専用の通信インフラにアクセスする手段の拡充が望まれる。

Bloss ではノード数を 1 から 16 まで増やしながら並列化効率を測定した。結果を **図 4** に示す。 並列化効率は 1 ノード (2 ランク) 時を基準に算出した。Bloss には各ランクで処理が重複する箇所があり、本質的に並列化効率の低下が避けられない。このアルゴリズム由来の並列化効率 (ideal) をグラフ中にあわせて示した。この曲線との差分が、通信由来の並列化効率の低下分となる。Bloss における MPI 通信はメッセージサイズの大きな集団通信が主体であり、並列効率は通信性能を素直に反映している。一方、NPB の SP と異なり、Bloss の MPI 通信は比較的、粗粒度であることから、通信性能の低い HVM でも並列化効率が極端に悪くなることはない。通信量自体は決して小さくないものの、粗粒度を前提としたアルゴリズムは十分、仮想環境での実行に耐えると考えられる。

5. 最 後 に

本稿では、仮想化が計算機性能に及ぼす影響を、ディスク IO、 MPI 通信性能、計算性能のそれぞれについて調べた。ディスク IO 性能について準仮想化マシンは実ハードウェアとほぼ同等の性能を示し、クリティカルな用途にも十分耐えると考えられる。一方、完全仮想化環境では実ハードウェアの半分程度の性能となり、アプリケーション次第ではディスク IO についても準仮想化ドライバの導入が有効と考えられる。仮想環境における MPI 通信性能は最善でも実ハードウェアの半分程度で、今後の改善が強く望まれる分野である。InfiniBand など MPI 通信専用ハードウェアの優位性は大きく、仮想環境からこれら通信インフラに直接アクセス可能になれば、状況は大きく改善されると思われる。計算性能はアプリケーションの性質に左右されるものの、大体 5 ～ 15 % のオーバヘッドと見込まれる。クラウド環境では利用計算資源の増設が容易であり、またより高性能な計算資源の選択も可能なことから、この程度のオーバヘッドであれば十分吸収可能と思われる。HPC 分野におけるクラウドコンピューティングは、現状ではアプリケーション次第と言わざるを得ない。MPI 通信性能が弱いので、細粒度な並列アプリケーションは不利である。一方、粗粒度並列を前提に設計されたアプリケーションは、現状でもそう困難なく性能を発揮できると考えられる。

謝辞 仮想化技術全般について産総研の広淵氏にご助言いただいた。また実験環境の整備では (株) 創夢の大田氏にご尽力いただいた。謹んで感謝の意を表する。

参 考 文 献

1) Amazon Web Services Blog: New Amazon EC2 Instance Type - The Cluster Compute Instance (2010). `http://aws.typepad.com/aws/2010/07/the-new-amazon-ec2-instance-type-the-cluster-compute-instance.html`.
2) Takemiya, H., Tanaka, Y., Nakada, H., Sekiguchi, S., Ogata, S., Kalia, R. K., Nakano, A. and Vashishta, P.: Sustainable Adaptive Grid Supercomputing: Multi-scale Simulation of Semiconductor Processing across the Pacific, *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing* (2006).
3) Ikegami, T., Maki, J., Takami, T., Tanaka, Y., Yokokawa, M., Sekiguchi, S. and Aoyagi, M.: GridFMO – Quantum Chemistry of Proteins on the Grid, *Proceedings of the 8th IEEE/ACM International Conference on Grid Computing (Grid 2007)*,

Vol.2010-ARC-192 No.14
Vol.2010-HPC-128 No.14
2010/12/17

pp.153–160 (2007).
4) : IOZone. http://www.iozone.org.
5) : Intel MPI Benchmarks. www.intel.com/software/imb.
6) : NAS Parallel Benchmarks. http://www.nas.nasa.gov/Resources/Software/npb.html.
7) Asakura, J., Sakurai, T., Tadano, H., Ikegami, T. and Kimura, K.: A numerical method for polynomial eigenvalue problems using contour integrals, *JSIAM Lett.*, Vol.1, pp.52–55 (2009).
8) Ikegami, T., Sakurai, T. and Nagashima, U.: A filter diagonalization for generalized eigenvalue problems based on the Sakurai-Sugiura projection method, *J. Comp.Appl.Math.*, Vol.233, pp.1927–1936 (2010).
9) Sakurai, T. and Sugiura, H.: A projection method for generalized eigenvalue problems using numerical integration, *J. Comp.Appl.Math.*, Vol.159, pp.119–128 (2003).