Vol.2009-IOT-7 No.2
2009/10/9

教育用PCシステムを用いた大規模分散計算フレームワークの実現に向けて

奥村 勝[†1]

オープンソースの大規模分散処理ソフトウェアである Hadoop と，既存の教育用 PC システムの組合わせにより，大規模分散処理環境を構築し動作させる検証を行った．この結果，ネットブート型 PC システムと Hadoop を組み合わせて，1,000 台の PC からなる Hadoop クラスタを構成し，大規模分散処理環境を構築できることを確認した．

Towards Large Scale Distributed Computing Framework with Educational PC System

Masaru Okumura[†1]

In this paper, I report the Hadoop cluster that consisted of 1000 PC was constructed with netboot PC. I verified the large scale distributed processing environment that combined an existing educational PC system with Hadoop that was the open-source software for large scale distributed processing.

1. はじめに

Google に代表されるような日々発生する大規模なデータを高速に処理することで提供される情報サービスが実用化し，日常の生活に欠かせないものとなっている．これらのサービスは，数千台から数万台という大規模な計算資源を組み合わせ，分散処理することで実現しているが従来の並列処理や分散処理とは異なるアプローチが採られている．従来の PC クラスタやグリッド技術を用いた並列処理，分散処理は主に計算処理を利用の主目的としていたが，これらの仕組みを利用するには並列処理や分散処理に特有の問題を理解しつつプログラミングを行う必要があったり，台数に応じた効果を得るには様々な問題分析や最適化を必要としてきた．これに対し，Google に代表されるようなクラウド・コンピューティングでは，プログラミングモデルを制約し，単純化することで並列，分散処理に特有の問題を除去し，処理を行うための計算機数をスケールさせることで性能の向上を図る仕組みを用いている．

一方，大学のような教育機関においては，情報処理教育等の実施のための PC 環境として教育用 PC システムが多く導入されている．その PC 台数は，大学の規模などにもよるが，100 台程度から多い場合は数千台にも及ぶ．しかしながら，教育機関という組織の運用上，夜間や夏季，冬季などの長期休暇期間は，それらの PC の多くは遊休状態であることが多い．また，導入台数が数百台以上と多い場合，メンテナンス性の点から，PC の構成を容易に変更できるネットブート型のシンクライアント方式等の仕組みを採用していることが多い．これは，場合よっては通常の講義利用以外への PC システムの転用が容易に行える機能を備えているともいえる．

本研究では，大規模データ処理を実現するための仕組みである大規模分散計算フレームワークを既存の教育用 PC システムのリソースを活用して実現し，大学が保有する PC リソースの有効化を図ると供に，今後ニーズが増加すると予想される大規模データ処理環境を学内の利用者に提供する仕組みを検討している．これに先立ち本稿では，オープンソースの大規模分散処理ソフトウェアである Hadoop と既存の教育用 PC システムの組み合わせにより大規模分散処理環境を構築し，動作させた検証実験について報告する．この結果，ネットブート型 PC システムと Hadoop を組み合わせ，1,000 台の PC からなる Hadoop クラスタを構成して動作させ，100GB 分の数値データの並べ替えを約 5 分で処理できることを確認した．

2. Hadoop

2.1 Hadoop の概要

Hadoop[2] とは，Google の基盤技術を基に開発されたオープンソースの大規模分散データ処理用ソフトウェアである．Google の基盤技術[1] である大規模分散ファイルシステム Google File System，大規模分散計算フレームワーク MapReduce，大規模分散データベース BigTable に対応する機能が，Hadoop ではそれぞれ HDFS(Hadoop Distributed File System)，HadoopMapReduce，hBase というソフトウェアにより提供されている．

†1 福岡大学 総合情報処理センター
Information Technology Center, Fukuoka University

Hadoop の目的は，一般的なハードウェアやソフトウェアを備えた計算機を多数用いて分散処理環境を構築し，大規模データ処理のための分散計算フレームワークを提供することである．これまでの MPI を用いた PC クラスタなどの並列分散処理環境とは異なり，Google やそれに基づく Hadoop では MapReduce というプログラミングモデルを用いることで，性能やプログラミングの自由度は犠牲になるものの，利用者が分散処理に特有の様々な問題を意識することなくデータ処理を行える仕組みを提供している．また，スケールアウトにより処理能力を拡大することを基本方針としており，高性能な計算機でなくても，処理に活用できる計算機数を単純に増加させることで処理能力の向上を図ることができる仕組みになっている．Hadoop の利用事例としては，4,000 台のサーバにより，32,000CPU コア，16PB のディスクを備えた Hadoop クラスタの運用を行っている Yahoo!の事例などがある[3]．

2.2 Hadoop のシステム構成

Hadoop は大きく分けて，図 1 に示すようにデータを分散管理する分散ファイルシステムである Hadoop Distributed File System（以下，HDFS) と，HDFS 上で分散処理を行う HadoopMapReduce（以下，MapReduce）から構成される．さらに，HDFS は NameNode と DataNode，MapReduce は JobTracker と TaskTracker というサーバから構成される．これらはマスタ・スレーブ構成となっており，NameNode と JobTracker がマスタとして機能し，それぞれ HDFS のメタデータの管理と MapRedcue ジョブの実行管理を行う．DataNode と TaskTracker はスレーブとして機能し，それぞれ HDFS の分散データの保持と MapReduce ジョブの計算処理を行う．

一般的には計算機を多数用意し，DFS や MapReduce のスレーブノードとして Hadoop クラスタを構成するが，Hadoop は Java で実装されているため，Java を実行可能なプラットフォームであれば異機種混在であってもスレーブとして組み入れることが可能である．Hadoop クラスタをシンプルに構成するならば，スレーブを同一のハードウェアや OS からなる機種で統一して構成する方式となる．この場合，スレーブは IP アドレスだけが異なれば良く，ソフトウェア構成や設定に関してはまったくの同一構成（クローン）であっても構わない．

Hadoop クラスタの利用は，図 1 に示すようにクライアント端末から Hadoop クラスタに対して行い，次のような流れで処理を行う．

(1) クライアントから HDFS への処理対象データの書き込み
(2) MapReduce ジョブのを JobTracker へ投入
(3) JobTracker は各スレーブの TaskTracker へジョブを分散配置し，各 TaskTracker は HDFS 上のデータへアクセスして処理を実行
(4) 処理結果を HDFS へ書き込む
(5) ジョブ終了後，HDFS からクライアントへ結果を読み出す

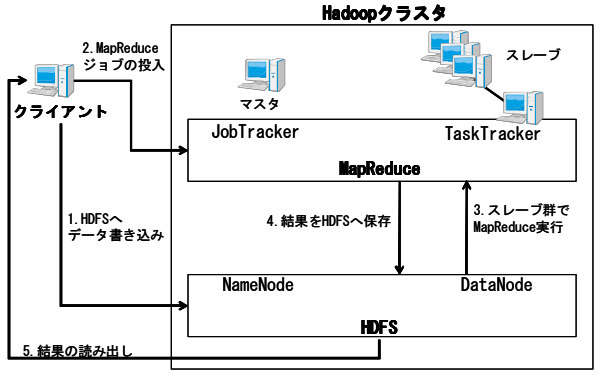

図 1 Hadoop のシステム構成

3. 福岡大学における教育用 PC システム

福岡大学総合情報処理センターが提供する教育用 PC システムは，教育研究システム FUTURE3[4] のうち，情報処理教育などの PC を用いた各種講義，演習を行うための教育設備であり，2009 年 8 月現在 19PC 教室，約 1,336 台の PC から構成されている．本学の教育用 PC システムの特徴としては，ネットブート型 PC クライアントを採用し，WindowsXP と Linux のデュアルブート構成とすることで，幅広い教育ニーズへの対応を図ると同時に，1,000 台を越える PC 群のメンテナンスコストを抑えた運用を実現している点である[5]．

各 PC 教室のシステム構成を図 2 に示す．VID(Virtual Image Distributor) と呼ばれるネットブートシステムを採用しており，PC50 台に 1 台の割合で配置される IO サーバ上にクライアント PC 用の起動イメージ (Virtual Image) を配している．各 PC は起動時に，IO サーバ上の起動イメージをネットワークを介してアクセスし，イメージを読み込むことで OS やアプリケーションを起動する．なお，現 FUTURE3 の教育用 PC システムでは HDD を搭載しないディスクレス PC 構成となっている．また，ネットブート型であるため，通常の講義運用時は講義用のソフトをセットアップした Windows の起動イメージで各 PC を起動するが，必要に応じて IO サーバ上の起動イメージを修正したり，異なる設定の OS 起

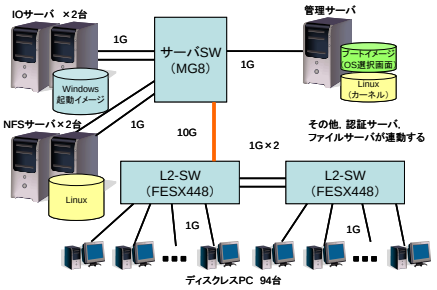

図 2　PC 教室のシステム構成

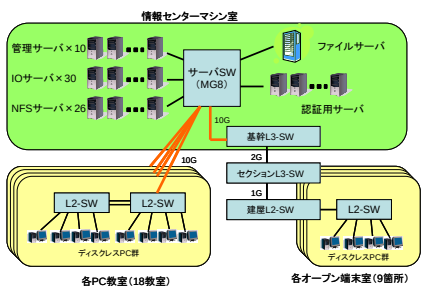

図 3　教育用 PC システム全体構成

動イメージと差し替えることで，全 PC の OS の設定やアプリケーションを容易に変更することができる．

各 PC 教室の基本構成は，図 2 で示した構成となっており，19 教室からなる教育用 PC システム全体では，この構成を図 3 のように水平展開し，管理サーバ 10 台，IO サーバ 30 台，NFS サーバ 26 台，その他認証用サーバなどを用いて，約 1,300 台の教育用 PC システムを構成している．また，管理の面からこれら全てサーバは，総合情報処理センターのマシン室に集中配置し，各 PC 教室にはディスクレス PC とネットワーク機器のみを配する構成としている．

4. ネットブート型 PC クライアントによる Hadoop クラスタの検証

本稿では，既存の教育用 PC システムと Hadoop を活用することで，教育用 PC システムを活用した大規模計算フレームワークの実現性や課題の発見を行うことを目的に，ネットブート型 PC クライアントによる Hadoop クラスタの構築検証を行った．

4.1 ネットブート型 PC クライアントによる Hadoop クラスタの構築

2.2 節で述べた Hadoop の構成上の特徴であるスレーブが，ホスト名や IP アドレスなどのネットワーク設定を除き同一の内容で構成できること，また 3 節で述べた起動イメージの変更により，教育用 PC システムのソフトウェア構成や設定を容易に変更できる仕組みを組み合わせ，ネットブート型 PC システムで Hadoop クラスタを構築し，動作することを検証した．なお，事前にネットブート型ではない通常の PC を用いて数台規模の Hadoop クラスタを構築し，Hadoop の設定や運用に必要なコマンドなどの確認を行い，一般的な構

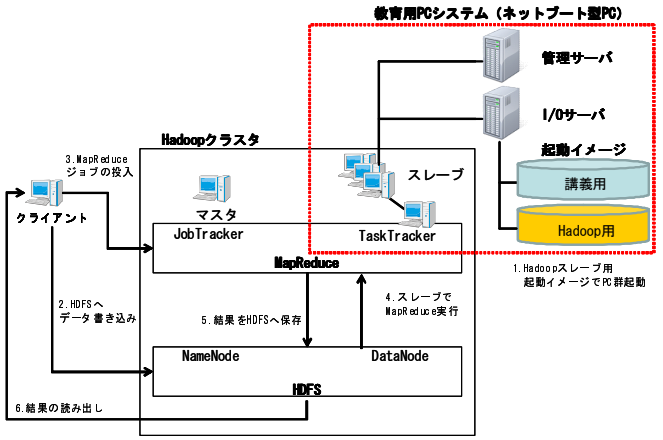

図 4　教育用 PC システムを用いた Hadoop クラスタの構成

成の PC 環境で Hadoop が動作することを確認している．

ネットブート型 PC を用いた Hadoop クラスタのシステム構成について述べる．図 4.1 に示すように，Hadoop クラスタを構成するマスタとスレーブのうち，台数が必要となるスレーブについては，教育用 PC システムのクライアント PC を用いることとする．そのため，IO サーバ上に通常の講義用の起動イメージとは別に，WindowsXP に Hadoop のスレーブとして動作するようソフトウェアを組み込んだ Hadoop 用の起動イメージを準備した．また，マスタについては，別途 PC を 2 台用意し，それぞれ NameNode,JobTracker 用とし，マスタとして動作するよう設定した．マスタ，スレーブ，IO サーバのマシン構成を表 1 に示す．

スレーブ群で唯一異なる個々の IP アドレスとホスト名については DHCP により解決している．IO サーバ上の Hadoop 用イメージで起動するスレーブは DHCP により，MAC アドレス毎に指定された IP アドレスが割り当てられ起動する．割り当てられた IP アドレスに伴いホスト名（FQDN）も一意に定まるため，マスタはスレーブとして利用する PC をホスト名のリストのみで管理できる．

検証ではまず，95 台の PC からなる PC 教室で実施した．別途用意したマスタを起動し，スレーブについてはスレーブ用起動イメージを用いて 95 台の PC を起動した．各 PC の起動後，DFS および MapReduce のデーモンを起動するコマンドを実行し，マスタの管理画

表 1 検証環境の構成

構成	Hadoop(Master)	Hadoop(Slave)	VID IO サーバ
CPU	Pentium4 3.4GHz	Pentium4 3GHz	Xeon 3.2GHz
MEM	3GB	1.5GB	1GB
HDD	1TB	ディスクレス	134GB
NIC	1000Base-T	1000Base-T	1000Base-T
OS	WindowxXP Pro.	WindowsXP Pro.	WinowsServer2003
Hadoop	v0.20.0	v0.20.0	-
Java	JDK1.6.0u14	JDK1.6.0u14	-

面にて全スレーブ 95 台上で必要なデーモンが動作していることを確認した．なお，スレーブ用 PC がディスクレス構成であるため，Windows の C:ドライブ上への書き込みや HDFS や MapReduce に利用する領域は，クライアント PC の搭載メモリの一部をキャッシュ領域として割り当て確保した．このため WindowsXP としては 512MB のメモリで動作させた[⋆1]．

実際に 95 台の Hadoop クラスタが正常に動作することを確認するために，Hadoop に付属のサンプルプログラムを用いて動作確認を行った．マスタから，HDFS に対し容量の異なるログファイルを書き込み，HDFS 上のログファイルに対し，サンプルの grep プログラムにて文字列検索を実施した．PC95 台での grep の実行結果を表 2 に示す[⋆2]．20GB,30GB のログファイルは 10GB のファイルを複数個連結したものである．表 2 より，10GB のファイルに対する検索結果の行数が，20GB，30GB でそれぞれ，2 倍，3 倍となっていること，また参考として行った PC1 台での同じログファイルへの grep の結果[⋆3]と結果が一致することからも，95 台のスレーブが分散して grep を行い，正常な結果を出力できていることが確認できた．

以上の結果より，事前に行った HDD を内蔵した一般的な PC による Hadoop クラスタと同様に，ネットブート型 PC を用いて一つのスレーブ用起動イメージから多数のスレーブを起動した PC 群で，Hadoop クラスタを構成できること，また構成した Hadoop クラスタを用いて正常に分散処理が行えることを確認した[6)]．

4.2 教育用 PC システムを活用した大規模 Hadoop クラスタの構築

前節の検証結果から，図 2 のようなネットブート型 PC クライアントを Hadoop のスレーブとして動作させて Hadoop クラスタを構成，動作できることが確認できた．そこで，より大規模な分散処理を実現するためにスレーブ数を増やし，演算実行ノード数を拡大することを想定し，教育用 PC システム全体を Hadoop クラスタとして利用することを試みた．なお，現時点では教育用 PC システム全体で Hadoop クラスタを構成した場合，通常の講義用の WindowsXP は利用できず，一般の講義利用や学生の自学自習に影響が出るため，今回は教育用 PC システムのメンテナンス日 (日曜日) の作業後の空き時間に，動作検証目的で一時的に Hadoop クラスタを構成した．

⋆1 搭載メモリ 1.5GB のうち，1024MB をキャッシュ領域とし，残る 512MB を OS の使用領域とした
⋆2 処理時間は JobTracker の管理画面での表示結果．ジョブキューの前処理および後処理の時間も含む
⋆3 Hadoop ではなく，Cygwin に付属の grep で実行

表 2 PC95 台での grep の処理結果

サイズ (GB)	95 台 (sec)	結果 (行)	(参考)1 台 (sec)
10	121	45,400	218
20	135	90,800	470
30	181	136,200	721

表 3 想定 PC 数と実起動 PC 数

PC 教室数 (室)	想定 PC 数 (台)	DFS(台)	MapRed(台)
16	1,053	1,004	1,001

教育用 PC システムは，3 節で述べたように，図 2 のような教室単位のネットブートシステムを水平展開し，図 3 のようにして 1,300 台の PC 群を構成しているため，Hadoop クラスタの規模を拡大するには，各教室の IO サーバに 4.1 節で用いた Hadoop のスレーブ用起動イメージをそれぞれコピーして配置するだけである．マスタには，スレーブホストのリストに 1,300 台の PC を登録した．今回の大規模クラスタの構築では，1,000 台規模の構成を一つの目標とし，16 の PC 教室を利用し，1,053 台の PC を対象として検証を行った．

16 の PC 教室はキャンパス内の各建物に物理的にも分散しているため，スレーブの起動およびシャットダウンは，総合情報処理センターのマシン室よりネットワーク経由でコマンドを送出して実施した．表 3 に想定 PC 数と実起動 PC 数を示す．想定した 1,053 台に対し，起動コマンドを遠隔実行した．その後，マスターより対象 PC に対し，HDFS および MapReduce デーモンの起動コマンドを実行した．

その結果，HDFS については 1,004 台が，MapReduce については 1,001 台が動作していることを図 5，図 6 に示す管理画面より確認した．DFS デーモンが 1,004 台で，MapReduce デーモンが 1,001 台で起動していることから，実際の分散処理の実行は DFS と MapReduce デーモンが両方動作する 1,001 台を用いて行った．動作検証には，Hadoop に付属の

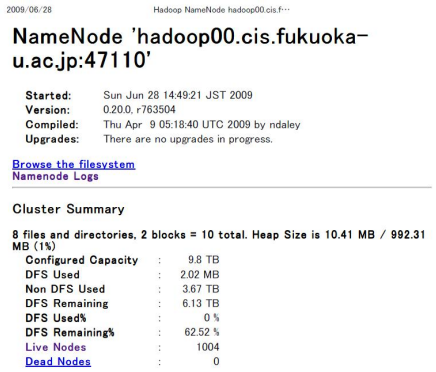

図 5 DFS 管理画面

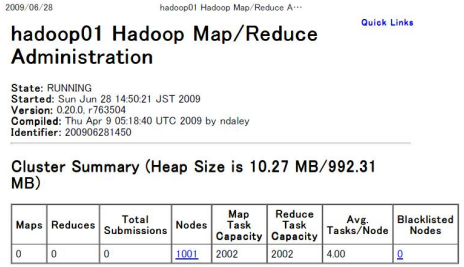

図 6 JobTracker 管理画面

表 4 RadnomWriter と Sort の実行結果

プログラム名	1001 台 (sec)	(参考)1 台 (sec)	(参考)1 台 (h:m:s)
RandomWriter	99	9,610	2h 40m 10s
Sort	299	47,980	13h 18m 10s

RandomWriter と Sort プログラムを用いた．RandomWriter は，各スレーブの PC 上に指定したサイズのランダムな数値データを生成するプログラムであり，Sort は指定した数値データを整列処理させるプログラムである．

Sort 処理の場合，並び替え前のデータに加えて，並び替え後のデータも HDFS へ出力される．しかしながら，ディスクレス PC の構成ではメモリキャッシュの関係で，さほど大容量のデータ処理は実行できないため，今回は RandomWriter により，1,001 台のスレーブ上にそれぞれ 100MB のランダムな数値データを生成させ，クラスタ全体として HDFS 上に 100GB 分の数値データを準備した[⋆1]．次に Sort により，この 100GB の数値データの並べ替え処理を実行した．各コマンドの実行結果を表 4 に示す．また，参考としてスレーブと同程度の性能の PC1 台（HDD 内蔵）をスレーブとして RandomWriter および Sort を実施した場合の処理時間も示す．この結果，PC1,001 台で分散して並べ替え処理を実行した場合 299 秒で Sort が完了し，PC1 台の場合と比較すると 160 倍の性能向上となった．

⋆1 1 レコードは 10byte のキーと 90byte の値なので，10 の 9 乗レコード

4.3 検証結果

4.1 で述べた PC95 台の Hadoop クラスタの構築と動作検証から，Hadoop クラスタを構成する上で一般的に作業量の増加するスレーブ台数の増設を，一つの起動イメージから多数の同一環境 PC を構成することができるネットブート方式のシステム的な親和性が高く，単純なシステム構成で容易にスレーブ台数の拡張を行えることが確認できた．また，4.2 節で述べた 1,000 台の PC を用いた大規模 Hadoop クラスタの構築検証でも，そのメリットは活かすことができ，各 PC 教室の起動イメージを変更するのみ作業のみで Hadoop クラスタの規模を 1,000 台までスケールさせることができた．これらの構築，検証結果から規模をスケールさせることで処理能力の拡大を柔軟に拡大できる大規模分散計算フレームワークを既存の教育用 PC システムを活用することで構築できることが，原理的に可能であることは予測できたが，現実的に可能であり，高い処理能力が得られることとを示せた．

5. 関連研究

教育用 PC システムを活用した分散処理システムとしては，パラメータサーベイのような計算処理を主目的とした広島大学におけるキャンパス・グリッドの事例がある[7]．夜間遊休資源となる WindowsXP ベースの教育用 PC システムのうち，500 台をリモート制御により LinuxOS でグリッドのジョブ実行用ノードとして再起動し，ジョブ管理システムにより割り振られたジョブを実行する仕組みである．また，翌朝には再度リモート制御により，通常の教育用 PC システムとして復帰する．これらの切り替えは自動化され，数分で用途を切り替えられるものである．広島大学の事例も，本稿で報告した検証環境と同様にネットブート型 PC の利点を用いて教育用 PC システムを活用したものである．

また，ネットブート型でない教育用 PC システムの活用例としては，1CD Linux を用いた方法などが提案されている[8),9)]．これらの方法では，PC に内蔵の HDD から OS を起動するのではなく，1CD Linux と呼ばれる CD-ROM から起動可能な Linux システムを用いることで，既存の PC 環境へ影響を与えることなく，別目的のシステムとして活用してする仕組みであり，対象を教育用 PC システムに限定せず，短時間で専用目的の PC クラスタや PC グリッドに切り替えて利用することができる．

紹介した事例[7)–9)] はいずれも従来の並列分散処理の技術により PC クラスタを構成するものであり，用途としては数値計算を主な対象としている．そのため，利用するには MPI などの並列プログラミングに関する知識が必要となるなど，そのシステムの利用を学内等で普及させる上では難しい点があった．また，既存の計算機資源を活用する点でも共通してい

Vol.2009-IOT-7 No.2
2009/10/9

るが，計算機資源を排他的に専用システムに切り替えるため，夜間や教育用 PC システムを利用していない期間しかシステムを運用できないなど，一時的には大規模分散処理が行えるものの運用サービス面では常設の計算機によりもサービスレベルが低下してしまうという制約がある．

6. おわりに

検証の結果，ネットブート型 PC システムの特徴である同一起動イメージで多数の PC クライアントを管理，運用するという仕組みがオープンソースとして広く利用可能な大規模分散計算フレームワークである Hadoop と組み合わせることで，一般的に台数の増加に伴い作業量の増加するスレーブ台数の拡大を単純な構成で行えることが確認できた．また，サンプルプログラムを用いて行った分散処理の結果から，スレーブ台数の増加により処理能力を向上させることができることからも，多数の PC を備えた教育用 PC システムの遊休資源を活用することで，処理能力を柔軟にスケールさせることができることも確認できた．

今後は，大規模分散計算フレームワークを日常の教育用 PC システムの運用と併用して，サービス提供を実現するための課題の検討や解決を図りたい．

参考文献

1) 西田圭介，Google を支える技術，技術評論社，(2008).
2) Hadoop,http://hadoop.apache.org/
3) Yahoo!における 4000 ノードの Hadoop クラスタ，http://developer.yahoo.net/blogs/hadoop/2008/09/scaling_hadoop_to_4000_nodes_a.html
4) 西原壱彦，藤村丞，奥村勝，新教育研究システム FUTURE3 の概要，情報処理教育研究集会予稿集, (2005).
5) 奥村勝，藤村丞，1000 台規模のディスクレス PC システムの構築と運用，情報処理学会研究報告，2008-DSM-23，pp.61–66，(2008).
6) 山邉大樹，奥村勝，教育用 PC システムを利用した Hadoop クラスタの動作検証，電気関係学会九州支部連合大会講演論文集，(2009).
7) 次世代キャンパスグリッド，http://jp.fujitsu.com/featurestory/2005/1116hiroshima-u/
8) 柴田良一，1CDLinux を用いたグリッド上での分散処理による構造最適化に関する基礎的研究，情報処理学会第 66 回全国大会講演論文集，(2004).
9) 手島翼，奥村勝，吉村賢治，1CDLinux を用いたグリッド環境の構築，福岡大学工学集報　No.79，pp.35–45 (2007).