社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－ARC－177 (15)
2008－HPC－114 (15)
2008／3／5

分散クラスタ環境におけるユーザ間資源分配の公平性評価支援システム

太田健介[†1] 合田憲人,[†2,†1]

共有計算機資源を複数のユーザが利用する環境では，ユーザ間の資源分配の公平性を考慮する必要があり，そのためにはユーザの公平性に関するモニタリングが必要になる．しかし，従来のPCクラスタ上のモニタリングツールはシステム稼働率等に焦点が置かれているため，公平性をを評価するための十分な情報を収集していない．本稿では，分散クラスタ環境でのユーザ間の公平性モニタリングツールを提案する．また，本システムが与えるクラスタへの負荷を測定する性能評価実験と，実際の分散クラスタ上のユーザの利用履歴情報に対して本システムを用いた評価実験を行い，本システムの使用が想定される状況下における本システムの有用性を示す．

Fairness Evaluation System of Resource Sharing among Users on Distributed Clusters

Kensuke Ota[†1] and Kento Aida,[†2,†1]

Fairness is one of issues that administrators of shared computer systems need to take account. Although monitoring tools help administrative issues on computer systems, existing monitoring tools focus on system statuses, e.g. utilization, and do not collect information enough to evaluate fairness among multiple users. This paper presents a monitoring tool for fairness among users on distributed PC clusters. The experimental results show that the presented tool gives users effective information to evaluate fairness with low overhead.

1. はじめに

近年，様々な研究機関や企業で地理的に離れた複数のクラスタをネットワーク越しに繋いだ大規模共有計算機資源を使って，より大きな計算機パワーを必要とする研究・開発が可能になりつつあるる[1)2)]．このような共有計算機資源を複数のユーザが同時利用する環境では，ユーザ間の資源分配の公平性を保つ必要があるが，ユーザ間の公平性に配慮した運用は，まだ確立されていない．

公平性に配慮した運用を行うためには，そのためのモニタリングが必要になるが，従来のPCクラスタ上のモニタリングツールではシステム稼働率や計算機ごとの資源使用量等に焦点が置かれており，公平性を評価するための十分な情報を収集することができない．そのため，特定ユーザによる計算機資源の占有が発生したような場合でも，それが管理に反映されづらいという問題がある．

公平性を評価するためのモニタリングツールがあれば，計算機管理者はあらかじめ設定された公平性指標に基づいて評価した各ユーザの計算機資源利用情報を参照することができる．これにより，より良い資源分配ポリシーの策定やユーザ同士の公平性の確保が可能になる．さらに，様々な環境下で得られる公平性指標による評価データの蓄積が体系的な公平性指標の確立に繋がることも考えられる．

本稿では，分散クラスタ環境でのユーザ間の公平性モニタリングツールを提案する．本システムは，定期的に計算機資源の監視を行い各ユーザの計算機資源使用状況を定量的な公平性指標により評価し，その評価値を計算機管理者に分かりやすい形式で可視化し提供する．可視化については，各ユーザの計算機資源利用履歴の公平性指標による評価値をグラフ化する機能や，評価値をWebブラウザで閲覧する機能を持つ．さらに，本システムではテンプレートへ数値を入力するだけで新たな公平性指標を作成可能な入力インターフェースを提供し，計算機管理者が新たに公平性指標を容易に組み込むことができる．このように複数の公

†1 東京工業大学
Tokyo Institute of Technology
†2 国立情報学研究所
National Institute of Informatics

平性指標について評価・比較を可能にすることで，本システムでは，それぞれの環境に適した公平性指標の模索を可能とすることを目指している．

性能評価では，本システムを導入した際のクラスタへの負荷を測定する性能評価実験を行い，クラスタ内の計算ノード上での負荷が非常に小さく，実用上の問題ないことを確認した．また，実際に利用されている分散クラスタにおけるユーザの利用履歴情報を本システムを用いて評価し，本システムの使用が想定される状況下において本システム提供する情報が計算機管理上，有用であることが確認された．

2. 分散クラスタ環境

本稿が対象とする分散クラスタ環境は，複数のサイトに分散された複数の PC クラスタから構成される．ユーザは，本環境上の全ての PC クラスタ上の計算ノードへ直接ログインすることや，計算ノード上でジョブを実行することが可能であり，バッチスケジューラによって利用される計算ノードの制御を行う一般的な大規模計算機システムとは異なり，本分散クラスタ環境ではユーザの自由度がかなり大きい．そのため，公平性を考慮した運用の重要性が非常に高い．このような分散クラスタ環境の実例としては，PlanetLab[3)4)]，Grid'5000[2)]，InTrigger[1)] が挙げられる．

PC クラスタの利用状況監視に用いられるソフトウェアは，数多く存在する．広くクラスタ監視に用いられているソフトウェアの例として ganglia[5)] などが挙げられる．これら一般的なリソースモニタリングツールは，計算機の利用状況を監視し，障害発生の検知やロードアベレージや CPU 使用率などに焦点を当てて監視を行いシステム全体の稼働率を上げることを目的としている．そのため，各ユーザの計算機利用状況を監視する機能が十分に提供されていない．

3. 公 平 性

本稿で述べる公平性とは，計算機資源の利用状況を，ロードアベレージやシステム稼働率といったシステム側からの観点ではなく，ユーザ側の観点からその "公平さ" を評価したものを意味する．

公平性に配慮した運用は，従来の計算機センター等で用いられている従量制課金による方法[6)] と，フェアシェアによる方法[7)] に分けられる．後者のフェアシェアでは，ユーザ毎の過去の計算機資源利用量からその後のユーザの計算機資源利用可能量を決定する．具体的には，直近の利用量が多いユーザほど，直後の利用可能量が少なく制限される．

3.1 公平性指標

本節では，提案システムに実装されている公平性指標の具体的な説明を行う．現在，実装されている指標は以下の 2 つである．

3.1.1 従量制指標

従量制指標は，従来の計算機センター等で用いられている従量制課金に相当する指標である．従量制課金において，課金額を決定する際に考慮される要素として"計算機資源使用量"と"計算機の性能"が挙げられる．計算機資源使用量が多ければ線形に課金額は増え，使用した計算機の性能が高ければ，やはり課金額は増える．そこで，本システムでは以下のようなパラメータを設定する．

- **RV(Real Value):** ユーザの計算機資源使用量を示す値
- **CV(Conclusive Value):** 計算機毎の使用時間に対する課金額を考慮した最終的な値

RV は，ユーザの計算機資源使用量を測定したものである．RV の初期値は 0 で，ユーザが計算機資源を使用すればする程増大する．

時刻 t におけるユーザ u の RV，$\sigma_r(u,t)$ は全ての時間間隔 δt において以下の式で計算される．

$$\sigma_r(u,t) = \sigma(u,t-\delta t) + \rho(u,t) \quad (1)$$

$\rho(u,t)$ は時刻 t においてユーザ u によって使用される計算機資源量を示す．

CV は，計算機に応じた課金係数を RV に乗算した値である．CV は，RV に対して線形の関係がある．特に条件がない場合は，課金係数は 1.0 で，CV と RV は等しい．

時刻 t においてユーザ u の CV，$\sigma_\epsilon(u,t)$ は次のように計算される．

$$\sigma_\epsilon(u,t) = \sigma_r(u,t) \times f(x) \quad (2)$$

$f(x)$ は計算機 x の課金係数を示す．

3.1.2 Condor の FairShare 指標

2 つ目の指標は，クラスタのジョブスケジューリングシステム Condor でジョブの実行優先順位を決定するために用いられている fair share アルゴリズム[7)] における公平性指標である．

Condor では，ジョブへのマシンの分配を決定するために，以下に示すプライオリティを用いる．

- **RUP(Real User Priority):** ユーザの計算機資源使用量を示す値
- **EUP(Effective User Priority):** ユーザが利用できる計算機資源量を決定する値

RUP は，ユーザの計算機資源使用量を測定したものである．RUP の初期値は 0.5 で，定常状態でユー

ザの RUP はそのユーザによって使われる計算機資源量に釣り合う．また，RUP には半減期 **PRIORITY_HALFLIFE** が設定されている．例えば，**PRIORITY_HALFLIFE** − 86400 秒 (一日) で RUP − 10 ならば，RUP は 1 日後には 5，2 日後で 2.5 と減少する．

時刻 t におけるユーザ u の RUP，$\pi_r(u,t)$ は全ての時間間隔 δt において以下の式で計算される．

$$\pi_r(u,t) = \beta \times \pi(u,t-\delta t) + (1-\beta) \times \rho(u,t) \quad (3)$$

$\rho(u,t)$ は時刻 t におけるユーザ u によって使用される計算機資源量を示す．$\beta = 0.5^{\delta t/h}$ であり，h は **PRIORITY_HALFLIFE** によって定義される半減期を示す．

EUP は，ユーザが利用を許される計算機資源量を決定するのに用いられる．ユーザの利用可能な計算機資源量は EUP の値に反比例する．例えば，ユーザ A，B，C の EUP が 20，10，5 である場合，それぞれのユーザへの計算機資源の割り当て比は，ユーザ A:ユーザ B:ユーザ C = 1:2:4 となる．EUP は，RUP に対して線形の関係にあり，ユーザ毎にプライオリティ係数を設定することができる．通常は全てのユーザのプライオリティ係数は 1.0 で，EUP と RUP は等しい．

時刻 t におけるユーザ u の EUP，$\pi_e(u,t)$ は次のように計算される．

$$\pi_e(u,t) = \pi_r(u,t) \times f(u,t) \quad (4)$$

$f(u,x)$ は計算機 x におけるユーザ u のプライオリティー係数を示す．

本稿では，従量制指標および Condor の FairShare 指標において，各ユーザが使用した CPU 時間の合計値を計算機資源量とする．CPU 時間は各ユーザの実行プロセスが終了した際に検出され，公平性指標の評価値が計算される．つまり，$\rho(u,t)$ を以下のように定義する．

$$\rho(u,t) = \sum_{k=1}^{n(u,t)} CPUTime(u,t,k) \quad (5)$$

$n(t)$ は時刻 $t-\delta t$〜t の間に終了したユーザ u のプロセス数であり，$CPUTime(u,t,k)$ は時刻 $t-\delta t$〜t の間に終了したユーザ u のプロセス k の CPU 使用時間を示す．

4. 公平性評価支援システム

本節では提案システムの概要を述べる．本システムは，各ユーザの計算機資源利用状況を監視し，公平性指標により評価した結果を時系列グラフにして表示する．公平性指標は，本システムに実装されている指標の中から選択して用いる他，計算機管理者が自ら追加した指標を用いることもできる．

4.1 システム動作概要

本システム全体の一連の処理を図 1 に示す．システム内部の動作は以下のとおりである．

(1) 各クラスタ内に 1 つ設置された情報集積サーバ (Archiving Server) が，クラスタ内の監視対象である計算ノード (以降，監視対象ノード (Monitoring Node) と呼ぶ) に，ユーザの計算機資源利用状況を問い合わせる．この処理は定期的に繰り返される．

(2) 各監視対象ノードは情報集積サーバからの問い合わせがあった場合，設定されている公平性指標によってユーザの計算機資源利用量を評価し，その結果を情報集積サーバに通知する．

(3) 情報集積サーバはクラスタ内の全監視対象ノードの問い合わせた後，分散クラスタ上の遠隔地にある他のクラスタの情報を収集し，クラスタ間で情報の同期を行う．

(4) 情報集積サーバは，収集した各ユーザの計算機資源利用量の公平性指標による評価結果からグラフ，HTML ファイルを生成する．

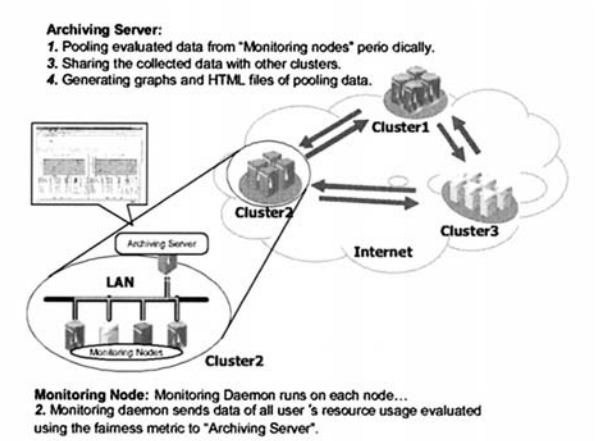

図 1 公平性評価支援システム動作概要

4.2 システム構成

本節では，4.1 節で説明したシステムを実現するためのシステム設計と実装について述べる．本システムの構成図を図 2 に示す．本システムは，情報集積サーバ (Archiving Server) と監視対象ノード (Monitoring Node) から構成される．

また，本システムは Linux(Debian,Redhat) 上で動作し，Perl とラウンドロビンデータベースツールの一

種である RRDTool[8] を用いて実装されている．

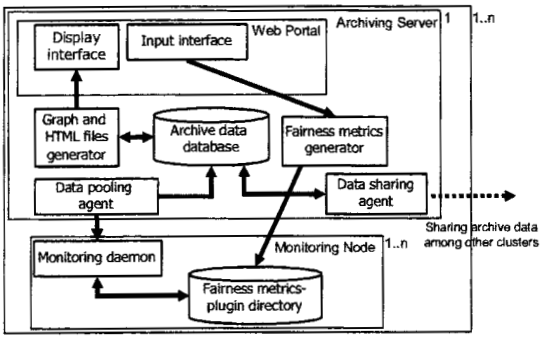

図 2　公平性評価支援システム構成図

4.3　情報集積サーバ

情報集積サーバは，各クラスタ上の 1 ノードで動作する．情報収集エージェント (Data pooling agent) が，定期的に監視対象ノードから，各ユーザの資源利用量の公平性指標による評価結果を取得し，時系列データベース (Archive data database) に保存する．そして，情報共有エージェント (Data sharing agent) によって遠隔地にある他のクラスタと時系列データベースの同期を行う．集積した情報をもとに可視化エージェント (Graph and HTML files generator) によってグラフ，HTML ファイルが生成され，Web ブラウザからの閲覧が可能になる (Display interface)．また，Web インターフェース (Input interface) から，利用者独自の指標をシステムに組み込む機能も提供される (Fairness metrics generator)．

4.4　監視対象ノード

監視対象ノードは各クラスタ上の全計算ノードで動作する．監視デーモン (Monitoring daemon) が常駐し，各ユーザのプロセス ID を管理している．そして，情報集積サーバからの要求が来ると，各ユーザの計算機資源利用量に対して公平性指標による評価を実行した結果を情報収集エージェントに通知する．

4.5　インターフェース

情報集積サーバは，集積した各ユーザの計算機資源利用量の公平性指標による評価結果を Web ポータルによって提供する．また，Web ポータルは新規公平性指標を追加するためのインターフェースを持つ．Web ポータルの構成と機能を表 1 に示す．

表 1　Web ポータルの構成と機能

トップページ	トップページは，システムの利用方法やプラグインの追加方法についての解説資料が置かれ，基本的なシステム利用を助ける．
グラフ閲覧用ページ	グラフ閲覧用ページは，ユーザの計算機資源利用量の公平性指標による評価結果のアーカイブから生成したグラフを，ユーザ毎に分類して表示する．インデックスページからホスト別，ユーザ別のリンクを辿ることで，容易に目的のユーザ情報を得ることができる．
新規公平性指標入力ページ	新規公平性指標入力ページは，新規に公平性指標をシステムに組み込むためのインターフェースを提供する．公平性指標は，スクリプト言語等で書かれた実行ファイルとして実装される必要がある．そのため，入力フォームに入力された情報を元に CGI を用いて指標スクリプトを生成する．公平性指標の入力方法は以下の 3 種類が用意される．
1. 直接入力方式	利用者が実行可能なスクリプトを入力する．入力されたものを，そのまま公平性指標としてシステムに組み込む．
2. ウィザード方式	あらかじめ公平性指標のテンプレートが用意されており，利用者が入力フォームに従って必要なパラメータを入力するだけで，指標スクリプトを自動生成し，それをシステムに組み込む．直接入力方式に比べて自由度は下がるが，容易に公平性指標を作成することができる．
3. 詳細ウィザード方式	上記のウィザード方式と同様に，あらかじめ公平性指標のテンプレートが用意されており，入力フォームに従って必要なパラメータを入力する方式．ただし，用意されているテンプレートがウィザード方式よりシンプルで入力項目が多くなっている分，作成できる指標の自由度も高くなる．

5.　実　　験

本節では，本システムを評価するために行った 2 種類の実験について述べる．

5.1　性能評価

本システムの実用性を示すために，実際の分散クラスタ環境 InTrigger[1] の一部のクラスタを実験環境として，本システムの性能評価実験を行った．

本実験では，NAS Parallel Benchmarks (NPB)[9] を用いて，本システム稼動時・非稼動時におけるベンチマーク・プログラムの実行時間を測定する．それら実行時間の差が大きければ，本システムが計算機に与える負荷が大きいということになり，差が小さければ，本システムが計算機に与える影響は小さいと言える．

表 2　性能評価実験クラスタ詳細

情報集積サーバ	1 台
監視対象ノード	35 台
OS	Debian
CPU	Intel(R) Core(TM)2 CPU 6400 2.13GHz
Memory	4GB
MPICH Version	1.2.7p1
NPB Version	3.2.1

性能評価実験に用いたクラスタの詳細を表 2 に示す.

5.1.1　情報集積サーバ負荷測定結果

本実験では，情報集積サーバ上で情報集積サーバのプログラム群が全て動作している場合と全て停止している場合とで，NPB の実行時間をそれぞれ測定した．具体的には，NPB の全ての対象問題の問題サイズ A, B, W に対して，プロセス数 1 として NPB プログラムを生成し，mpirun コマンドで 10 回実行した平均値を算出した．

本実験の結果，情報集積サーバ・プログラム動作時は，非動作時に比べて，対象問題間で差はあるものの平均で 9.8 % (最大で 50 秒) 実行時間が増大したことが確認された．今回の実験では，この情報集積サーバは監視対象ノード，全 35 ノードを監視している．しかし，さらに監視対象ノードが増加すれば，情報集積サーバの負荷が高くなり，今回の測定結果以上に NPB 実行時間が増大することが考えられる．

以上の結果より提案システムでは，各クラスタ内の 1 ノードを情報集積サーバ専用ノードとして設置し，本ノードでは計算を行わないことが推奨される．

5.1.2　監視対象ノード負荷測定結果

本実験では，クラスタ内の 35 台の監視対象ノード上で監視デーモンが動作している場合と，停止している場合とで，全監視対象ノード上で NPB を並列実行した時の実行時間をそれぞれ測定した．具体的には，全ての対象問題の問題サイズ A, B に対して，プロセス数 64 として NPB プログラムを生成し，mpirun コマンドで 10 回実行した平均値を算出した．

監視デーモン動作時は，非動作時に比べて，平均で 8.2 % (最大で 3.6 秒)NPB 実行時間が増大した．以上の結果より，監視対象ノードにおける本システムの負荷は非常に小さいことが確認された．

5.2　有用性評価

本システムの有用性を示すために，Grid Workloads Archive 上で公開されている Grid5000 の運用履歴[10] から抽出したユーザ毎の計算機資源利用履歴を，本システムを用いて公平性指標による評価を行った．公平性指標として，3.1 節で紹介した従量制指標と FairShare 指標を用いた．

図 3，図 4 は，Grid5000 のユーザ利用履歴から無作為に選択した 1 週間分の利用履歴について，本システムが従量制指標または FairShare 指標を適用して評価した例である．グラフは，ユーザ毎の指標による評価値を示したものであり，x 軸は時間 (日)，y 軸は公平性指標による評価値を示す．

(1)　**公平性評価支援システム利用例 1**

ここでは，実際に計算機管理者が，利用者の週ごとの計算機資源利用量をなるべく一定基準以下に抑え，過度な使用の抑制し計算機資源利用の平均化を図りたいという要求を想定する．

本システムが利用実績を指標によって評価し，結果を図 3 のように可視化し提供することで，週前半に多量の計算機資源を利用し評価値が高い図 3 中① のユーザ user165，user569，user570 に対して利用制限を行えばよいことがわかり，週後半の利用を抑制するといったことが可能になる．また，計算機資源が空けば，図 3 中② のユーザ user267，user571，user572 のように積極的に利用するユーザが現われ，長期的な計算機資源利用の平均化も期待できる．

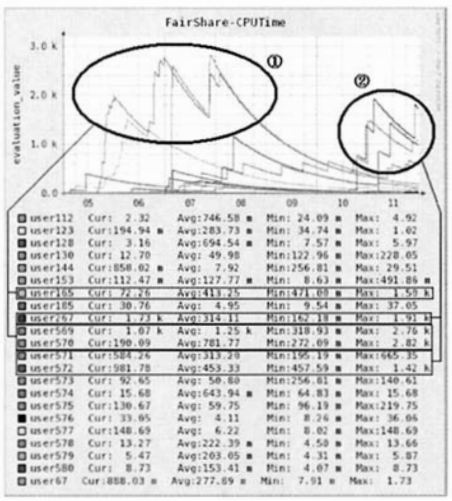

図 3　Grid5000 ユーザの計算機資源利用実績 (1 週間分) の FairShare 指標による評価結果 (例 1)

(2)　**公平性評価支援システム利用例 2**

この例では，計算機管理者が，各ユーザの計算機資源利用傾向を調べたいという要求を想定する．

分散クラスタ管理者は，グラフから直感的にユーザの計算機資源利用傾向を知ることができる．例えば，図4中の①においてuser277とuser431の期間最後の時点での評価値はほぼ同じ値になっているが，user277のグラフは階段状になっており，実行時間の短いジョブを連続して実行しているということが分かる．また，user431のグラフは急に評価値が高くなっており，1つの大規模なジョブを長時間実行させていたということが分かる．以上より，最終的な計算機資源利用量は同じ場合でも，ユーザ毎の詳細な利用傾向を把握することが可能である．これは，本システムが数値データとしてではなく，グラフという可視化され直感的にデータの傾向を把握しやすい状態で公平性指標による評価値を提供しているためである．

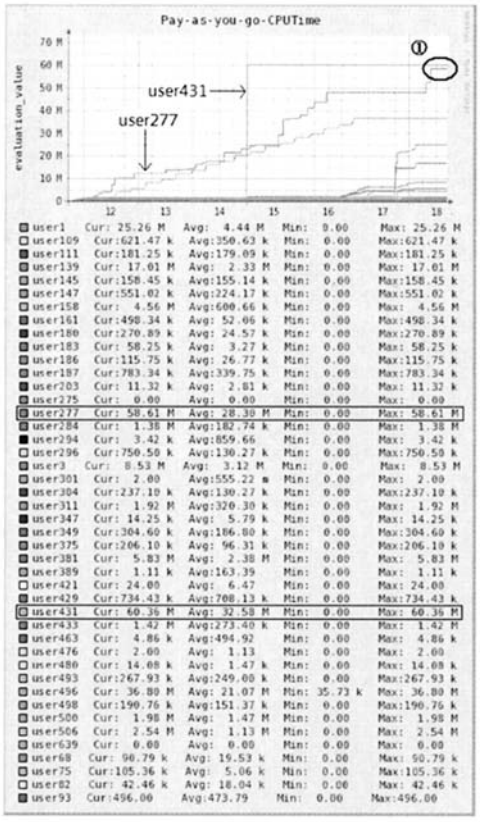

図4 Grid5000ユーザの計算機資源利用実績（1週間分）の従量制指標による評価結果（例2）

6. まとめ

本稿では，分散クラスタ環境におけるユーザ間の公平性モニタリングツールを提案した．本システムは，各ユーザの計算機資源利用状況を定量的な公平性指標を用いて評価した結果を可視化し計算機管理者に提供する．さらに，Webインターフェースから，利用者独自の指標をシステムに組み込む機能を提供する．本システムの性能評価を行った上で，実際の分散クラスタ上でのユーザの利用履歴情報を本システムを用いて評価し，本システムがユーザに公平な資源分配ポリシーの策定に役立つことを示した．

また，本システムの運用において，監視対象ユーザ数・監視対象ノード数の増加に伴う情報集積サーバの負荷増大により，スケーラビリティに制限が生じる問題がある．よって，今後は情報集積サーバへの負荷を減らすため，実装面での見直しを行っていく予定である．

謝辞

本研究の一部は，科学研究費補助金特定領域研究「情報爆発に対応する新IT基盤研究支援プラットフォームの構築」による．

参考文献

1) Intrigger. http://www.infoplosion.nii.ac.jp/infoplosion/html/S/A02-1/index.html.
2) Grid'5000. https://www.grid5000.fr/mediawiki/index.php/Grid5000:Home.
3) Planetlab. http://www.planet-lab.org/.
4) Larry Peterson, Tom Anderson, David Culler, and Timothy Roscoe. A blueprint for introducing disruptive technology into the internet. *SIGCOMM Comput. Commun. Rev.*, Vol. 33, No.1, pp. 59–64, 2003.
5) Ganglia monitoring system. http://www.ganglia.info/.
6) Tsubame grid cluster. http://www.gsic.titech.ac.jp/ccwww/tgc/kakin.html.
7) Condor version 6.7.20 manual. http://www.cs.wisc.edu/condor/manual/v6.7/3_5User_Priorities.html.
8) Rrdtool - about rrdtool. http://oss.oetiker.ch/rrdtool/.
9) Nas parallel benchmarks. http://www.nas.nasa.gov/Resources/Software/npb.html.
10) The grid workloads archive. http://gwa.ewi.tudelft.nl/pmwiki/pmwiki.php?n=Main.Home.