京都大学におけるグリッド研究の現状と今後の計画について

平野 彰雄 *

1 はじめに

京都大学学術情報メディアセンター (以下、本センターという) では、 2001 年から、他の６全国共同利用センターと協調してグリッドコンピューティング環境構築の調査研究を進めてきた。

本稿では、まず、本センターが他センターとの接続実験の前段階として行ってきたグリッド環境構築のためのミドルウエア Globus Toolkit[1] および Globus 対応のメッセージパッシングライブラリ MPICH-G2[2] の移植とセンター内のシステム間で進めてきた実証実験の内容と結果について報告する。

また、 2002 年 8 月以降、名古屋大学情報基盤センター、北海道大学大型計算機センターおよび東京大学情報基盤センターとの間で Super SINET 経由での接続試験を開始した。その成果と現状について報告する。

さらに、グリッドコンピューティング構築に向け、より実際的な研究を行うために調達したグリッド研究用システムの概要を説明する。

最後に、センター内および他センターとの間で計画しているグリッド実験計画について紹介するとともに、本センターが推進するグリッド研究計画並びに共同研究推進体制について述べる。

2 センター内での実証実験について

2.1 システム構成と Super SINET 接続

本センターのグリッド研究用システムには、まず、スーパーコンピュータ Fujitsu VPP800/63 (以下、 VPP800 という) および共有メモリ型の並列計算機 Fujitsu GP7000F/900 (以下、 SPP という) がある。これらの２台のシステムは図 1に示すような構成で Super SINET に接続され、他センターとのグリッド実験に使用している。

また、これらのシステム以外にも、センター内でのグリッド実験のために Fujitsu PRIMEPOWER200 (以下、 gridws という) および Linux を搭載した 2 台の PC を使用しており、これら３台のシステムは、館内の LAN に接続されている。

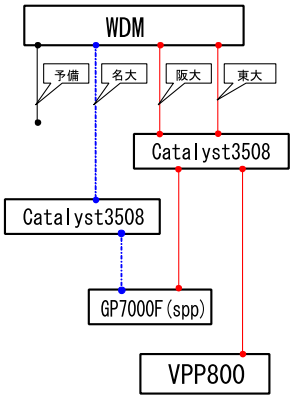

図 1. Super SINET 接続図

表 1に、本センターのグリッド研究用システムの構成とインストールしているソフトウエアのバージョンを示す。なお、この間に幾度かソフトウエアのバージョンアップを行っており、表 1では、最新のバージョンを載せている。

2.2 Globus の移植および評価

本センターが、最初に Globus Toolkit を移植したのは、 2001 年 12 月である。当時、最新のレベルであった Globus2.0 βを SPP および Linux にインストールして、グリッド研究のために Globus 機能およびコマンドについて、調査を開始した。

2002 年 3 月には、富士通 (株) による VPP800 への Globus の移植が完了し、 Globus1.1.4 が提供された。なお、 VPP800 への導入は、 63 台ある PE(Processor Element) の中で、 P-PE および 1 台の IMPE の 2 台 PE に、 Globus1.1.4 をインス

* ひらの あきお (京都大学 学術情報メディアセンター)

表 1. システムの構成とソフトウエア

システム	OS	Globus	MPICH-G2	Network
VPP800(P-PE)	UXP/V	1.1.4	1.2.4	LAN
VPP800(IMPE)	UXP/V	1.1.4	1.2.4	SINET
GP7000F/900(SPP)	Solaris8	2.0	1.2.4	SINET
PRIMPOWER200(gridws)	Solaris8	2.2	1.2.4	LAN
PC	Linux	2.0	1.2.4	LAN
PC	Linux	2.0	1.2.4	LAN

トールして、VPP800 の PE 間でのテスト環境も確保している。

また、VPP800 への Globus1.1.4 の導入に合わせて、SPP をはじめ他システムも Globus2.0 へのレベルアップを行い、異機種間でのテストを開始した。テストの結果、同じバージョンの Globus 同士の間では問題が無いが、Globus2.0 をクライアント、Globus1.1.4 をサーバとする組合わせでは、バッチジョブの結果を取り出す globus-job-get-output コマンドでエラー (error code 39) が発生することが明らかになった。

このエラーは Globus Project[1] の Error FAQ にも掲載されている障害であり、Globus2.0 側のコマンドスクリプトで、サーバが VPP800 の場合には、環境変数名を `GLOBUS_LOCATION` から `GLOBUS_INSTALL_PATH` に変更を行った。

2.3 MPICH-G2 の移植および評価

2002 年 7 月、富士通 (株) より VPP800 用に MPICH-1.2.1 が提供されたのを契機に、MPICH-G2 を用いた複数システム間を跨る MPI プログラムのテストを開始した。まず、VPP800 で、PE 内に閉じた MPI プロセス間での検証、2PE 間に跨る MPI プロセス間で動作を検証した。

つぎに、異機種間でのテストのために SPP および gridws へ MPICH-G2 を移植した。移植したのは、その時点での最新のバージョンであった MPICH-1.2.4 である。テストの結果、SPP と gridws という同じ機種間では、動作することが確認できたが、異機種となる SPP と VPP800 との間では、MPI プロセス間の通信においてデッドロックが発生することが判明した。

この現象を切分ける目的で、新たに 2 台の PC に Linux を載せ、さらに Globus2.0 および MPICH-1.2.4 のインストールを行いテストを実施した。

テストの結果、Linux 同士では問題なく動作するが、異機種となる VPP800 や SPP などでは、Linux 側の MPI プロセスでエラーが発生した。エラーメッセージから原因を調査したところ、Linux に対する MPICH-G2 の実装においてバグがあり、既に PATCH が提供されていたので、これを適用することで、SPP と Linux という異機種の組み合わせでは、MPI の動作を確認できた。

VPP800 と他のシステム間で発生するデッドロック現象は、MPICH-G2 のバージョンが異なるための非互換の可能性を考え、センターで VPP800 に MPICH-1.2.4 をインストールした。バージョンを一致させることで、デッドロック現象は回避できたが、VPP800 との間では、新たに別のエラー事象は発生する場合があることが明らかになった。詳細については、現在調査中である。

なお、計算グリッドにおいて Fortran のサポートが必須であるといえるが、VPP800 の MPICH-G2 ではサポートされていない。本センター独自に、Solaris8 および g77 という組合わせで Fortran をサポートすることができたので、VPP800 でも同じように試みたがインストールできなかった。

MPICH-G2 において Fortran が使えないのは、グリッド研究にとって致命的ともいえる。しかし、MPICH-G2 が Fortran をサポートしていない訳ではなく、インストールに失敗するだけなので、移植に関するノウハウなどの情報を整備が必要であると考える。

3　他センターとの接続実験

3.1　名大センターとの接続実験

2002 年 8 月、名古屋大学情報基盤センター (以下、名大センターという) のスーパーコンピュータ Fujitsu VPP5000/64 (以下、VPP5000 という) との間で SuperSINET 経由での接続実験を開始した。名大センターの VPP5000 は、本センターの VPP800 の後継機種であるので、この接続実験は同じ機種のスーパーコンピュータ同士を Super SINET を介して接続するものである。

名大センターとの接続試験に関係するシステムおよびソフトウエア、バージョンを表 2に示す。

表 2. 名大センターとの接続試験のシステム構成

センター	システム	アドレス	Globus	MPICH
名 大	VPP5000	172.22.0.4	1.1.4	1.2.1
京 大	VPP800	172.22.0.5	1.1.4	1.2.1
	SPP	172.22.0.6	2.0	1.2.4

3.1.1　Globus および MPICH-G2 の検証

表 3にテストを実施した Globus コマンドと結果を示す。基本的に同じメーカーの同じスーパーコンピュータ同士の接続であり、 Globus のバージョンも同じなので、問題なく完了した。

表 3. Globus コマンドとテスト結果

コマンド名	機 能	結 果
globusrun -a -r	リモート認証の確認	○
globus-job-run	コマンドの実行	○
globus-job-submit	バッチジョブの投入	○
globus-job-status	〃　状態表示	○
globus-job-get-output	〃　結果の取出し	○
globus-job-cancel	〃　キャンセル	○
globus-job-clean	〃　結果の削除	○

また、両センターのスーパーコンピュータ間を跨る MPI プログラムも問題無く動作することが確認できた。一方、 Super SINET を跨る MPI プロセス間での通信性能は、現状では余り期待できないことも明らかになった。

3.1.2　ネットワーク性能評価

MPICH-G2 での通信性能の問題を整理する目的で、 Netperf[3] によりネットワークの性能を試みたが、両センターのスーパーコンピュータが運用システムであり、システム負荷が高く、正確なデータを得ることができなかった。

なお、両センターのスーパーコンピュータ間の RTT (Round Trip Time) を実測すると約 20 ミリ秒であり、 Globus および MPICH-G2 などが TCP(Transmission Control Protocol) を用いて通信を行っているので、ネットワークレベルでのチューニングが必要であることが明らかとなった。

3.1.3　リモート可視化実験

グリッドコンピューティングでは、他のセンターの特色あるアプリケーションをリモートから活用できる枠組みの構築も重要なテーマである。

今回、名大センターでサービスされている 3 次元流体解析プログラム α -flow およびリアルタイム可視化ツール VisLink を使用して、解析結果を本センターの汎用可視化ツール AVS により表示するリモート可視化の実験も行った。

図 2にシステム構成と処理の流れを示し、図 3に可視化結果を示す。

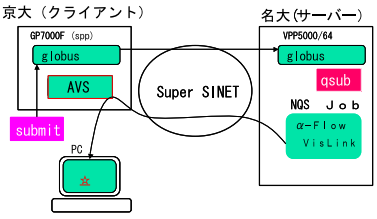

図 2. リモート可視化構成図

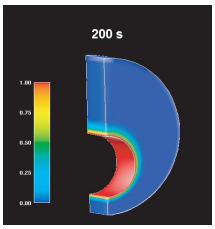

図 3. 可視化結果

3.2 北大センターとの接続実験

2002 年 12 月、北海道大学大型計算機センター (以下、北大センターという) との接続試験を開始した。北大センターのシステムの構成を表 4に示す。

表 4. 北大センターのシステム構成

システム	アドレス	Globus
SR8000	172.22.16.4	2.0
SR8000-Compact	172.22.16.2	2.0
SGI Onyx	172.22.16.5	2.0

実施したテストは、 Globus コマンドによる基本機能の確認である。

北大センターの SR8000 とは、異機種間でのテストであったが、 SR8000 をクライアントとし、本センターの VPP800 および SPP へのテストは、 Globus のバージョンの違いによる問題を除けば 1 日で一応完了した。また、本センターのシステムをクライアントとしたテスト、また、北大センターの SR8000 以外のシステムに対するテストも基本的に完了した。

また、 2003 年 1 月、本センターのメディアコンピューティング研究分野の西村直志 教授は、土木学会主催の［計算力学フォーラム］において「ネットワークコンピューティングの可能性」と題して講演され、その中で本センターの SPP、 gridws および北大センターの SR8000 を使用してグリッドコンピューティングの実演を行われた。

3.3 東大センターとの接続実験

2003 年 1 月、東京大学情報基盤センター (以下、東大センターという) との接続試験を開始した。東大センターのグリッド研究用システムを表 5に示す。

表 5. 東大センターのシステム構成

システム	アドレス	Globus
SR8000-Compact	172.22.18.10	2.0

なお、東大センターとの接続実験では、東大センターが独自に上げられた認証局 (CA: Certificate Authority) を本センターの SPP 上の Globus2.0 に追加し、 Globus コマンドでの基本的な動作確認を行なうことができた。

表 6. 調達物品の諸元

計算サーバ	
Fujitsu PRIMEPOWER HPC2500	
CPU(1.3GHz)	96 台
ピーク性能	499.2 GFlops
メモリ	192GB
Disk	3TB
ルーター	
Catalyst4006	
1000Base-LX MMF	4 Port
1000Base-SX SMF	4 Port
1000/100Base-T	24 Port
サブシステム	
IBM pSeries 630/6E4	
CPU(1.0GHz)	4 台
ピーク性能	16.0 GFlops
メモリ	8GB
簡易 3D 装置	
裸眼で 3D 表示	

4 グリッド研究システム

4.1 システム構成と諸元

本センターで Super SINET に接続される VPP800 および SPP は、いずれも利用者サービスを行なっているシステムであり運用の妨げになるような実験はできないので、本格的な実験を行うためにグリッド研究システムを新たに調達した。

調達したシステムの概要を図 4に示し、各システムの諸元を表 6に示す。

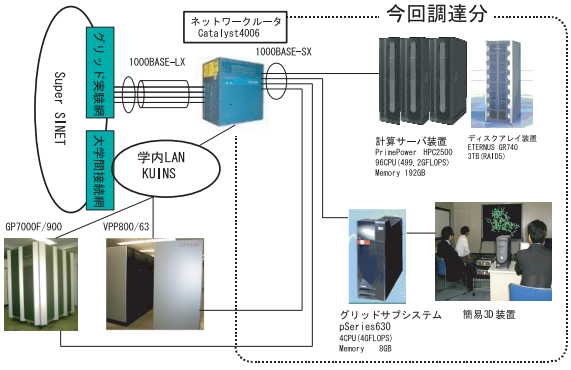

図 4. グリッド研究システムの構成

4.2 ネットワーク構成

現在、本センターでの Super SINET の接続は、図 1に示すように Catalyst3508 を使用してレイヤー２で接続している。

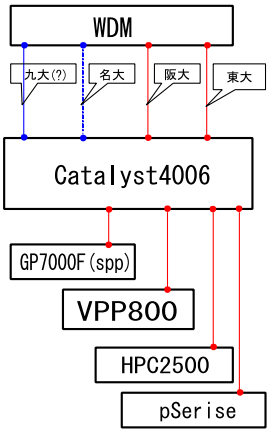

図 5. ネットワーク構成 (案)

これを図 5に示すように、今回、調達した Catalyst4006 を用いてレイヤー３に移行し、本センターに割当てられたグリッド研究用サブネット (172.22.16.0/24) を立ち上げ、これまでセンターの館内 LAN に接続していたシステムも含めて、色々なシステムおよび機器を収容することを検討している。

他センターとも必要な調整を行ない、今年度中には、移行したいと考えている。

5 今後の課題と実験計画

5.1 バージョンアップと版数管理

Globus Toolkit は、Globus2.2 が最新のバージョンであり、これのバグフィックス版 Globus2.2.4 が 2003 年 2 月 14 にリリースされている。

表 1に示すように、既に gridws は Globus2.2(2.2.3) にバージョンアップを行い、他のバージョンの Globus との間でテストを行ったが、非互換など問題は無いようなので、他のシステムも順次 Globus2.2 にバージョンアップを考えている。

また、MPICH-G2 も Globus2.2 の新機能を取入れた、対応バージョン MPICH-1.2.5.1 が 2003 年 1 月 31 日にリリースされている。Home Page[2] では、プロセス間での集団的な通信において、複数の TCP コネクションを確立することで性能向上を図っていることが明示されているのでバージョンアップを行い評価することを考えているが、但し、下位のバージョンとは互換性がない事も明記されている。

ただし、非互換は７センター間での接続試験を進める上で、障害となる可能性がある。したがって、当面、MPICH-G2 は全てのバージョンのインストールを行い、バージョン毎にディレクトリを分けて管理することを考えている。

5.2 ７センター間での接続試験

これまでに名古屋大学情報基盤センター、北海道大学大型計算機センター、東京大学情報基盤センターのスーパーコンピュータとの間で接続試験を始めることができたが、残る東北大学シナジーセンター、大阪大学サイバーメディアセンター、九州大学情報基盤センターとの間での接続試験を早期に開始したいと考えている。

5.3 東大、北大、京大での実験計画

これは東大センター、北大センター、京大センターの間で実験を計画しているものであるが、現在、本センターが Super SINET に接続しているシステムは利用者サービス用のシステムであり、高負荷となるような試験やシステムの再起動がいるような実験はできない。したがって、グリッド研究システムが導入される３月以降に行う計画である。

３センターで予定している実験は、つぎのようなものである。

1) Super SINET 経由での
   NFS(Network File System) の性能評価
2) アプリケーションから見た通信性能

5.4 TCP 通信性能のチューニングと評価

3.1.2でも述べたが、Globus および MPICH-G2 などは通信に TCP を用いているが、TCP の設計、実装から既に＋数年が経過しており、今日のギガビッ

ト級の高速ネットワークには必ずしも適合していない。

その典型が TCP のウインドウサイズの問題であり、 TCP では、ウインドウサイズ分のデータを送信すると、 ACK 応答を受信するまでつぎのデータを送信を行わない。ウインドウサイズの標準は、 64KB であり、 RTT が 20 ミリ秒の場合、 64KB/20 ミリ秒、すなわち 25.6Mbps が一つの TCP コネクションでのデータ転送速度の眼界である。

この問題は、既に 1990 年代前半に広く認識されており、「TCP window scaling」 (RFC1323,1992 年)[5] が拡張仕様として標準化されている。しかし、これはあくまでのオプショナルな拡張仕様であり、高速インターフェースを実装している計算機でも、サポートしていることは少ない。

一方、 OS レベルではサポートされていることも多く、幸い、本センターが導入するグリッド研究用システムの OS である Solaris では、動的にパラメータ値の変更などが可能である。

また、アプリケーション側の対応として、たとえば MPICH-G2 では、 v1.2.2.3 以降の機能として、環境変数 MPICH_GLOBUS2_TCP_BUFFER_SIZE により TCP のバッファサイズが指定できるようになっている。さらに、 v1.2.5.1 では、バッファサイズなどがチューニングできない状況を想定して、２点間の通信に並列に複数の TCP コネクションを張ることで性能を向上させるための拡張も提供されている。

しかし、これらの値をどのように設定すると最も高性能が得られるかは、実際のネットワーク環境や通信パターンにも依存する。したがって、７センター間グリッドのように、複数の Gigabit Ethernet のインターフェースを持つスーパーコンピュータからなる各サイトが 10Gbps 級の Super SINET の回線を用い、数ミリ秒の遅延で接続される環境において、様々な通信パターン毎の最適なパラメータを調べることが、グリッドの本格利用のまえに検討し、チューニングすることが重要である。

また、このような目的に簡単に利用できる性能評価、チューニングのソフトウエアやツールを整備することも重要である。

5.5 NGB を用いた計算性能の評価

NGB(NAS Grid Benchmarks) は、グリッド環境における性能評価のためのベンチマーク定義 [7] である。 NGB は、並列計算機の性能評価のためのベンチマークとして著名な NPB(NAS Parallel Benchmarks) を開発した NASA で設計され、現在 NGB3.0 βが公開されている [7]。

NGB は、 NPB において定義された BT(Block Tridiagonal simulated CFD application)、 SP(Scalar Pentadiagonal simulated CFD application)、 LU(LU factorization)、 MG(Multi Grid) および FT(Fast Fourier Transform) を各サイトで実行されるタスクとし、それらの間のデータの受け渡しの関係をデータフローグラフ (DFG) により定義している。

ED(Embarrassingly Distributed) は、図 6 に示されるように、複数の SP の問題が相互に通信の必要なく異なるパラメータで独立に解かれるもので、 NPB における EP(Embarrassingly Parallel) と同様である。

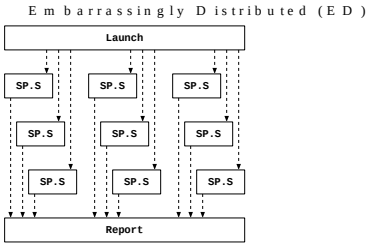

図 6. ED の DFG

HC(Helical Chain) は、図 7 に示すように BT → SP → LU の順でタスク間でデータが受け渡されこれが繰り返される。

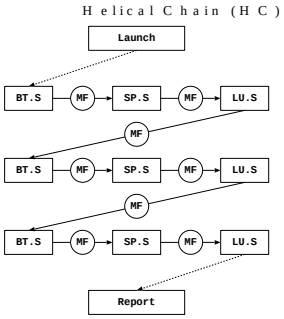

図 7. HC の DFG

VP(Visualization Pipe) は、計算途中で可視化する際に必要とされる計算パターンを想定して設計されたもので、BT を主ソルバ、MG をポスト処理、FT を可視化処理に模して、図 8に示すような DFG としている。

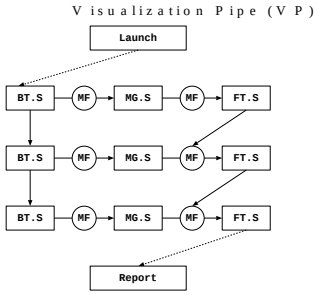

図 8. VP の DFG

MB(Mixed Bag) は、VP と同じ考え方で可視化のプロセスを模しているが、図 9に示すように、タスク間で受け渡されるデータの量や各タスクの計算負荷が非対称である点が特徴である。すなわちグリッド環境においてスケジューラがタスクと DFG を実際の計算資源に効率よく割り当てるのを困難にしている。

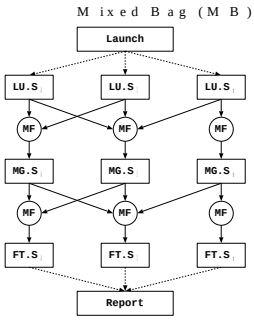

図 9. MB の DFG

NGB は、「紙と鉛筆」式のベンチマーク定義であり、アルゴリズムやプログラミング言語、ツールキットなどについての定めはない。

NGB 3.0 βには、参考の実装として、各タスクの Fortran による (非並列の) 実装と、DFG に沿ったタスク間のデータの受け渡しをファイルを介して逐次的に行うシェルスクリプトが添付されている。ベンチマークの評価においては、出力される結果の値の正当性と全体のターンアラウンド時間のみが考慮される。

NGB は、公開されて間もないため、まだ実際のグリッド環境におけるベンチマークレポートは集成されていないが、NPB と同様に今後続々と報告が集まり標準的なベンチマークとなることが期待される。

7センター間グリッドにおいても、NGB のような客観的なベンチマークスィートによる性能評価に早期に取り組み、その高い能力を数値で、世界に示せるようにすることが必要であると考える。

6 研究計画と推進体制

6.1 研究計画について

グリッドコンピューティング実現する上では、これまでに述べたネットワークやミドルウェアなどの基盤部分での調査研究に加えて、アプリケーションなど利用技術での調査研究を進めることも肝要である。したがって、本センターでは、導入したグリッド研究用システムおよび7センター間グリッドとも連携し、つぎのような研究計画を予定している。

6.1.1 計算グリッド

本センター研究開発部コンピューティング部門の教官を中心に、学内外の研究者に広く共同研究を募り、以下のような研究開発を計画している。

1) 並列プログラミングに関する研究

グリッド環境での並列プログラミング手法に関する調査研究

2) 並列プログラムの処理効率に関する研究

グリッド環境での並列処理および最適なジョブ割付け法に関わる調査研究

3) 並列ライブラリに関する研究

グリッド環境での大規模行列計算に関するアルゴリズムの検討とライブラリの開発に関わる調査研究

6.1.2 可視化グリッド

本センターは、IT プログラム「スーパーコンピュータネットワーク上でのリアルタイム実験環境」(代表者 北陸先端科学技術大学院大学 松澤 照男 教授) [8] に研究協力機関として参画し、「協調ビジュアルデータマイニングのテレイマージョン環境の構築」を担当している。このプロジェクトとも連携して、可視化グリッドに関する研究開発を計画している。

6.1.3 アプリケーション

大規模計算を必要とする研究者と共同で、実アプリケーションの立場からグリッドコンピューティングの利用法および有効性についての調査研究を行う。予定している共同研究分野は、つぎのものである。

- 核燃焼プラズマに関する共同研究

  参加研究組織等:
  京都大学、九州大学、山口大学他

- フィジオーム[1]

  参加研究組織等:
  医学研究科、情報学研究科、薬学研究科、
  再生医科学研究所

6.2 共同研究推進体制について

本センターでは、学内、学外の研究者と広く交流し共同研究を推進するために研究委員会を設置し、つぎの専門委員会を発足させている。

- グリッド研究専門委員会

  委員長 金澤 正憲 教授

- ビジュアライゼーション研究委員会

  委員長 小山田 耕二 助教授

[1] 医学生物学と計算機科学の力を駆使して分子から個体に至る各階層の明らかにし、ダイナミックな生体機能を時間軸および空間軸に沿って統合的に解明する新しい学問分野

7 まとめ

以上、京都大学学術情報メディアセンターにおけるグリッド研究の現状および今後の研究計画について報告した。

7センター間グリッドのためには、今後の計画としてあげたネットワークなど基盤部分での実証実験をはじめ多くの課題があるが、実現に向けて他センターとも協調して研究を進めたい。

参考文献

[1] Globus Project Home Page, http://www.globus.org/

[2] MPICH-G2 Home Page, http://www3.niu.edu/mpi/

[3] Netperf Home Page, http://www.netperf.org/netperf/NetperfPage.html

[4] S. Floyd, H. Henderson, The NewReno Modification to TCP's Fast Recovery Algorithm, IETF RFC 2582, Apr. 1999.

[5] V. Jacobson, R. Branden, D. Borman, TCP Extensions for High Performance, IETF RFC 1323, May 1992.

[6] NAS Parallel Benchmarks Home Page, http://www.nas.nasa.gov/Software/NPB/

[7] R. F. Van der Wijngaart, M. Frumkin, NAS Grid Benchmarks Version 1.0, NASA Technical Report NAS-02-005, July 2002.

[8] VizGrid Home Page, http://www.vizgrid.org/