2005－QAI－14（4）
2005／1／19

Gridにおける大量データ送信に適した品質保証方式

野呂正明† 長谷川一郎† 馬場健一‡ 下條真司‡

†〒567-0047 大阪府茨木市美穂ヶ丘5-1 大阪大学サイバーメディアセンター
下條研究室内 NICT 大阪JGNⅡリサーチセンター
‡〒567-0047 大阪府茨木市美穂ヶ丘5-1 大阪大学サイバーメディアセンター
E-mail: †{noro, ichiro}@ais.cmc.osaka-u.ac.jp, ‡{shimojo, baba}@cmc.osaka-u.ac.jp

あらまし 近年、Gridで利用するアプリケーションのデータサイズは年々大規模化している。大規模なGridでは、分散計算の品質を確保するため、Gridの広域網に対して専用回線や帯域を確保したMPLS網を利用しているが、大きなコストを要する。本報告では、比較的安価に利用できるDiffservのAFサービスを利用して、フロー単位で品質を確保しながら、契約帯域を動的に変更することにより、計算資源の利用効率の向上と公平性の確保を実現するDataGrid向けのQoS制御方式について報告する。

キーワード QoS, Diffserv, Grid, データ転送

QoS Control Method for Large Scale DataGrid Applications

Masaaki Noro† Ichiro Hasegawa† Ken-ichi Baba‡ Shinji Shimojo‡

† Osaka JGNII Research Center , NICT 5-1 Mihogaoka, Ibaraki, Osaka, 567-0047 Japan
‡ Cybermedia Center , Osaka University 5-1 Mihogaoka, Ibaraki, Osaka, 567-0047 Japan
E-mail: †{noro,ichiro}@ais.cmc.osaka-u.ac.jp, ‡{shimojo, baba}@cmc.osaka-u.ac.jp

Abstract Data size of Grid has recently increased enormously. Large scale Grid environment (like OptIPuter etc.) plans to use "dark fiber" or VPN path with bandwidth reservation technique for guaranteeing "Quality of Computation". But, Cost of this approach is very high. So, We are studying about QoS method for large scale grid environment which guarantee quality of data transmission with reasonable cost. This method reserves bandwidth for each data flow (like GridFTP) using AF service of Diffserv environment. And, user application regularly adjusts own reserved bandwidth based on throughput of their data transfer. This method achieves well-balanced performance of fairness and throughput. In this paper, we discuss this method and evaluation using "NS2".

Keyword QoS, Diffserv, Grid

1. はじめに

近年の傾向として、Gridにおける処理データの大容量化が進行し、1つの組織で準備できる計算能力を超える処理が増加している。これに対応するため、Gridにおいては仮想組織(VO;Virtual Organization)を構成するが、これらの組織間を広域網を介して接続して、計算データを転送する。従って、転送性能が計算性能に大きく影響する。

特に、観測装置から出力されるデータを処理する場合、データを発生させる速度以上の性能で計算を実施する必要があり、各処理に時間的な制約が存在する。例えば、テンサイエンスでは、高精細の電子顕微鏡による試料のスライスデータや、MRIによる人体の断層撮影の画像データ等が観測装置等から発生し、そのデータを実時間で処理することが必要である。電子顕微鏡による観測の場合、1つの試料あたり、数十Mbyte程度の画像データを100以上発生させる。このような要求に対応するためには、データ転送の品質を広域ネットワークにおいて保証することが非常に重要である。

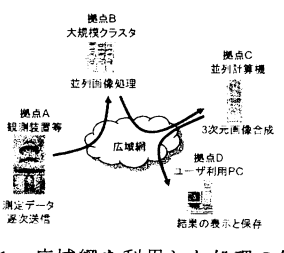

図 1：広域網を利用した処理の例

2. 従来技術の問題点と研究の目標

現在、VO 環境でデータ転送の品質を確保するために利用される方法は以下の二つに分類できる。

2.1. 広帯域な専用ネットワークの利用

OptIPuter[17]等では VO を構成するサイト間の接続にダークファイバ等、大容量かつ占有可能な経路を準備することにより、通信品質の問題を軽減している。この方法は非常に高価であり、一般的には適用は困難である。

2.2. ネットワーク資源の予約

フローに対して帯域を割り当て、必要な性能を確保する方法がある。要求された性能を確保するため、フローに対して固定された帯域を確保する方法や、Diffserv の最低帯域保証（AF サービス）を利用して、品質を保証する方法がある。

2.2.1. 最低帯域保証による品質確保

最低帯域を保証する方式として、Diffserv の AF による方法を説明する。図 2のように、Diffserv では、対象ネットワークを DS ドメインと呼ぶ。DS ドメインと他のネットワークの境界となる DS エッジにおいて事前に設定された情報に基づき、流入する IP パケットのヘッダの情報に従って、パケットの DSCP (Diffserv Code Point)に値を書き込み、DS ドメイン内に転送する。

DS ドメイン内部の装置（コアデバイス）では、各 IP パケットの DSCP の値に応じて各パケットに対する操作を差別化することで品質を制御する。

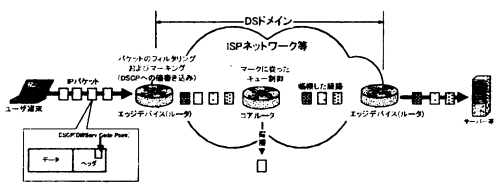

図 2：Diffserv 概念

AF では、図 3に示すように、ユーザとの間で保証する帯域を事前に契約する。次に、流入するパケットの流入量をエッジルータにおいて監視し、契約帯域内のパケットは Green に、契約帯域を上回るパケットを Red にマーキングする。DS ドメイン内部のコアルータは、輻輳状態になると Red から先に破棄し、Green を極力廃棄しないことで、ユーザに対して最低帯域を保証する。

実際には、帯域の契約段階において、Green のパケットが破棄されないように、DS ドメイン内のルータ間のリンク帯域設計を行っておく必要がある。

さらに、図 3のように、RIO と呼ばれるルータのキュー管理技術では、実際に輻輳状態（ルータの出力バッファのキュー長が規定を超える）となる前に、Red を先行して破棄することで、TCP のフロー制御メカニズムを利用し、輻輳を抑制することが可能となっている。

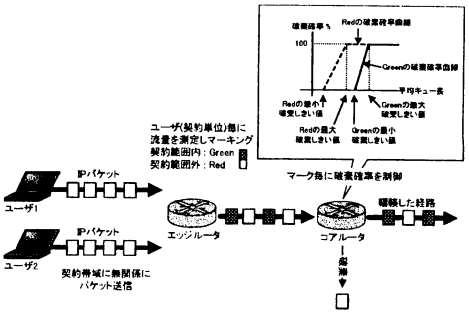

図 3：AF サービス

ここで、Grid のファイル転送に対して、フロー単位に AF で最低帯域保障を実施した場合を想定する。ネットワークの利用率が低い場合には、フロー単位に帯域を制御しない場合と同様に、次のような公平性に関する問題がある。

時間差を伴って同一の契約帯域を持つ、2 つの TCP フローが発生したと仮定する。この場合、先行して発生したフローが余っている帯域の大部分を利用してしまい、要求した帯域に対して実際に得られるスループットは非常に偏ったものとなってしまう。

2.2.2. 固定帯域割り当てによる品質確保

各フローに要求帯域を固定的に割り当てる場合について説明する。図 4のように、各フローに対してバッファを割り当てると共に、パケットのスケジューラに対して設定を行い、各フローに対して契約した値のスループットを提供する。

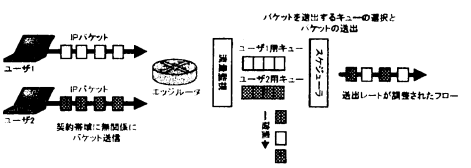

図 4：固定帯域割り当て

この方式で Grid のファイル転送に帯域制御を実施する場合、ルータにおいて、制御対象となるフロー数だけバッファを準備すると共に、バッファの長さ等のパラメータ設定を頻繁に変更する必要があり、ルータの負荷が非常に高くなる。

また、契約帯域以上のスループットを提供しないため、公平性は確保されるものの、未利用となる帯域が発生し、資源の利用効率はあまり良くない。

そのため、図 5のように、帯域に余裕のある場合に空き帯域の利用で、新規フローの要求を満たすことが

可能な場合でも、帯域を固定で割り当てているために、新規フローの帯域獲得の要求が呼損となる場合がある。

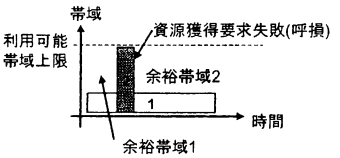

図 5：帯域の無駄による呼損の例

2.3. 研究の目標

我々の研究グループでは、帯域の利用率の向上と、公平性を両立させるため、ファイル転送のフローに対して AF サービスで品質を保証する方式について研究を行っている。

3. Grid におけるファイル転送と品質

まず、Grid におけるファイル転送に対して実際に提供すべき品質とファイル転送間の公平性について検討する。Grid では新規のファイル転送の発生時の品質確保要求が拒絶された場合、そのファイル転送（呼）のオーナーである Job が失敗するとは限らず、他の計算資源の割り当てや、失敗した呼が属する Job のスケジュールを Grid のスケジューラと連携して修正することで Job の失敗を回避することも可能である。しかし、Grid のスケジューラとネットワークの QoS 機構の連携による全体での効率等の評価は本報告では対象としない。

一般的に、ネットワークで提供する品質の指定方式としては、ジッタ、遅延、パケットの損失率の 3 つの指標がある。ただし、専用の Grid ネットワークを構成する場合と違い、一般の ISP の環境を利用した場合は経路を指定することができないため、品質の契約時は帯域を指定することとなる。

ここで、あるファイル転送に対して帯域を指定した場合、利用プロトコルとファイルのサイズが既知であるため、帯域の指定は転送の終了予定（希望）時間を指定するのとほぼ同義である。さらに、ファイル転送は非インタラクティブトラフィックであるため、各フローに対して保証するスループットを契約した帯域をフローの生存期間中、常に一定にする必要はない。実際に保証する品質は、フローの生存期間の平均で契約した条件を満たしていれば良いと考えられる。

公平性は同一契約帯域を持つ複数フローの実効スループットで議論する必要がある。2.2.1のように、個別フロー毎に帯域を契約する方式では、同一条件下でも、フローの発生するタイミングにより、得られる平均スループットが異なる。そのため、より多く帯域を利用したフローのスループットを規制して他のフローに分配できる方式が必要となる。

4. 提案方式

帯域の利用効率と公平性の両立のため、Grid の個別のフローに対して AF で最低帯域を保証しつつ、呼の開始時に契約した帯域をファイルの転送が進むにつれて、見直す方式を提案する。

4.1. 提案方式概要

本方式は図 6に示すように帯域を利用し、定期的に必要帯域を見直して、契約帯域を徐々に削減する。

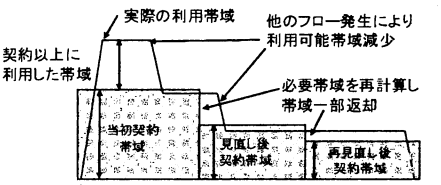

図 6：提案方式の帯域利用法

図 7は本提案方式の実現例である。ファイル転送の発生時に、送信側端末はルータを管理するサーバに要求帯域を通知し、帯域を契約する。サーバはルータの設定を変更し、フローに最低帯域を保証する。また、個々のパケットに対してはエッジルータにおいて、契約帯域内かどうかでマーキングを実施する。

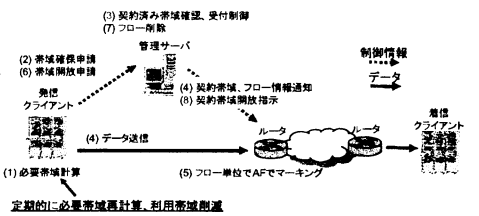

図 7：提案方式の実現例

本方式では、AF を用いたフローに対する最低帯域保証に加えて、予定以上にスループットを得られた場合に、契約帯域を削減していくことで、以下のような効果を得ることを目指している。

- 個別のフロー単位で最低帯域保障を実施することにより、経路の利用効率を向上させる。
- 必要帯域の削減により、空き帯域を創出し、新たな呼の受理可能性を増加させる。
- 送信端末で送信済みデータ量を監視し、定期的に必要帯域を見直すことで、呼（フロー）の間のスループットの不公平性を軽減する。

4.2. 帯域の見直し手法

提案方式では、想定した帯域以上にスループットが

得られた場合にのみ、契約帯域の見直しを実施する。そのため、契約帯域の増加はないので、各フローがいったん契約した帯域はフローの終了まで維持されるか、減少するかしかない。よって、帯域要求の失敗による呼損はフロー発生時しか起こりえない。すると、帯域の再契約処理は残りデータを転送しながら実施してもかまわないため、帯域見直しに伴うデータ転送の中断は発生しない。

なお、本提案手法では、事前に定められたブロックサイズ分だけファイル転送が終了した時点で必要帯域の見直しを実施する。

4.3. スループットの予測と契約帯域の算出方法

契約帯域の見直しには、要求した帯域に対してどの程度の性能が得られたかを計算する必要がある。TCPではフローに対してある帯域を契約していても、その性能が得られない。そのため、契約帯域に対して期待できるスループットを計算し、その値に基づいて必要帯域の見直しを行う。

4.3.1. 契約帯域とスループットの関係

まず、輻輳回避フェーズにおけるTCPの性能について議論する。TCPの挙動は単純化すると、パケットの破棄により、ウィンドウサイズが1/2となり、そこから1RTT時間毎にウィンドウサイズが1パケットサイズ分だけ増加するという動作を繰り返す。(図 8)

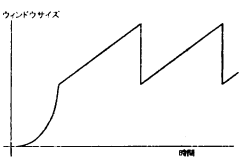

図 8：TCPのウィンドウサイズの変化

次に、各種の変数を以下のように定義する。

T ：経路のRTT （秒）

P ：パケットサイズ(byte)

B ：契約帯域(bps)

W_{max} ：可能な最大ウィンドウサイズ(byte)

W_{min} ：小さくなったウィンドウサイズ(byte)

また、パケットロスによるウィンドウサイズの減少から、元の大きさまで復帰するまでの1周期分の様子を図 9に示す。

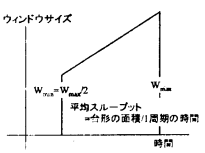

図 9：ウィンドウサイズ変化（1周期分）

すると、ウィンドウサイズの回復に必要な時間は式1で表すことができる。

$$(W_{max} - W_{min})T / P \qquad \text{式 1}$$

さらに、最大のウィンドウサイズ、最小のウィンドウサイズの関係は式2であるので、この区間での平均スループットは式3となる。

$$W_{min} = W_{max} / 2 \qquad \text{式 2}$$

$$S = 8(W_{max} + W_{min}) / T = 6W_{max} / T \ \text{(bps)} \qquad \text{式 3}$$

W_{max}(byte)のデータがT （秒）の間に、経路を流れて、それが可能な最大帯域となることから、契約帯域BとW_{max}の関係は式4で計算できる。

$$W_{max} = \frac{BT}{8} \ \text{(byte)} \qquad \text{式 4}$$

すると、スループットは以下の式5で与えられる。

$$S = 3B / 4 \qquad \text{式 5}$$

しかし、実際には、TCPのヘッダや再送、さらにはスロースタートの影響や、契約帯域を越えてもすぐにパケットは破棄されないため、アプリケーションレベルのスループットはこれを多少前後した値となる。

4.3.2. 契約帯域の見直し

提案方式では、以上の式を利用して契約済みの帯域と転送に成功したデータ量を比較した上で契約済み帯域を削減可能かどうか判断する。

時刻tにおいて契約済み帯域が削減可能かどうかを初期の契約帯域BとファイルサイズF、現在の契約帯域B_tに基づいて判断をする。

まず、BとFおよびTから予測されるスループットS_0を計算する。次に、S_0からファイル転送の予想終了期限t_{lim}を計算して記憶しておく。

ここで、時刻tにおいて転送しなければならない残りのファイル容量をF_tとすると、F_tと$t_{lim}-t$から必要となる最低のスループットS_tを計算する。このS_tを用いて、必要となる契約帯域B_t'を計算し、B_tとB_t'を比較して、再契約が可能かどうか決定する。

5. 評価

定量評価はNS2によるシミュレーションで実施するが、現在のところは以下の3つの評価を予定している。

- 各フローの初期契約帯域に対して得られた実際のスループットの分散 （公平性）
- 単位時間あたりの、実際に送信が終了したデータ量 （性能）
- 単位時間あたりの新規に発生したフロー（呼）の帯域確保要求が失敗する確率 （呼損率）

さらに、提案方式と比較する対象として以下の3つの方式を予定している。

- 品質制御、呼制御を実施しない方式
- 個別フローに最低帯域保障のみを実施する方式
- 個別フローに固定の契約帯域を割り当てる方式

5.1. 評価モデル

図 10と図 11は評価モデルである。評価モデルには、Gridの広域網を構成するネットワークとファイル転送のフローの性質の条件の2つがあるが、ネットワークに関しては、JGNⅡを想定し、アクセス網、基幹網共にギガビットレベルのネットワークを設定している。

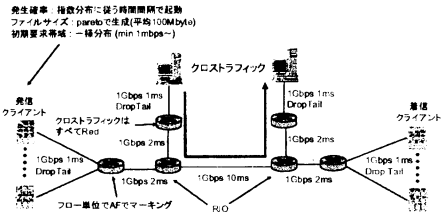

図 10：AF（最低帯域保障）方式の評価モデル

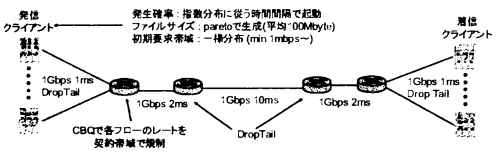

図 11：従来方式の評価モデル

また、発生するファイル転送に対しては以下のような性質を想定している。

フローの発生確率

多数のユーザが同時に利用する環境を想定した場合、各ユーザは独立にJobを投入するものと考えることができるため、現在はフローの発生時間間隔を指数分布としている。

ファイルサイズ

テレサイエンス分野の画像処理アプリケーションでは、処理対象の画像サイズは1枚あたり数十Mbyte、場合によっては1画像あたり100Mbyteを越える。観測器装置の品質や測定対象の形状により、データ容量は異なるものの、本評価では平均100Mbyteの Pareto分布を利用し、10Mbyteから100Mbyte程度のファイルを多数発生させ、かつ極少量ながらGbyteを超えるようなファイルの転送を発生させている。

図 12に実際に発生した各フローのファイルサイズの分布の例を示す。このグラフを見てもわかるように、20Mbyte前後に多くのファイルが集中しており、最大では26Gbyteのファイルも存在し、全体として平均は94Mbyte程度となっている。

ファイル転送が要求する初期帯域

今回の評価モデルでは、ルータの処理遅延等を除いた純粋な経路のRTTは32msとした。RTTの値が32msで、経路の物理的要因によるパケットの損失確率を10^{-12}と想定した場合の1本のTCPコネクションが出しうる最高性能は約5Mbps程度である。

そのため、今回の評価では呼の発生時に、2Mbps～5Mbpsの間の一様分布の乱数を発生させ、各フローが要求する初期の要求帯域として利用している。

その他の項目

要求帯域の見直しのタイミングを定めるファイル転送のブロックサイズは20Mbyteとしている。さらに、ファイル転送にはftpを想定し、1つのTCPフローでデータを転送するとともに、ウィンドウサイズ等は標準のTCPの値を利用している。

ただし、大規模なGrid環境では、広帯域かつ、遅延の大きなネットワーク（帯域と遅延の積の大きなネットワーク）でデータ転送の性能を引き出すために、TCPのウィンドウサイズを変更したり、データ転送に複数のTCPフローを利用したりする対策も行われている。

そのため、標準的なTCPによるシミュレーションの他にウィンドウサイズを変化させたTCPを利用した場合や、複数のTCPを利用した場合の比較や、転送するファイルおよびデータ転送の1ブロックのサイズを変化させた場合の評価についても現在検討している。

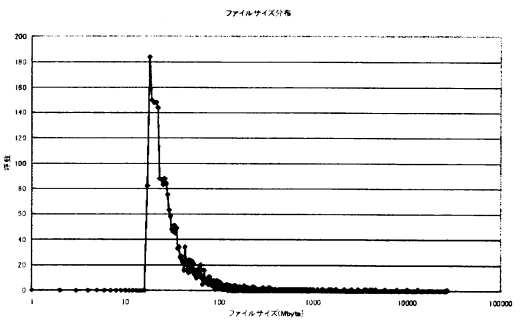

図 12：ファイルサイズ分布

5.2. 評価状況

表 1に現在の評価データを示す。データ収集時間は約2時間分であり、表中のそれぞれの項目は以下のような値である。

平均フロー数 ：ファイル転送中のフロー数の平均

平均契約帯域 ：全体の帯域に対して契約済み帯域の比率（平均値）

総データ転送量 ：測定時間内に発生し、かつ同時間内に終了したファイル転送の転送データ量の総和

公平性 ：当初の契約帯域に対して、実際に得られたスループットの比率の分散

呼損率 ：ファイル転送（呼）のうち、ネットワークの帯域不足により、帯域契約に失敗した呼の率

発生確率 (session/分)	方式	平均フロー数	平均予約帯域 (%)	総データ転送量 (Gbyte)	公平性		呼損率
					平均	分散	
48	提案方式	96.19	24.59	391.10	1.448	0.150315	0.00
	帯域固定	157.13	50.25	376.42	0.878	0.000039	0.00
	AFのみ	105.22	35.67	428.97	1.452	0.155627	0.00

表 1：評価データ

現在は収集データ量が少ないため、正確な判断はできないものの、実際に送信できたデータ量も公平性も他の従来方式の中間値となっているが、AF だけの方式に対して得られた性能が低いのと、公平性の改善がごく少ない値となってしまっている。

これは、平均の契約帯域を見てわかるように、Green となるパケットが Red のパケットに対して少ないため、DS ドメイン内部での RIO の設定を注意深く行わないと、Red のパケットロスのために、性能が出にくくなっていると予想できる。さらに、平均のフロー数が少ないことも影響していると思われる。

次に、公平性であるが、個別にフローに対して AF で最低帯域を保証した場合と比較してあまり差が無い。これは、負荷が低い場合は TCP の性能限界とフロー数の積に対して帯域が大きいため、差が出にくくなっていると考えられる。これについても、負荷率をさらに大きくして、データを収集する必要がある。

6. まとめ

大規模な DataGrid アプリケーションを想定した、ネットワークの品質保証方式および、実施中の定量評価の評価方法、評価モデルについて報告した。現在複数の従来方式と提案方式の定量的評価をすすめている。

今後は方式のチューニングを進めて、さらにデータ収集を大規模に実施して詳細な評価を実施する。

なお、具体的なアプリケーションの動作について詳細に把握するため、他の研究グループと共同で実際のアプリケーションの発生する通信の統計情報も実環境を用いて収集することも計画している。

文　献

[1] T. Li, Y. Rekhter, "A Provider Architecture for Differentiated Services and Traffic Engineering (PASTE)", RFC2430, October, 1998.

[2] K. Nichols, S. Blake, F. Baker, D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC2474, December, 1998.

[3] D. Grossman, "New Terminology and Clarifications for Diffserv", RFC3260, April, 2002.

[4] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, "An Architecture for Differentiated Service", RFC2475, December, 1998.

[5] J. Heinanen, F. Baker, W. Weiss, J. Wroclawski, "Assured Forwarding PHB Group", RFC2597, June, 1999.

[6] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC2205, September, 1997.

[7] R. Braden, L. Zhang, "Resource ReSerVation Protocol (RSVP) -- Version 1 Message Processing Rules", RFC2209, September, 1997.

[8] K. Kompella, J. Lang, "Procedures for Modifying the Resource reSerVation Protocol (RSVP)", RFC3936, October, 2004.

[9] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC3209, December, 2001.

[10] D. Awduche, A. Hannan, A. Xiao, "Applicability Statement for Extensions to RSVP for LSP-Tunnels", RFC3210, December, 2001.

[11] S. Floyd, V. Jacobson, "Link-sharing and Resource Management Models for Packet Networks", IEEE/ACM Transactions on Networking, Vol.3, No.4, pp. 365-386, August, 1995.

[12] 鶴正人, 熊副和美, 尾家祐二, "長距離高速通信のための TCP 性能改善技術の動向", 情報処理学会誌, vol.44, No.9, pp.951-957, September, 2003.

[13] S. Floyd, V. Jacobson, "Random Early Detection gateways for Congestion Avoidance", IEEE/ACM Transactions on Networking, Vol.1 No.4, pp. 397-413, August, 1993,.

[14] W. Feng, D. Kandlur, D. Saha, K. Shin, "Understanding and improving TCP performance over networks with minimum rate guarantees", IEEE/ACM Transactions on Networking, Vol.7, Issue.2, pp.173-187, April, 1999

[15] The Globus Alliance , "Grid FTP" http://www-fp.globus.org/datagrid/gridftp.html

[16] VINT project, "ns2", http://www.isi.edu/nsnam/ns/

[17] OptIPuter, http://www.optiputer.net/