4-2　広域分散計算における大規模ファイル転送向けデータ転送技術

4-2 QoS Control Method for Large-scale Data Transfer in DataGrid Applications

野呂正明　馬場健一　下條真司
NORO Masaaki, BABA Ken-ichi, and SHIMOJO Shinji

要旨

近年、Grid サービスに関する研究開発が盛んである。Grid サービスでは、Web インターフェースを使うことにより、ユーザが利用する際の敷居を下げることができる。現在、高性能な Grid では、広域網として高性能な専用ネットワークを利用するのが主流である。しかし、新規参入するユーザにとって、専用ネットワークはコストが高すぎることが問題である。

一方、処理するデータサイズは増加しており、計算の実行時にファイル転送がボトルネックとなる。パイプライン的にデータ処理を行うアプリケーションでは、性能確保のために、ファイル転送にネットワークの帯域を割り当てると、アプリケーションの並列度が経路の帯域で制約される。そこで、多くのユーザが同時利用する環境を前提として、Diffserv の最低帯域保証技術を利用し、データ転送のスループットを保証しながら、同時に実行できるデータ転送数を増加させる方式を提案した。

Many people works about Grid service technology recently. It is difficult to adopt private circuitas wide area link for many users in Grid service environment. It is helpful for scheduler of Grid that data transfer reserves bandwidth in pipeline processing applications. On the other hand, data size of Grid has increased enormously. Large-scale data transfer with bandwidth reservation should bebottleneck. Therefore, it is useful to increase numbers of data transfer within same network resource. We propose QoS control method for large-scale data transfer over Diffserv network.

[キーワード]
グリッド，ファイル転送，区分化サービス，品質制御
Grid, File transfer, Diffserv, QoS control

1　はじめに

近年、標準化団体[1]を中心に Grid サービスの研究開発が盛んである。Grid サービスでは、簡易な Web のインターフェースで利用できるため、新規ユーザの急速な増加が期待される。その反面、このような環境では、サービスコストが重要であり、OptIPuter[2] のような従来の Grid 環境が採用してきた高速専用ネットワークの利用は困難である。そのため、一般の IP ネットワークを利用し、計算サービスを実施するサイトの計算資源の利用効率の向上や、アプリケーションの実行効率を向上する技術が求められる。

一方、Grid の処理データは大容量化し、数百メガバイトの大量に処理するアプリケーションも存在する。この種のアプリケーションでは、広域ネットワークでのデータ転送が性能を左右する。大量データをパイプライン処理する場合は、計算中にバックグラウンドで次のデータを転送することが有効である。この種のアプリケーションで大量データを効率よく処理するには、ファイル転送の性能保証が必要である。しかし、このような場合は、帯域不足により、同時処理できるファイル数が制約される。この問題に対応するため、

Diffserv の最低帯域保証技術を利用してより多くのデータ転送を実現方式について提案した[3]。本稿では提案方式とシミュレーションソフト[4]による評価結果について述べる。

2 従来技術とその問題点

Grid において各ファイル転送の性能を保証する場合、帯域確保の失敗はジョブの再スケジューリングや、処理遅延の原因となる。そのため、各ファイル転送に対する帯域確保要求は極力受理することが望まれる。

2.1 帯域を固定的に割り当てる方法

Diffserv[5] の EF サービス[6] や CBQ[7] を利用し、契約帯域の範囲内でデータ転送を行う方式では契約帯域以上のスループットを得られないため、図 1 のような場合にフロー 2 の帯域獲得要求が失敗(呼損)する。しかし、GridFTP[8] のようにデータ転送のスループットを大きく取ることは珍しくないため、このような状況が発生する確率は高いと考えられる

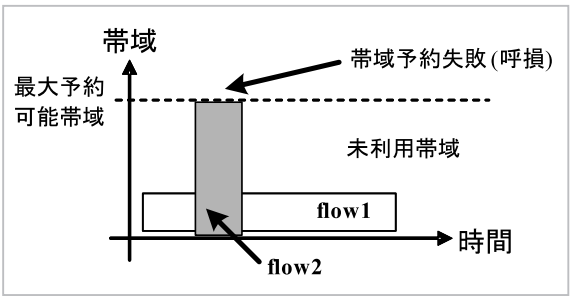


図1 バースト的な帯域要求による呼損

2.2 最低帯域保証による方法

次に各ファイル転送に最低帯域保証[9] を適用する方式を検討する。この方式では、ネットワークが輻輳していない場合に各セッションは契約以上にスループットを得ることができるため、固定した帯域を割り当てる方法と比べ、契約済みの帯域も減少し、呼損率は改善する。しかし、この場合でも図 1 のフロー 2 のような場合は呼損する可能性がある。

3 大規模ファイル転送のための QoS 制御方式

まず、分散計算環境におけるファイル転送に対して実際に提供すべき品質について検討する。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 の処理時間の増加もしくは Job を、よりネットワーク資源に余裕のある場所に再割付する必要があるなど、好ましくない結果をもたらす。

また、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三つの指標がある。ただし、専用のネットワークを構成する場合と異なり、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。

以上の議論から、Job と初期に契約した帯域を目標スループットととらえ、そのスループットを確保しつつ、全体の帯域を有効に利用することで、新たな帯域確保の要求をより多く受け入れることができれば、CPU も含めた全体の計算資源を有効に活用することが可能と考えられる。

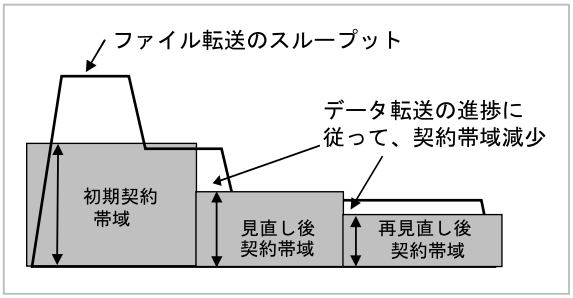


図2 提案方式におけるデータ転送

3.1 提案方式

Diffserv の最低帯域保証サービス(AF サービス)を利用し、個別のフローのスループットを保証しつつ、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。本方式では、以下のような効果を得ることを目指している。

- 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。

図 2 は本方式におけるデータ転送の TCP フローのスループットと契約帯域の関係を示している。本方式では、ユーザから指定されたスループットを達成するのに必要な帯域を計算し、その値を用いて最低帯域保証サービスで契約を実施する。この時、契約帯域と同時に TCP の性質を利用して終了予定時刻を計算しておく。次に、固定長のデータを転送した後、転送残りデータ量と終了予定時間までの残り時間から再度契約すべき帯域量を算出する。

もし、ネットワークが輻輳していなければ、データ転送のスループットは目標スループットを上回っているので、契約すべき帯域量は現在の契約値より小さくなる。契約すべき帯域量が現在の契約を下回る場合のみ、契約を変更して空き帯域を増加させる。なお、極端にネットワークが輻輳している等の理由でデータ転送のスループットが目標を下回った場合でも、契約帯域を増加させない。これは、データ転送中に契約帯域を広げようとした場合の予約の失敗を避けるためである。

本方式では、最低帯域保証サービスを利用していることと、契約の変更は減少のみしか行わないという特徴のため、各端末はデータ転送中に帯域契約のシグナリングのためデータ転送を中断したり、他の端末との同期をとったりする必要がまったくない。そのため、実装が容易になるとともに、スケーラビリティ確保が比較的簡単にできる。

3.2 目標スループットからの契約帯域の算出

提案方式では、達成すべきスループットから契約すべき帯域の値やデータ転送の終了予定時刻を算出する必要がある。そのため、AF 環境における契約帯域と TCP のスループットについて議論する(詳しい議論は[10]参照)。

TCP はパケットロスを検知すると、ウィンドウサイズを 1/2 とし、その後、1RTT ごとにウィンドウサイズを 1 パケットサイズ分だけ増加させる。そのため、パケットの連続ロスや契約帯域内のパケットの破棄がない限り、図 3 のようにウィンドウサイズが変化する。DiffservAF では、マーキングはスループットの一定時間の平均値に基づくため、スループットが契約帯域を超えてもすぐには RED にならない。そのため、TCP のスループットのピーク(パケットロスの発生時)は契約帯域と同等以上となる。すると、TCP が輻輳回避フェーズで動作しており、AF サービスにおいて契約が守られている環境での TCP のスループット S と契約帯域 B の関係は $S \geq 3B/4$ で表すことができる。

次に、この関係が成立しない場合を検討する。これは、目標スループットに対して、ファイルサイズが小さい場合と、ファイルサイズが大きくても、パケットの連続ロスにより、TCP が輻輳回避フェーズからスロースタートフェーズに頻繁に移行する場合に発生する。しかし、本方式のターゲットアプリケーションでは数百メガバイトやそれ以上のサイズのファイルを多数転送するため、このような状況はまれにしか発生しないと想定しており、現在のアルゴリズムでは契約帯域は目標スループットの 4/3 倍としている。

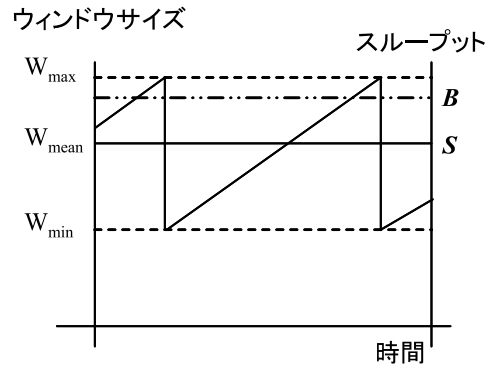


図3 輻輳回避フェーズでの TCP のスループットと契約帯域の関係

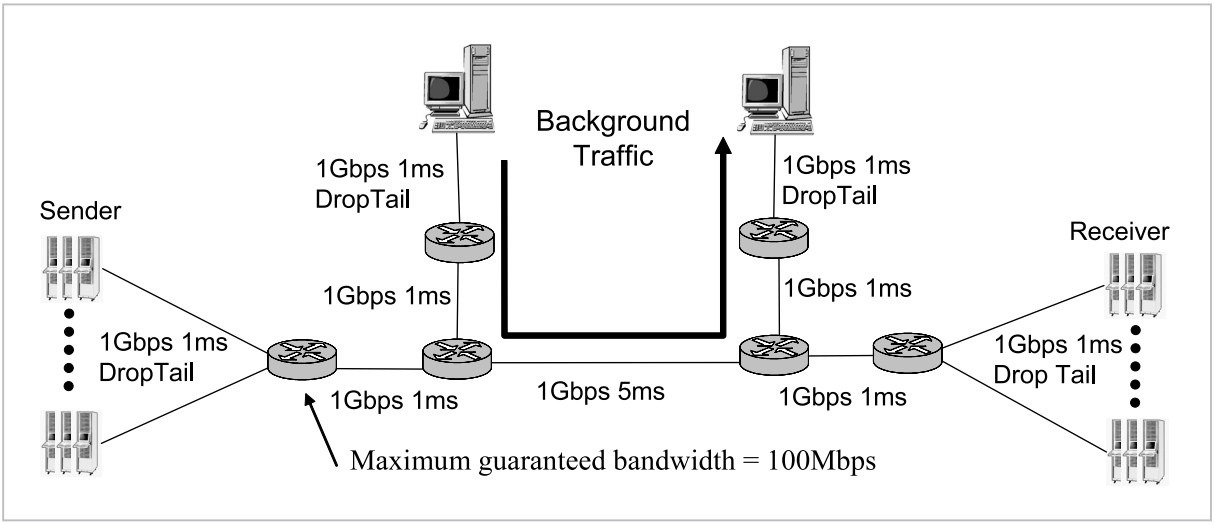


図4 シミュレーションモデル

4 評価

NS2 によるシミュレーションにより、以下の指標で提案方式の性能を評価した。

(1) 呼損率：新規に発生したファイル転送（セッション）の帯域確保要求が失敗する確率
(2) スループット：単位時間当たりの、全フローで実際に送信できたデータ量
(3) expired flow：フロー開始時に設定した目標スループットを達成できなかった確率

なお、提案方式の比較対象として以下の二つの方式がある。

(A) Simple EF：ファイル転送ごとに帯域を契約し、各ファイル転送は契約帯域を超えないよう、エッジにおいてシェーピングを施してファイルの転送を実施する方式。
(B) Simple AF：個別のファイル転送ごとに Diffserv の AF で必要帯域を契約し、帯域の契約を転送終了まで変更なしに保持し続ける方式。

4.1 評価モデル

図 4 に本稿で用いた評価モデルを示す。なお、各種パラメータは以下のとおりである。

(1) ファイル転送の発生確率：ポアソン過程
(2) ファイルサイズ：平均 100Mbyte の Pareto 分布
(3) ファイル転送が要求する初期帯域：1Mbps から最大スループットの間の一様分布
(4) その他の項目：各ファイル転送は 20Mbyte 送信するたびに帯域の見直しを実施
(5) バックグラウンドトラフィック：ない場合と、最大 1Gbps の Pareto 分布の 2 通り

4.2 シナリオ 1

ここでは、ファイル転送に TCP Reno を利用したシナリオで評価を実施している。なお、各グラフの横軸はファイル転送のフローの発生確率である。

(1) 呼損率

図 5、図 6 はそれぞれバックグラウンドトラフィックがない場合とある場合の呼損率のグラフである。提案方式は、負荷が高くなった場合でも、各フローが得られるスループットに応じて利用帯域を削減するため、新規のフローが受理され、呼損率が低くなる。

(2) スループット

バックグラウンドトラフィックのない場合とある場合の総スループットは 図 7、図 8 である。提案方式はどちらも他の方式より良い性能である。これは、高負荷でも呼損率が低いため、より多くのフローを受け付けることで、より多くのデータを転送するためである。

(3) Expired flow

フローの要求したスループットを守れなかった比率であるが、表 1 のようにすべての方式がファイル転送の発生確率によらず、ほぼ一定の値を示した。

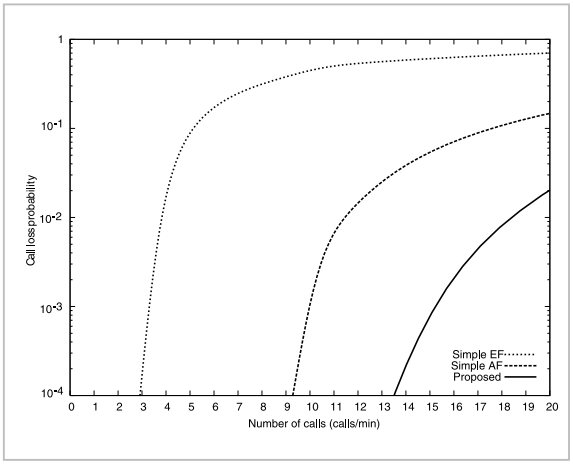


図5　呼損率（無し）

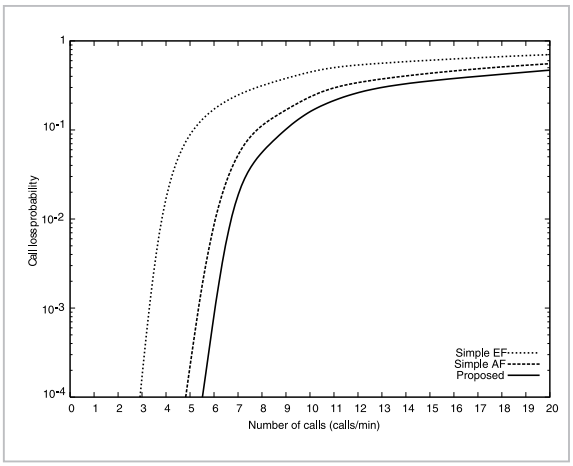


図6　呼損率（Pareto 分布）

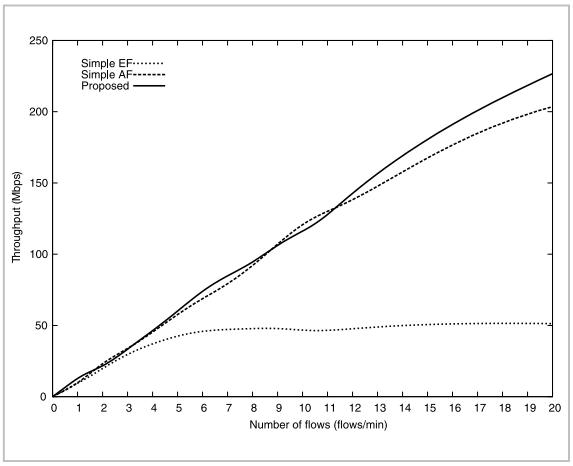


図7　総スループット（無し）

Simple EF では常に 1 %程度契約が守られない。これは、要求帯域に対してファイルサイズが小さいためである。これに対し、提案方式と Simple AF はバックグラウンドトラフィックの有無にかかわらず、スループットが目標を上回り、Simple EF より良い結果を示している。

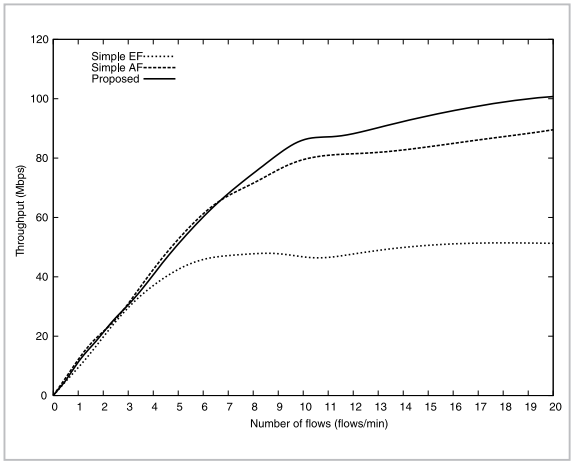


図8　総スループット（Pareto 分布）

表1　スループットが要求を下回った比率

	バックグラウンドトラフィック		
方式	なし	Pareto分布	CBR
Simple EF	約1%	約1%	約1%
Simple AF	0%	約0.5%	約0.5%
提案方式	0%	約0.5%	約0.5%

4.3　シナリオ 2

この評価では、GridFTP 等を用いた場合を想定し、TCP のウィンドウサイズを調整してスループットが 2 倍、3 倍となる状態を作り出したほかは、シナリオ 1 と同じである。図 9、図 10 が呼損率、図 11、図 12 が総スループットである。どちらも、スループットが 3 倍の場合に、提案方式と Simple AF の差がほとんどない。これは、TCP のスループットと予約帯域が大きいため、系に存在する TCP の数が少ないためである。少数の TCP で大量のパケットを発生すると、破棄確率が上がり、結果としてスループットが低下することで、提案方式の効果がなくなっていく。

次に、実際のスループットが要求を下回った確率を表 2 に示す。スループット 2 倍、3 倍のどちらの場合でも、提案方式は他の方式と同等以上の性能を示している。

以上の結果から、データ転送に TCP を用いている場合、ボトルネックリンクに同時に存在する

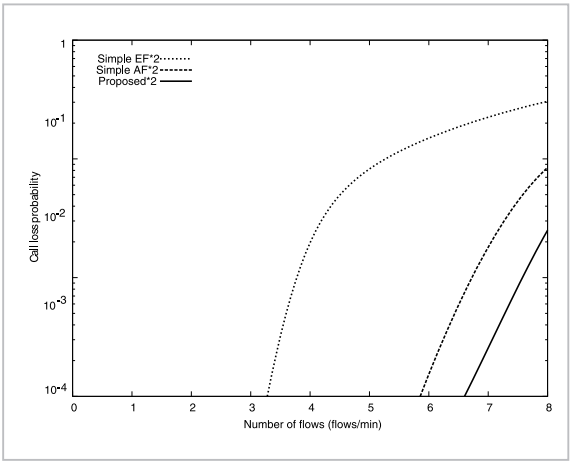


図9 呼損率（スループット２倍）

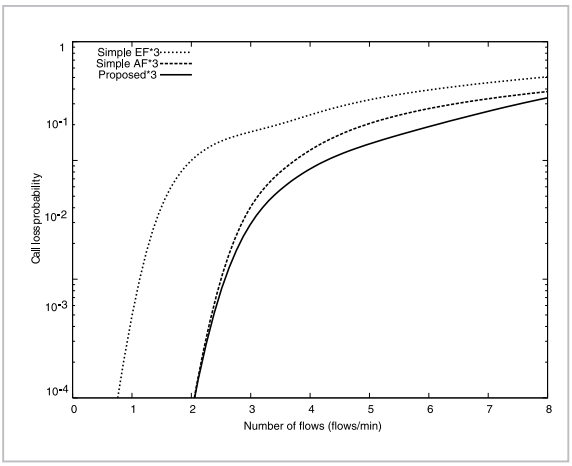


図10 呼損率（スループット３倍）

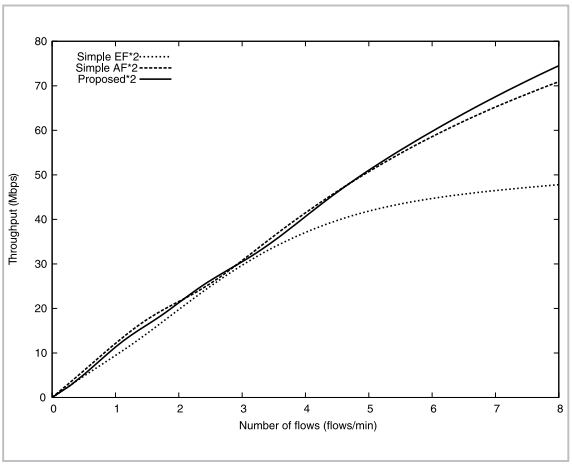


図11 総スループット（スループット２倍）

TCP フローの本数が少ない環境ではデータ転送のスループットをより大きくできる方式を組み合わせる必要があることが分かった。

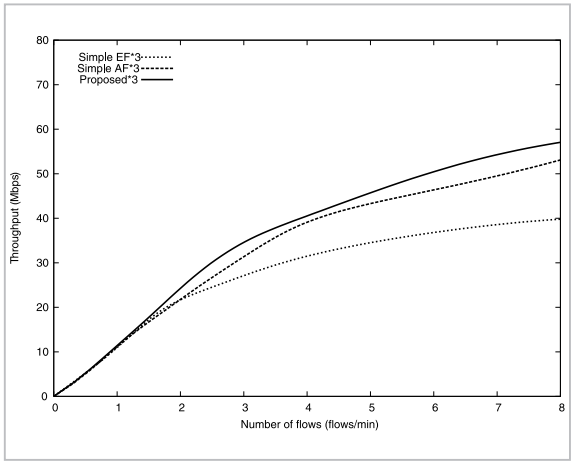


図12 総スループット：（スループット３倍）

表2 スループットが要求を下回った比率

	フロー単位のスループット	
方式	2倍	3倍
Simple EF	約1%	約1%
Simple AF	0.5%以下	0.8%未満
提案方式	0.5%以下	約0.8%未満

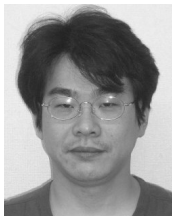
5 まとめ

大規模なファイル転送を伴う Grid アプリケーションのための QoS 制御方式を提案し、シミュレーションでその評価を実施した。通常の TCP を用いた場合は、バックグラウンドトラフィックの有無にかかわらず、提案方式は他の方式に比べて呼損率、スループット共に良い性能であり、要求スループットを守れなかった比率も他の方式と同等以上の性能を示す。一方、ウィンドウサイズを調整した評価から、1 セッション当たり 1 本の TCP でデータ転送を実施した場合、TCP の弱点のため提案方式の効果があまり得られない結果となった。

今後の課題としては、最低帯域保証サービス環境で、よりスループットを得られるデータ転送プロトコルと提案方式の組合せでの性能評価や、実際のアプリケーションの特性を反映したシナリオでの性能評価を実施する予定である。

参考文献

1 Global Grid Forum, http://www.gridforum.org/ .
2 OptIPuter, http://www.optiputer.net/ .
3 野呂，長谷川，馬場，下條，“Grid における大量データ送信に適した品質保証方式”，信学技報，IA2004-22，2005 年 1 月.
4 VINT project, "ns2", http://www.isi.edu/nsnam/ns/ .
5 S.Blake, D.Black, M.Carlson, E.Davies, Z.Wang, and W.Weiss, "An Architecture for Differentiated Service", RFC2475, December, 1998.
6 V.Jacobson, K.Nichols, and K.Poduri, "An Expedited Forwarding PHB", RFC2598, June, 1999.
7 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.
8 The Globus Alliance, "Grid FTP", http://www-fp.globus.org/datagrid/gridftp.html .
9 J.Heinanen, F.Baker, W.Weiss, and J.Wroclawski, "Assured Forwarding PHB Group", RFC2597, June, 1999.
10 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.



野呂正明（のろまさあき）
拠点研究推進部門大阪 JGNⅡリサーチセンター専攻研究員
インターネット、品質制御、グリッド技術

馬場健一（ばばけんいち）
拠点研究推進部門大阪 JGNⅡリサーチセンター専攻研究員　博士（工学）
広帯域通信網，コンピュータネットワーク，光ネットワークシステムの性能評価に関する研究

下條真司（しもじょうしんじ）
拠点研究推進部門大阪 JGNⅡリサーチセンター専攻研究員（大阪大学サイバーメディアセンター長・教授）　工学博士
マルチメディア応用システム、Peer-to-Peer、インターネット、ユビキタスネットワーク、グリッド技術