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

2006－ARC－167 (21)
2006－HPC－105 (21)
2006／2／28

グリッドにおける計算資源と光パスネットワーク資源のコアロケーション実験

竹房 あつ子[†1] 林 通秋[†2] 長津 尚英[†3]
中田 秀基[†1] 工藤 知宏[†1] 宮本 崇弘[†2]
大谷 朋広[†2] 田中 英明[†2] 鮫島 康則[†3]
今宿 亙[†3] 神野 正彦[†3] 滝川 好比郎[†3]
岡本 修一[†4] 田中 良夫[†1] 関口 智嗣[†1]

産業技術総合研究所，KDDI 研究所，日本電信電話は共同で，ネットワーク資源の事前予約を行うための標準ウェブサービスインタフェースを定める G-lambda プロジェクトを推進している．このインタフェースの有効性を確認するために，計算資源と光パスネットワーク資源を事前予約によりコアロケーションする実験を，情報通信研究機構の協力を得て行った．本稿では，前提とするコアロケーションシステムモデル，G-lambda プロジェクトにおけるグリッドでのネットワークサービスインタフェースの規定，産総研が開発した計算・ネットワーク資源を同時予約するスケジューラの概要，KDDI 研が開発した GMPLS で制御される光パスネットワークの提供機構の概要と，これらにより実現されたコアロケーションシステムプロトタイプの iGrid2005 での実証実験について報告する．

An Experiment of Co-allocating Computing and Lambda Path Resources on the Grid

Atsuko Takefusa,[†1] Michiaki Hayashi,[†2] Naohide Nagatsu,[†3]
Hidemoto Nakada,[†1] Tomohiro Kudoh,[†1]
Takahiro Miyamoto,[†2] Tomohiro Otani,[†2] Hideaki Tanaka,[†2]
Yasunori Sameshima,[†3] Wataru Imajuku,[†3] Masahiko Jinno,[†3]
Yoshihiro Takigawa,[†3] Shuichi Okamoto,[†4] Yoshio Tanaka[†1]
and Satoshi Sekiguchi [†1]

AIST, KDDI R&D Laboratories and NTT are promoting collaborative project G-lambda to establish a standard Web services interface between Grid and network resources. At iGrid2005, we conducted a experiment in cooperation with NICT, where our Grid scheduling system co-allocated computing and network resources with advance reservation through the interface. In this report, our co-allocation system model, the Web services interface (GNS-WSI) between GRS (Grid Resource Scheduler) and NRM (Network Resource Management System), overview of GRS which AIST has developed and NRM which KDDI R&D laboratories developed, and overview of the experiment are described.

1. はじめに

グリッド基盤技術の成熟により，異なる組織により管理される地理的に分散した計算資源上での大規模高性能計算が実行可能になった．しかしながら，現状では分散した計算資源上で実行される大規模高性能計算の多くは，通信遅延や帯域の不安定さにより著しい性能低下を引き起こすことのないパラメタ探索型やマスタ・ワーカ型のアプリケーションプログラムに限られる．これは，以下により広域分散グリッド環境で高性能計算を実行するのは困難なためである．

インターネットにより容易に広域計算が実行できるが，ベストエフォート型の通信を行うため，各アプリケーションプログラムの求める通信性能を保証できない．光ネットワーク技術を用いることで高品質通信が

†1 産業技術総合研究所 National Institute of Advanced Industrial Science and Technology (AIST)
†2 株式会社 KDDI 研究所 KDDI R&D Laboratories Inc.
†3 日本電信電話株式会社 NTT Network Innovation Laboratories
†4 情報通信研究機構 National Institute of Information and Communications Technology (NICT)

提供可能であるが，現状では計算に必要なネットワーク資源の構成・提供は自動化されておらず，E-mail や FAX 等により人同士で事前に交渉し，設定する必要がある．また，グリッド"スーパースケジューラ"は異なる組織間の分散した計算資源を必要に応じて確保・提供するものであるが，まだ実用段階に至っていない．現状では，他の組織の計算資源を利用するにはネットワーク資源同様，E-mail 等による事前の交渉が不可欠である．

産業技術総合研究所，KDDI 研究所，日本電信電話は共同で，ネットワーク資源の事前予約を行うための標準ウェブサービスインタフェースを定める G-lambda プロジェクト[1] を推進している．このインタフェースの有効性を確認するために，計算資源と光パスネットワーク資源を事前予約によりコアロケーションする実験を，情報通信研究機構の協力を得て行った．

本稿では，前提とする事前予約のためのコアロケーションシステムモデル，G-lambda プロジェクトにおけるグリッドでのネットワークサービスインタフェースの規定，産総研が開発した計算・ネットワーク資源を同時予約するスケジューラの概要，KDDI 研が開発した GMPLS で制御される光パスネットワークの提供機構の概要と，これらにより実現されたコアロケーションシステムプロトタイプの iGrid2005 での実証実験について報告する．

分散する計算資源間には高品質かつ安定した通信を可能にする光パスネットワークを用い，IETF で標準化が進められている GMPLS (Generalized Multi-Protocol Label Switching)[2] で制御することにより通信環境を柔軟に構成して提供する．GMPLS は分散した光クロスコネクト (OXC) 等のネットワーク機器を制御するプロトコル群である．必要に応じて計算資源間にユーザ専用の光パスネットワークを張ることにより，通信性能を保証することができる．

次に，G-lambda プロジェクトにおいてグリッドで柔軟に光パスネットワークを構築するため，ウェブサービスベースのネットワークサービスインタフェース GNS-WSI (Grid Network Service - Web Services Interface) を規定し，必要なネットワークの予約，修正，解放，予約状況の取得を可能にした．

また，分散した計算資源とネットワーク資源の事前予約によるコアロケーションを実現するため，GNS-WSI を用いた先行予約ベースのスケジューリングシステムを開発し，2005 年 9 月に米国サンディエゴで開催された iGrid2005 においてそのプロトタイプシステムの実証実験を行った．コアロケーションシステムは，主に両資源のコアロケーションを行うグリッド資源スケジューラ (GRS: Grid Resource Scheduler) と光パスネットワーク資源を管理するネットワーク管理システム (NRM: Network Resource Management system) からなる．実証実験では，ユーザが GUI(Graphical User Interface) から簡単な操作で計算機とネットワークに関する要求を GRS に送り，GRS が適宜 NRM に問い合わせながら適切な計算・ネットワーク資源を同時に事前予約し，予約された資源上でユーザの分子動力学アプリケーションを実行した．実験環境には，日本の 6 都市に配備された PC クラスタとその間の GMPLS ネットワークを用いた．本実験により，事前予約に基づくコアロケーションシステムとそのインタフェースの有効性・実現可能性を示した．また，資源予約サービスに関する技術的課題を明らかにした．

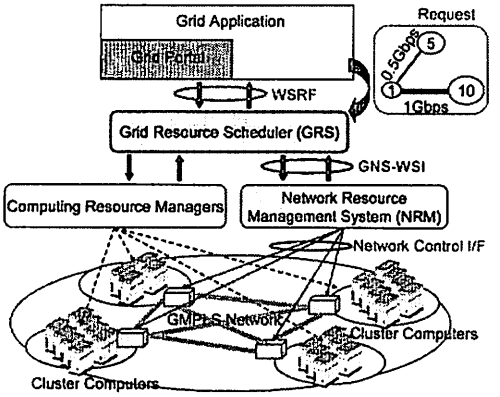

図 1　コアロケーションシステムの概要．

2. ウェブサービスベースのコアロケーションアーキテクチャ

図 1 にウェブサービスベースのグリッドコアロケーションシステムの概念図を示す．コアロケーションシステムはグリッド資源スケジューラ (GRS)，ネットワーク資源管理システム (NRM)，計算資源管理システム (CRM)，分散する実計算資源と計算資源間を結ぶ GMPLS で制御されるネットワークで構成される．GRS がユーザの要求に応じ，各サイトで計算資源を管理する OpenPBS[3] や N1 GridEngine[4] 等の CRM と，ネットワーク資源を管理する NRM にウェブサービスを介して連携し，適切な計算資源と資源間の光パスネットワーク帯域を選択し，事前予約する．

提供するネットワーク資源には，通信性能が保証可能な光パスネットワークを用いる．また，光パスネットワークの制御には IETF で標準化が進められている GMPLS 技術を用いる．GMPLS を用いることにより，分散した計算機や記憶装置等を要求されるバンド幅で適宜接続することができるため，計算効率と利用効率の向上が期待できる．

グリッドにおいて提供される資源は地理的に分散しているため，資源間を繋ぐネットワークは複数の都市や国を跨って接続され，複数のネットワークオペレータ

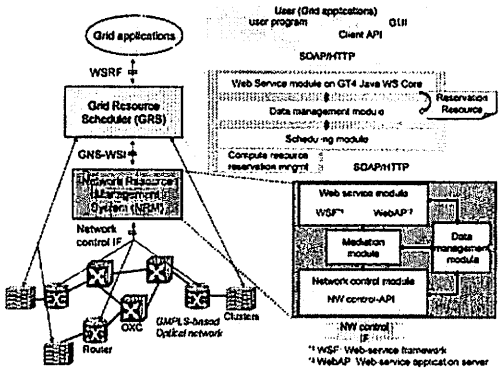

図 2　コアロケーションシステムアーキテクチャ.

が介在することが予想される．よって，GRS と NRM 間の連携で必要とされる情報交換には，ネットワークオペレータ間で共通のインタフェースを用いることが重要である．G-lambda プロジェクトでは，グリッドネットワークサービスウェブサービスインタフェース (GNS-WSI) としてネットワーク資源の確保のための基本的な機能を定義する．

2.1 GNS-WSI

2 拠点間を接続する光パスネットワークの提供を前提とした初期 GNS-WSI として，ウェブサービス記述言語 WSDL により以下のポーリングベースのオペレーションを定義した．

netResourceReservation　GRS から要求された仮想パスネットワークを OXC 間の物理リンクに割り当て，事前予約する．GRS はエンドポイントの拠点 ID，バンド幅，予約期間，遅延や耐故障性に関する要求を NRM に送信する．複数のパスを同時に要求することもできる．

netResourceModification　予約されたパスのバンド幅や予約期間等の予約情報を変更する．

netResouceRelease　予約されたパスを解放する．

netResourceReservationStatusQuery　要求された／予約されたパスの現在のステータスを返す．予約，変更，解放要求の結果は，このオペレーションにより取得することができる．また，確保されたパスが予約した時刻に実際に利用可能になったか否かの情報も取得することができる．

netResourceQuery　GRS の指定したパスが指定した期間に利用可能かどうかを返す．

netAvailableResourceQuery　GRS が指定した期間に利用可能なパスをすべて返す．

2.2 グリッド資源スケジューラ (GRS)

GRS はユーザの要求に応じて適切な高品質計算環境を提供するため，計算資源は各サイトの CRM に，ネットワーク資源は GNS-WSI を介して NRM に問い合わせて適切な資源を同時に確保する．GRS では事前予約，解放，予約情報取得のためのオペレーションを提供している．

GRS のプロトタイプとして，Globus Toolkit 4 (GT4)[5),6)] を用いて WSRF(Web Services Resource Framework)[7)] 仕様に基づく計算・ネットワーク資源を同時に確保するスケジューラを開発した．WSRF はウェブサービスにおけるステートフルリソースのモデリングとアクセスのための仕様であり，OASIS でその標準化が進められている．

コアロケーションシステムを構成する主要モジュール GRS と NRM の詳細を図 2 に示す．GRS は次のモジュールで構成される．

Client API, GUI　ユーザにプログラミング／グラフィカルインタフェースを提供する．

ウェブサービスモジュール　WSRF に基づく予約サービスインタフェースをユーザに提供する．このモジュールで個々の資源予約要求に対する予約リソースを生成し，データ管理モジュールへ格納する．ユーザにはこの予約リソースにアクセスするためのエンドポイントリファレンスを返す．これにより，ユーザは現在の予約状況や予約結果を取得することができる．

予約リソース　個々の予約要求に対する予約サービス情報を保持する．

データ管理モジュール　各予約処理の状況を管理する．格納データの持続性を保証している．

スケジューリングモジュール　ユーザの計算資源とネットワーク資源に関する要求と，予約期間とデッドライン等の時間に関する要求をもとに，NRM, CRM と連携して適切な資源を割当てる．要求された資源が先行予約できれば"予約成功"，できなければ"予約失敗"のステータスをデータ管理モジュールへ格納する．このステータスはウェブサービスモジュールを介してユーザに送られる．予約が失敗した場合は，ユーザが新しい条件を再度送信することになる．

計算資源予約管理　各サイトの CRM で管理される予約情報を管理する．

2.3 ネットワーク資源管理システム (NRM)

NRM では GMPLS で制御される光パスネットワーク資源を GNS-WSI を介して提供する．KDDI 研究所が開発した NRM は，下記の 3 つの機能モジュールから構成され，GRS に対して GMPLS によるネットワークサービスを提供する[8)]．

ウェブサービスモジュール　GNS-WSI を介して GRS 等のグリッドブローカリングシステムに GMPLS で制御される光パスネットワーク資源を提供する．これにより，GRS はネットワーク資源の発見，光パスの事前予約，予約の修正等を行うことができる．

中間モジュール　GRS からの要求と実際のネットワーク資源間を仲介し，GMPLS 制御光ネットワークを

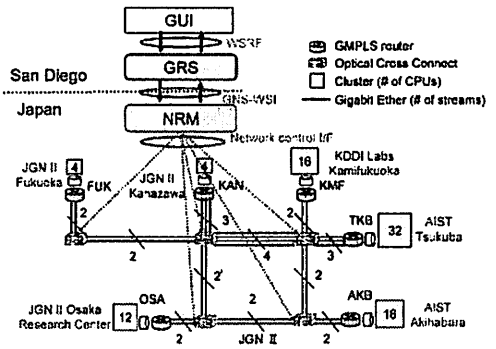

図 3　iGrid2005 実験環境.

計算資源間のパスとして仮想化する．また，光パスのスケジューリングも行う．

ネットワーク制御モジュール　ネットワーク制御インタフェースを介して GMPLS ルータを制御し，要求された光パスを利用可能にしたり，それらの状況や性能をモニタする．

3. iGrid2005 実証実験の概要

2005 年 9 月に米国カリフォルニア大学サンディエゴ校で開催されたグリッドに関するデモを中心とした国際会議 iGrid2005 において，本グリッドコアロケーションシステムの実証実験を行った．実験では，ユーザがグリッドの資源に関する要求を GUI から GRS へ送信し，GRS が NRM と連携して柔軟に適切な計算・ネットワーク資源を同時に事前予約する．予約が成功すると，予約された時刻に確保された資源を用いて科学アプリケーションプログラムを実行する．

図 3 に実証実験環境を示す．実験では，GRS と GUI はサンディエゴの会場の PC 上で，NRM は大手町の WS 上で実行した．図 3 に示すように，日本の 6 都市 (つくば，秋葉原，上福岡，金沢，大阪，福岡) に配置された計算資源とその間を結ぶ GMPLS ネットワークを用いた．本実験では，各計算資源サイトでは事前予約機能を持たない CRM を用いたため，全ての計算資源をコアロケーションシステムで占有し，先行予約の管理を GRS で行った．GMPLS ネットワークは光クロスコネクトと GMPLS ルータで構成されている．予約した資源上では，グリッドミドルウェア Ninf-G2[9] と Globus Toolkit 2 (GT2)[5] を用いて実装された分子動力学アプリケーションプログラムを実行する．実証実験にあたり，産総研，KDDI 研，NTT が GNS-WSI を規定し，産総研が開発した GRS，KDDI 研が開発した NRM を用いた．また，NICT が運営する JGNII のネットワークテストベッド上に KDDI 研，NTT が動作環境を構築した．

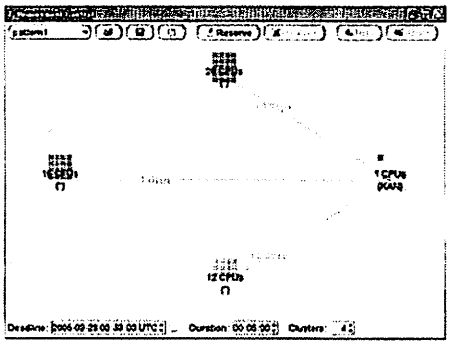

図 4　資源予約入力画面.

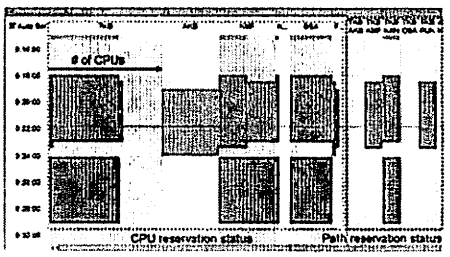

図 5　資源予約タイムテーブル (左：計算資源の予約 CPU 数，右：光パスネットワークの予約帯域).

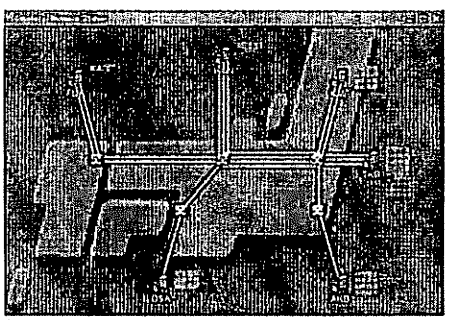

図 6　予約状況を表示する地図画面.

3.1 実証実験の詳細

図 4，図 5，図 6，図 7 に実証実験の表示画面を示す．実証試験では，図 4 の資源予約入力画面から簡単な操作でクラスタ数と各クラスタの CPU 数，その間のネットワーク構成と帯域，および予約期間とデッドラインに関する要求を入力し，GRS に送信する．GRS は通常要求された資源に対して実際の計算資源サイトとその間のネットワークを割当てるが，ユーザが一部の計算資源サイトを指定することも可能である．

ユーザの要求が GUI から GRS に送信されると，GRS が NRM とウェブサービスを介して連携して適切な計算・ネットワーク資源が事前予約し，その結果

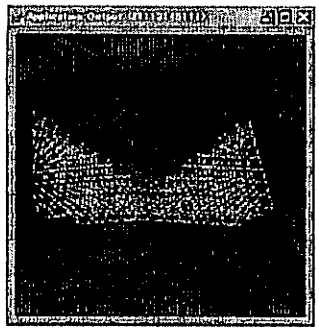

図 7　分子動力学アプリケーション実行結果画面.

が図 5 の資源予約タイムテーブルに反映される．ここでは，画面の左側に各サイトで予約された CPU の数，右側に予約された計算資源間の光パスネットワークバンド幅を示している．タイムテーブルの縦軸は時刻を示し，中央の横線が現在の時刻を表している．

図 6 は現在のネットワークの予約状況を表示する地図画面を示す．地図上に表示されるサイト間のパスネットワークは，実際の物理ネットワーク上で GMPLS の制御パケットを拾って表示させたものである．予約した時刻になると，NRM からの制御により割当てられた光パスが利用可能となり，地図上にパスが出力される．

図 7 は先行予約で確保され，動的に構築された高品質計算環境において分子動力学シミュレーションを実行した結果である．この実験により，地理的に分散した計算資源とその間の高品質ネットワークを GNS-WSI を介して統一的に扱い，予約ベースで必要に応じて柔軟に確保・提供できることを実証した．

3.2 議　　論

3.2.1 GRS-NRM 間の SOAP 通信遅延

GRS と NRM が連携してコアロケーションする際，GRS と NRM 間で複数の SOAP メッセージが送受信される．この際，通信遅延に加えて SOAP 通信にかかる遅延が加わるため，本実験のように通信遅延の大きい環境では，ユーザが GUI から資源要求を送信してから結果を受け取るまでの時間が長くなってしまう．iGrid2005 では数秒程度と許容できる時間であったが，大規模環境でより複雑な資源要求がされる場合は深刻な問題となる．GRS-NRM 間の通信回数をできるだけ少なくし，大規模環境にも耐えうるプロトコルを開発する必要がある．

3.2.2 光パスネットワークのモニタリング

本実験環境は GUI，GRS，NRM 等のプログラムモジュールと 6 拠点に分散したクラスタおよびその間の GMPLS ネットワークからなり，組み上げて安定した操作が可能になるまで多くの人手と時間を要した．特に，GMPLS 光パスネットワーク構築におけるデバッグ手法がまだ未熟であるため，予約した時刻に光パスネットワークが利用可能にならない等の不具合が発生した際にその原因究明に時間を要した．しかしながら，実証実験を目的として開発したネットワークの予約状況を表示する地図画面 (図 6) では，GMPLS でパスネットワークを利用可能にする際に GMPLS ルータ間で送受信されるメッセージパケットを直接拾って表示させていたため，比較的容易に原因箇所を特定することができた．よって，光パスネットワークレベルのモニタリングはデバッグ手法として非常に有効であり，今後その手法の改良と標準化が重要となる．

3.2.3 光パスネットワークのルーティング

GMPLS ネットワークでは，OSPF(Open Shortest Path First) ベースの経路計算手法を用いて動的に光パスの経路を設定することができる．この GMPLS ルーティングプロトコルの機能により，光ネットワークの障害耐力と資源利用効率を向上させることができる．しかしながら，GRS-NRM 間のネットワークサービスで事前にきめられた光パスの通信遅延に関する条件が，動的なルーティングにより変化してしまうため，本実験ではこの機能を利用しなかった．よって，本実験では GMPLS ネットワークにおける各光パスの経路を明示的に指定し，固定することにした．GMPLS ネットワークにおける動的なルーティング機能を有効に利用するためには，グリッドアプリケーションの通信遅延に関する要求条件を明らかにする必要がある．

3.2.4 耐故障性とエラー通知機能

資源予約の耐故障性はコアロケーションシステムの実用化に向けた重要な課題である．iGrid2005 の実験では，光パスの先行予約が成功したにも関わらず，設定上のバグや GMPLS メッセージングのエラーによる不具合が発生し，予約された時刻にパスネットワークが利用可能にならないことがあった．一方，Ninf-G と GT2 で実装されたアプリケーションプログラムにはこのエラーが検知できないため，プログラムの実行を開始して計算資源間の通信が確立されないまま待ち続け，複数プロセスが計算資源上に残ってしまった．この問題を解決するため，実験ではアプリケーションの実行開始時間を 10 秒程度遅らせるとともに，GRS が NRM に予約したネットワークが予約時刻に利用可能になったかどうか問い合わせて，何らかのエラーが発生した場合は GUI からユーザに通知するモニタ機能を用意した．今後，資源予約の耐故障性を強化するとともに，致命的なエラーが発生した場合に事前にユーザに通知したり，その際にユーザに何らかの補償をするメカニズムを開発する必要がある．

4. 関 連 研 究

グリッドで高品質なネットワークを提供する試みがいくつかある．CANARIE[10] はカナダの先進的インターネット開発の推進を目指した組織であり，教育目

的の光パスネットワークとしてCA*net4を配備している．CA*net4ネットワークの制御において，エンドユーザがネットワークをソフトウェアオブジェクトとして扱い，光パスの配備や再構成が容易にできるようにするため，UCLP (User Controlled LightPaths) ソフトウェアを開発している[11]．UCLPはウェブサービスを介してユーザが指定する個々の光パスネットワークを直接操作するものであり，事前予約機能や計算機とのコアロケーション機能は用意されていない．

VIOLAは本研究同様，グリッドで計算資源とネットワーク資源を同時に提供することを目的としている[12]．VIOLAではGGFにおいて標準化がすすめられているWS-Negotiation/-Agreement仕様に基づくメタスケジューラをUNICORE[13]ベースのグリッド環境で開発することを提案している．しかしながら，実装の詳細は明らかでない．

グリッドスーパースケジューラでは Silver[14]，CSF[15]，NAREGIスーパースケジューラ[16]など，複数開発されているが，その多くが実用段階には至っていない．また，計算資源の割当てが主体であり，性能が保証されたネットワーク資源と計算資源のコアロケーションを実現しているものはない．

5. まとめと今後の課題

本稿では，事前予約に基づくコアロケーションシステムモデル，産総研，KDDI研，NTTの共同プロジェクトG-lambdaにおけるグリッドでのネットワークサービスインタフェースの規定，産総研が開発した計算・ネットワーク資源を同時予約するスケジューラの概要，KDDI研が開発したGMPLSで制御される光パスネットワークの提供機構の概要と，これらにより実現されたコアロケーションシステムプロトタイプのiGrid2005での実証実験について報告した．iGrid2005実験では，ユーザがGUIから簡単な操作で計算機とネットワークに関する要求を入力することで，GRSがGNS-WSIを介してNRMと連携して適切な計算・ネットワーク資源を同時に事前予約し，柔軟かつ容易に高品質な計算環境を提供可能であることを実証し，その有効性を示した．また，資源予約サービスに関する技術的課題を明らかにした．

今後はG-lambdaにおいて複数GRSと複数NRMを考慮してGNS-WSIを洗練していくとともに，GGFやIETF等において標準化を進める．また，海外の光ネットワークとの連携実験も検討している

謝辞 iGrid2005実証実験にあたり，多大なご協力をいただいた谷村勇輔氏，武宮博氏，岡崎史裕氏（産総研），鈴木正敏氏，鶴沢宗文氏（KDDI研究所），大原拓也氏，築島幸男氏（NTT），下條真司氏，秋山豊和氏（NICT大阪JGNII RC），小林正朋氏（東京大学）に深く感謝いたします．

参考文献

1) G-lambda. http://www.g-lambda.net/.
2) GMPLS. RFC 3945.
3) OpenPBS. http://www.openpbs.org/.
4) N1 Grid Engine. http://gridengine.sunsource.net/.
5) Foster and Kesselman: Globus: A Metacomputing Infrastructure Toolkit., *International Journal of Supercomputer Applications*, pp. 115–128 (1997).
6) Foster: Globus Toolkit Version 4: Software for Service-Oriented Systems, *IFIP International Conference on Network and Parallel Computing, Springer-Verlag LNCS 3779*, pp.2–13 (2005).
7) OASIS Web Services Resource Framework (WSRF) TC: Web Services Resource 1.2 (WS-Resource) (2005). http://www.oasis-open.org/committees/wsrf/.
8) Hayashi, Miyamoto, Otani, Tanaka, Takefusa, Nakada, Kudoh, Nagatsu, Sameshima and Okamoto: Managing and Controlling GMPLS Network Resources for Grid Application, *Optical Fiber Communications Conference (OFC) 2006* (to appear).
9) Tanaka, Nakada, Sekiguchi, Suzumura and Matsuoka: Ninf-G: A Reference Implementation of RPC-based Programming Middleware for Grid Computing, *Journal of Grid Computing*, pp.41–51 (2003).
10) CANARIE. http://www.canarie.ca/.
11) Boutaba, Golab, Iraqi, Li and Arnaud: Grid-Controlled Lightpaths for High Performance Grid Applications, *Journal of Grid Computing, Special Issue on High Performance Networking* (2004).
12) Barz: Dynamic allocation of network resources in VIOLA, *VIOLA workshop* (2005).
13) Romberg: The UNICORE Architecture: Seamless Access to Distributed Resources, *Proceedings of the Eighth IEEE International Symposium on High Performance Distributed Computing*, pp.287–293 (1999).
14) Moab Grid Scheduler (Silver) Administrator's Guide version 4.0. http://www.clusterresources.com/products/mgs/docs/.
15) Community Scheduler Framework. http://sf.net/projects/gcsf.
16) 畑中，中野，井口，大野，佐賀，秋岡，中田，松岡:OGSAアーキテクチャに基づくNAREGIスーパースケジューラの設計と実装，情報処理学会研究報告 2005-HPC-102，pp.33–38 (2005).