10-IS03

学術グリッド基盤の構築・運用技術に関する研究

合田 憲人（国立情報学研究所）

概要

本研究では，我が国における e-サイエンスを活用した研究を促進することを目指し，学際大規模情報基盤共同利用・共同研究拠点に設置された計算機，およびこれらを接続する学術情報ネットワークである SINET3 から構成される実用的なグリッド基盤を構築し，グリッドミドルウェアの配備・運用技術およびグリッド環境上でのユーザ管理技術に関する研究を行う．グリッドミドルウェアの配備・運用技術については，Gfarm2 を用いたデータグリッド環境の配備方法を検討し，インストールソフトウェアの開発や実証実験を行った．また，SINET3 の 10Gbps の帯域を持つネットワーク上でファイルを高速に転送するための技術に関する研究を行い，実証実験を行った．グリッド環境上でのユーザ管理技術に関する研究では，Shibboleth 認証連携技術を用いたグリッド認証システムの構築を行い，実証実験を行うとともに，本システムを用いたユーザ登録業務の業務フローを検討した．

1. 研究の目的と意義

ネットワーク上に分散した様々な研究データを融合して処理することにより，未知の問題解決や科学的発見を行う新たな研究手法（e-サイエンス）が注目されている．e-サイエンスを実現するためには，ネットワーク上に分散した様々なデータを連携し，かつ高性能計算機群を用いてこれらのデータを高速に処理するための基盤が必要となる．このような背景のもと，1990 年代よりグリッド技術の研究が世界的に進められ，国内においても多くの基礎研究の成果が報告されている．これらの基礎研究の成果を利用し，グリッドを e-サイエンス実現のための基盤として活用するためには，様々なサイエンスアプリケーション分野で利用可能な実用グリッド基盤を構築・運用する必要がある．しかし，これを実現するための技術には未だ確立されていない部分も多く，解決しなければならない問題も残されている．

本研究では，我が国における e-サイエンスを活用した研究を促進することを目指し，学際大規模情報基盤共同利用・共同研究拠点に設置された計算機，およびこれらを接続する学術情報ネットワークから構成される実用的なグリッド基盤を構築・運用するための技術に関する研究を行う．具体的には，以下にあげる「グリッドミドルウェアの配備・運用技術に関する研究」および「グリッド環境上でのユーザ管理技術に関する研究」を行い，日本の実用グリッド基盤の構築，およびその運用技術の確立に貢献する．

一般に，計算機の構成やソフトウェア設定は運用組織毎に異なるため，グリッド基盤の構築では，組織間で差異を考慮した上で，適切にグリッドミドルウェアを配備・設定しなければならない．また，グリッド上で大規模データを共有するためには，大容量ネットワークの性能を活用した高性能データ転送サービスを提供する基盤の構築も重要となる．しかし，これらの基盤の構築・運用技術は未だ確立されておらず，運用ノウハウも蓄積されていない．「グリッドミドルウェアの配備・運用技術に関する研究」では，各拠点で運用される計算機（ストレージも含む）上へのグリッドミドルウェアの配備・運用方法，グリッド環境上の計算機（ストレージ）間で大容量ネットワークの性能を活用した高性能データ転送を実現するための技術に関する研究を行う．また，本研究で構築するグリッド環境の試験運用を通して，グリッド環境の運用ノウハウを蓄積する．

現在，情報基盤センター等の計算機システム管理は，その管理組織の運用方針に従って独立して行われており，ユーザアカウント管理も組織ごとに異なる．また，グリッド上でのユーザ認証方式

では，Grid Security Infrastructure (GSI)[1]に代表される証明書に基づく方式がデファクトスタンダードとなっており，世界的に利用されている．しかし現在，これらのユーザ管理を連携させる方式がなく，利用者は，利用する計算機システム毎のユーザ登録手続きのほかに，グリッド用証明書の申請手続きも行わなければならない.「グリッド環境上でのユーザ管理技術に関する研究」では，情報基盤センター等での計算機システムのアカウント登録およびグリッド用証明書の申請を一括して行うための技術および制度に関する研究を行う．

欧米では，早くから TeraGrid[2]や EGEE[3]に代表される実用グリッド基盤の構築・運用が進められており，すでに様々なサイエンスアプリケーションの分野で実用グリッド基盤として活用され，e-サイエンスを活用した研究が進められている．しかし日本では，グリッド技術の基礎研究については 1990 年代より多くの成果があるものの，グリッド基盤については実験的な運用のみであり，実用グリッド基盤については未だ運用されていない．このような e-サイエンスの基盤となる実用グリッド基盤の欠如は，サイエンス研究の国際的な競争力低下にもつながりかねない．本研究は，日本国内の実用的な e-サイエンス基盤として，学際大規模情報基盤共同利用・共同研究拠点からなる実用グリッド基盤の構築・運用を目的とするものであり，国内の e-サイエンス研究の促進，および e-サイエンスを利用した様々なサイエンスアプリケーション分野での研究成果の創出，さらには日本のサイエンス研究の国際的競争力の維持に貢献することを目指している．

2. 当拠点公募型共同研究として実施した意義

(1) 共同研究を実施した大学名

本研究は，北海道大学，東北大学，東京大学，東京工業大学，名古屋大学，京都大学，大阪大学，九州大学との共同研究として実施している．

(2) 共同研究分野

本研究は，「大規模情報システム関連研究分野」における研究として実施している．

(3) 当公募型共同研究ならではという事項など

本研究が目指すグリッド環境の構築および運用では，学際大規模情報基盤共同利用・共同研究拠点を構成する 8 大学に筑波大学を加えた 9 大学情報基盤センターの計算資源上へのグリッドミドルウェアのインストールや設定作業が必要であり，各拠点の教職員との協力体制を整備している．また，グリッド基盤のユーザ登録業務の試験運用を継続するため，各大学の全国共同利用担当者の協力を得ながら研究を進めている．グリッドミドルウェアの配備については，グリッドミドルウェア技術に関して深い知見を持つ九州大学，大阪大学，筑波大学，東京工業大学の研究者からの協力を得ながら進めている．さらに，実証実験のための流体音計算を実施するために九州工業大学と九州大学の研究者からも協力を得ている．

3. 研究成果の詳細

本研究では，学際大規模情報基盤共同利用・共同研究拠点に設置された計算機，およびこれらを接続する学術情報ネットワークである SINET3[4]から構成される実用的なグリッド基盤を構築・運用するための技術に関する研究を行うことを目的としている．本節では，最初に本研究において構築しているグリッド基盤について紹介するとともに，本基盤上で研究を進めているグリッドミドルウェアの配備・運用技術およびグリッド環境上でのユーザ管理技術に関する研究成果を報告する．

3.1. グリッド基盤

[1] http://www.globus.org/security/
[2] https://www.teragrid.org/
[3] http://www.eu-egee.org/
[4] http://www.sinet.ad.jp/

表 1 計算資源

拠 点	ハードウェア	コア数/ノード	メモリ/ノード	ノード数
北 大	DELL PowerEdge R200, HitachiHA8000/110W	2	2/4GB	27
東北大	NEC SX-9	16	1000GB	4
筑波大	Appro XtremeServer-X3	16	32GB	4
東 大	HitachiHA8000-tc/RS425	16	32GB	8
東工大	HP ProLiant SL390s	12	54-96GB	375
名 大	Fujitsu PRIMERGY RX200	2	2GB	6
京 大	Fujitsu HX600	16	32GB	4
阪 大	NEC SX-8R	8	64/256GB	8
	NEC SX-9	16	1000GB	8
	NEC Express 5800/120Rg-1	4	16GB	32
九 大	Fujitsu PRIMERGY RX200S3	4	8GB	12

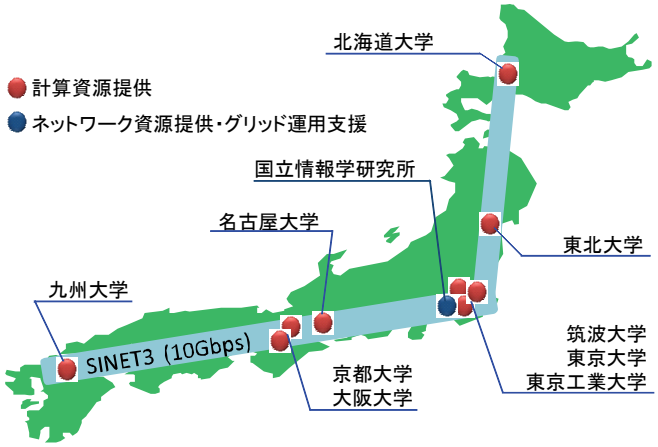

図 1 グリッド基盤構成組織

本研究で構築しているグリッド基盤は，図 1 に示す 9 大学の情報基盤センターおよび国立情報学研究所の 10 組織により構成されている．9 大学情報基盤センターは，資源提供サイトとしての役割を持ち，各センターが運用する計算資源の一部（表 1 参照）を本研究用に提供している．また，国立情報学研究所は，SINET3 内の 10Gbps のネットワークを提供するとともに，グリッドオペレーションセンター（GOC）としての役割を持ち，グリッドの運用に必要なサービス（具体的にはポータル，情報サービス，ジョブのブローカリング等）を提供するサーバのホスティングを行っている．

グリッド基盤の構築・運用では，グリッド基盤を構成するハードウェアやソフトウェアの整備とともに，グリッド基盤を運用するための人員体制を整備することも重要な課題である．本研究では，グリッド配備・運用タスクフォースの参加メンバーからなる運用組織を構成している．本タスクフォースは，グリッドミドルウェアの配備および運用技術に関する検討を目的として 2008 年に発足した組織であり，図 1 に示す組織の教職員から構成されている[5]．

本タスクフォースでは，グリッド配備技術のための研究として，グリッドミドルウェアを各計算資源にインストール・設定するためのソフトウェアツールを開発するとともに，各大学情報基盤センターの教職員間でのグリッドミドルウェア配備方法に関する情報共有を図っている．また運用技術については，グリッド用証明書の発行申請および各情報基盤センターへのユーザ登録申請を連携させたグリッドパックと呼ばれるサービスを提案している．グリッドパックは，グリッド用証明書の発行および（複数の）情報基盤センターのユーザ登録を一括して申請可能とするサービスであり，本タスクフォースでは，このための業務フローの策定，運用者用ソフトウェアツールの開発を行い，

[5] 小林泰三，天野浩文，青柳睦，合田憲人，“e-サイエンスを実現するグリッド技術：4.大学間連携グリッド基盤の運用”，情報処理，Vol.51 No.2, pp. 134-143，2010 年

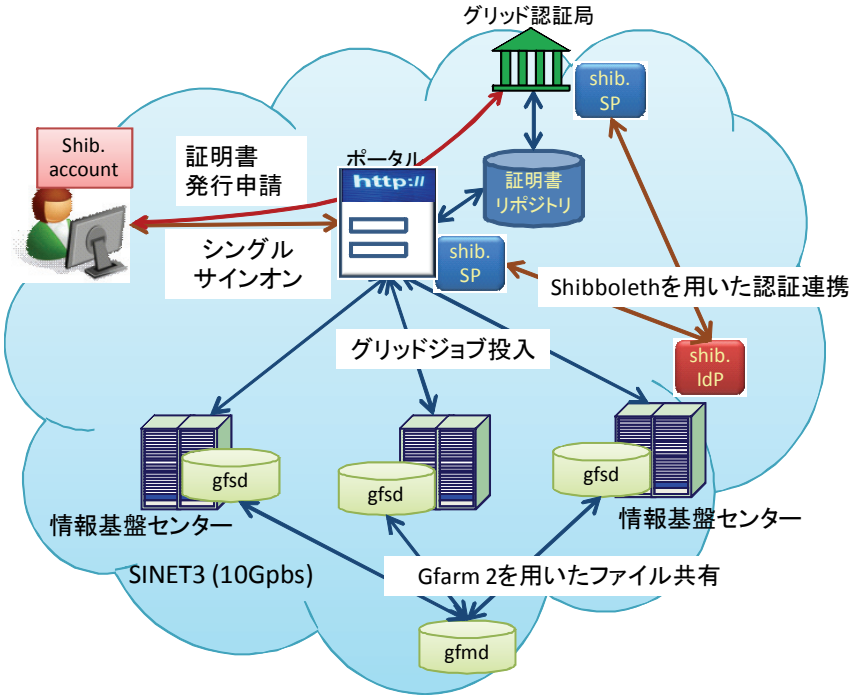

図 2 グリッド基盤上でのサービス

現在，運用実験を進めている．

本研究では今年度，図 2 に示す 2 つサービスをグリッド基盤上で構築することを目標としている．1 つ目のサービスは，情報基盤センターのストレージ群を SINET3 のネットワークで接続したデータグリッドサービスの構築であり，Gfarm2[6]を用いた環境構築を進めている．2 つ目のサービスは，情報基盤センターの認証とグリッド認証を連携させた認証基盤サービスの構築であり，Shibboleth[7]を用いた環境構築を進めている．これらのサービスが構築されると，ユーザは，情報基盤センターのローカルアカウントを用いてグリッド基盤にシングルサインオンし，計算ジョブの投入やグリッド上に分散したファイルの共有が可能となる．以下，それぞれのサービスの構築状況について報告する．

3.2. グリッドミドルウェアの配備・運用技術

本研究では，e-サイエンス研究に必要な機能として，情報基盤センターに設置された計算機資源（ストレージも含む）および SINET3 が提供するネットワークから構成されるデータグリッド環境を構築し，運用に必要なグリッドミドルウェアの配備方法や設定方法に関する検証を行うことを目的としている．現在，Gfarm 2.3.1 を用いて九州大学と大阪大学にファイルシステムノード（gfsd），国立情報学研究所に分散したファイルのメタデータを管理するメタデータサーバ（gfmd）を設置した環境が構築されている．また，これらの資源間は SINET3 L3 VPN サービスを用いて，10Gbps の帯域を持つネットワークにより接続されている．

3.2.1. Gfarm インストーラ

Gfarm2 に限らず，グリッドミドルウェアとして利用するソフトウェアは，グリッド向けにソフトウェア設定のインターフェースを整える必要がある．これは，グリッド基盤を形成する各基盤センターが計算機資源の管理単位として各々独立しているために必要になる問題であり，グリッド全体としての計算環境の整合性を確保するためには設定や運用に OOP の様な必要な情報の階層化と隠蔽が必要になってくるからである．今回

[6] http://datafarm.apgrid.org/
[7] http://shibboleth.internet2.edu/

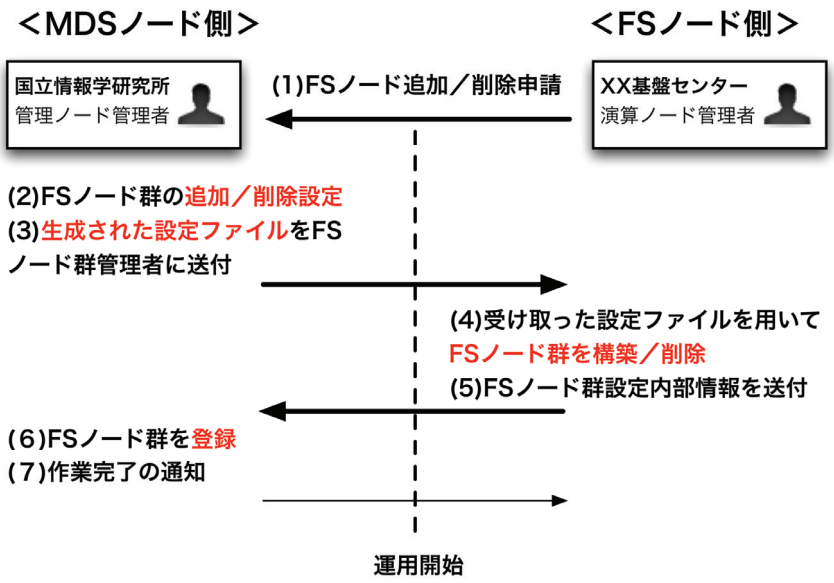

図 3 Gfarm インストール手順

Gfarm2 を配備するにあたって研究開発し実装した項目は以下の 3 つである．(1)管理ノード群(MDS, GFTP) とファイルシステムノード群(FS)の 2 つに Gfarm のサービスを分類して，(2)各情報基盤センターが提供するサービスに応じたメタ設定ファイルの生成と Gfarm のインストール／更新時などにメタ設定ファイルを利用する機構を開発実装し，(3)各基盤センター間でのインストール手順を制定した．

(1)は典型的なグリッド基盤の構造に合わせてあり，グリッド基盤全体を管理する役割を担う管理ノード群と，グリッド基盤に演算環境を提供する演算ノード群の分類を踏襲したものである．

(2)は(1)のサイト関係の下で各グリッドミドルウェアの動作状況整合性を確保するために必要な設定情報を集約したものをメタ設定ファイルとして定義している．このメタ設定ファイルで扱う情報は，hostname や IP などのノードの対外的な情報のみであり，OS やインストールされているソフトの情報などのノード内部に関する事柄は扱わないのが階層化の側面での重要な特徴である．実際に扱う情報は以下である．

- 管理ノード群情報
  - VO: serial number, name, hostname, port, location（内部 or 外部）
  - MDS: hostname, IP, location
  - GFTP: hostname, IP, location
  - 管理ユーザ: name, DN
  - テストユーザ: name, DN
- FS ノード群情報
  - FS: serial number, hostname, IP, DN, spool dir, RP name, location

これらの情報を該当情報基盤センター向けに組み合わせてメタ設定ファイルを自動生成する．現在の Gfarm2 構成では管理ノード群が 1 つで NII にあり，各基盤センターに FS ノード群が配備されている．また Gfarm2 の仕様と運用形態上すべてのサイトが全 Gfarm2 関連ノードの情報を必要とするために，基盤センター毎のメタ設定ファイルの差異はノードの location と FS ノードの通し番号になる．適切に生成され配置されたメタ設定ファイルの情報は，Gfarm2 の rpm パッケージに同梱したスクリプトを実行することにより読み込まれて Gfarm2 の設定に反映されるようになっている．この構成を採用することにより，rpm の pre, post section を利用したインストール／更新時の自動設定のみならず，何らかの都合でメタ設定ファイルのみを変更した場合でもスクリプトの再実行によりグリッド基盤全体の整合性を保つことを可能にしている．

表 2 チューニング前後の拠点間通信性能

From\To (Unit: Mbps)	東工大	阪大	NII	名古屋大学 (Max 1000Mbps)
東工大		1,296 3,320	4,128 4,472	176 544
阪大	1,496 4,208		1,336 3,328	208 672
NII	4,072 7,680	880 3,792		96 728
名古屋大学 (Max 1000Mbps)	296 816	368 832	440 824	

(3)は，(1)および(2)での設定情報の集約と階層化の結果として非常に簡素なものにまとまっている（図 3 参照）.

図中，太い矢印が管理者間のファイルの遣り取りを示しており，赤字がツールを用いた計算機上での管理作業を表している. FS ノードの申請を除けば一往復の手続きで作業は完結する.

3.2.2. サイト間高速ファイル転送

情報基盤センター間には SINET3 による最大 10Gbps の通信網が張り巡らされているが，あくまで通信経路を提供しているだけであり，性能を意識した通信を行うには，各拠点のルータ・スイッチ・計算機端末におけるパラメータチューニングが必要である．特に情報基盤センター群は北海道から九州まで広く分散しているため，拠点間通信の遅延が数十ミリ秒と大きく，この遅延に耐えうるパラメータチューニングを行わないかぎり，Gfarm2 によるデータグリッドのファイル転送はリンク性能を十分に活かせない.

本研究では，広帯域・高遅延の SINET3 上での高速データ転送を実現するための通信パラメータサーベイ・チューニングを，東京工業大学，大阪大学，東北大学，筑波大学，名古屋大学，国立情報学研究所に設置した性能測定サーバを用いて行った. 特に有効だったパラメータはサーバ NIC のパケットキュー長，TCP バッファサイズであった. その他, パケットスペーシングを行う帯域制御（今回はリンク性能に基づき静的に設定），TCP Segmentation Offloading 等の NIC 固有の機能の有効･無効, 割り込み頻度の調整等も行った結果，10Gbps 接続拠点間では，最大で標準パラメータ時の 4 倍程度の通信性能向上が得られた．結果の一部を表 2 に示す．表中，上段が標準パラメータ時の結果，下段がチューニング後の結果である．未だ理論性能から大きく乖離している結果が多いが，この理由としては，各拠点内部のルータ・スイッチの細かな設定までは対応出来ていないこと，我々がグリッド配備に用いている SINET3 ネットワークが共有回線であり，他の通信と共存していることが挙げられる．ただ，今後，ルータ・スイッチでのジャンボフレーム対応，通信状況に応じた動的なパケットスペーシングを行うことで，さらなる性能向上を達成できると見込んでいる.

3.2.3. 3 次元流体音数値計算による実証実験

OpenFOAM を用いた 3 次元流体音数値計算による実証実験を進めている．現在は OpenFOAM-1.7.1 をグリッド基盤の演算ノード環境に配備してテストを始めた状況であり, 今後，評価を進める予定である.

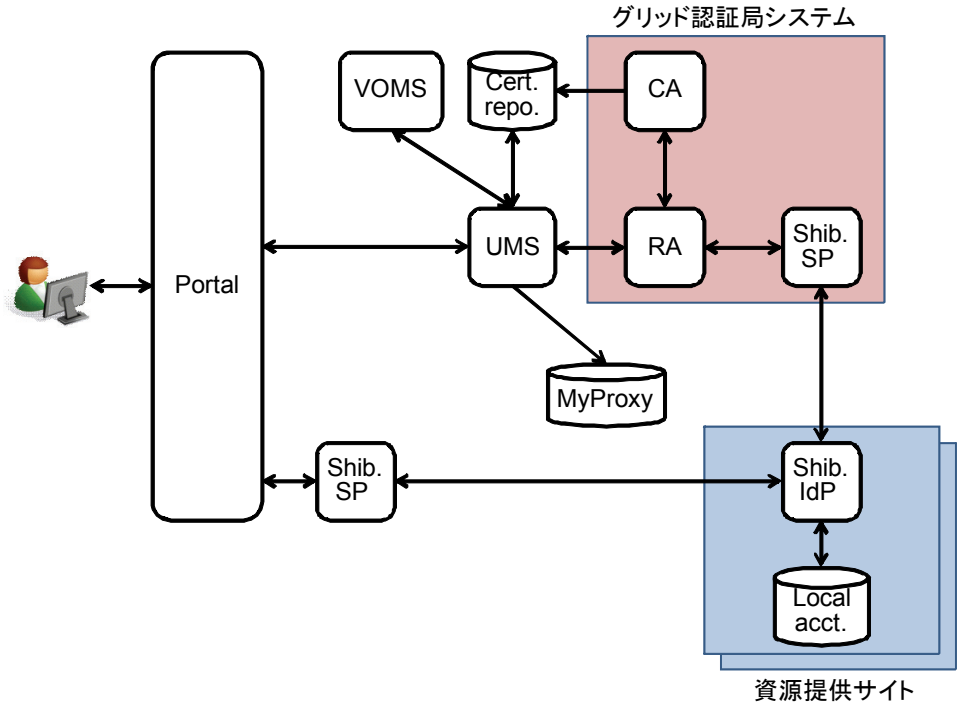

図 4　Shibboleth・グリッド認証連携システム

3.3. グリッド環境上でのユーザ管理技術

本研究では，情報基盤センターの認証とグリッド認証を連携させたシングルサインオン環境を構築し，その運用に必要な技術や業務フローの検証を行うことを目的としている．現在，大阪大学に設置されたグリッド認証局およびグリッドポータル，国立情報学研究所で運用されているグリッドサービス（情報サービスおよびジョブブローカリングサービス），大阪大学および九州大学に設置された計算ノードから構成される実証環境が構築されている．

3.3.1. ソフトウェアアーキテクチャ

図 4 は，本研究で構築した Shibboleth を用いた認証連携のためのソフトウェアアーキテクチャを示している．本環境では，ユーザは Shibboleth に対応した情報基盤センターのローカルアカウントを用いてポータル（portal）上でサインオンすることにより，グリッド用ユーザ証明書の取得，およびグリッド基盤上の資源利用の 2 つのサービスを受けることができる．

ユーザがグリッド用ユーザ証明書を取得するまでの手順は以下のとおりである．ここでは，ユーザが Shibboleth に対応した情報基盤センターのローカルアカウントを予め取得していることを前提とする．

(1) ユーザがローカルアカウントを用いてポータル（Portal）上でサインオンする．この際，Portal に接続された Shibboleth Service Provider (Shib. SP)が，ユーザが所属する資源提供サイトの Shibboleth Identity Provider (Shib. IdP)に問い合わせを行い，ユーザ認証が行われる．資源提供サイトの Shib. IdP は，サイトのローカルアカウントのデータベース（Local acct.）と接続されており，本データベースに登録されているユーザの認証を担当する．

(2) 認証終了後，ユーザがグリッド用ユーザ証明書発行のメニューを選択すると，User Management Server （UMS）が起動され，ライセンス ID が発行される．ライセンス ID は，次ステップでユーザが証明書発行手

続きを行う際に必要となる．UMS は，ユーザ証明書の取得や後述の Proxy 証明書発行等の処理を行うソフトウェアである[8].

(3) グリッド認証局システムは，登録局サーバ（RA）と認証局サーバ（CA）上で動作するソフトウェアから構成されている．RA は，Shib. IdP への問い合わせ（認証）後，ユーザが入力したライセンス ID を検証し，これらが正しい場合は，CA に対して証明書の発行を依頼する．CA は証明書を発行し，証明書リポジトリ（Cert. repo.）に格納する．

本グリッド基盤では，GSI を用いたグリッド認証を行っている．ユーザが本環境にシングルサインオンすると，ユーザ証明書から Proxy 証明書が作成され, Proxy 証明書用リポジトリ（MyProxy[9]）に格納される．これらの処理の手順は以下の通りである．

(1) ユーザがローカルアカウントを用いてポータル（Portal）上でサインオンする．Shibboleth を用いた認証手順はユーザ証明書取得と同じである．

(2) 認証終了後，ユーザが Proxy 証明書発行のメニューを選択すると，UMS が起動され，ユーザにユーザ証明書を活性化するためのパスフレーズおよびユーザが所属する VO 情報の入力を求める．

(3) パスフレーズの検証後，UMS は，VO 管理を行うソフトウェアである Virtual Organization Membership Service (VOMS[10])からユーザの VO 属性情報を取り出し, VO 属性つき Proxy 証明書を生成する．生成された Proxy 証明書は MyProxy に格納される．

[8] http://www.naregi.org/
[9] http://grid.ncsa.illinois.edu/myproxy/
[10] https://twiki.cnaf.infn.it/twiki/bin/view/VOMS/WebHome

3.3.2. 業務フロー

グリッド配備・運用タスクフォースでは，グリッド用ユーザ証明書および各情報基盤センターのローカルアカウントを一括申請するためのグリッドパックの運用実験を進めている．3.3.1 節が示した Shibboleth を用いた認証連携を実現することにより，グリッドパックにおける業務フローの一部をオンライン化することが可能である．

例えば，現在のグリッドパックでは，ユーザはグリッドパック利用のためのアカウント（GPID）を（情報基盤センターのローカルアカウントとは別に）取得する必要があるが，Shibboleth 認証連携機能を利用することにより，GPID の取得は不要となる．また，現在，ライセンス ID の発行では，グリッド認証局の運用者とユーザ間で電話等でのコミュニケーションをとる必要があるが，Shibboleth 認証連携機能を利用することにより，オンライン処理によりライセンス ID の発行が可能となる．

以上により，ユーザの負担を軽減するだけでなく，グリッド認証局や情報基盤センターの運用者の業務（主に書類処理）を大幅に削減でき，業務のスケーラビリティを向上することが期待される．現在，本タスクフォースにおいて，Shibboleth 認証連携を導入したグリッドパックの詳細な業務フローおよび必要なソフトウェアの仕様について検討中である．

4. これまでの進捗状況と今後の展望

本研究では，学際大規模情報基盤共同利用・共同研究拠点の計算機および SINET3 からなるグリッド基盤上でグリッドミドルウェアの配備・運用技術およびグリッド環境上でのユーザ管理技術に関する研究を行った．

グリッドミドルウェアの配備・運用技術については，Gfarm2 を用いたデータグリッド環境の配備方法を検討し，インストールソフトウェアの開発や実証実験を行った．また，SINET3 の 10Gbps

の帯域を持つネットワーク上でファイルを高速に転送するための技術に関する研究を行い，実証実験を行った．今後は，Gfarm2 ファイルシステムノード数を増やし，より大規模な実証実験を行う予定である．また，大規模データを扱うアプリケーションプログラムによる性能評価も進める予定である．

グリッド環境上でのユーザ管理技術に関する研究では，Shibboleth 認証連携技術を用いたグリッド認証システムの構築を行い，実証実験を行うとともに，本システムを用いたユーザ登録業務の業務フローを検討した．今後は，Shibboleth IdP を運用する資源提供サイトを増やし，より大規模な実証実験を行う予定である．また，ユーザ登録業務フローの詳細設計や，grid-mapfile 自動生成のためのソフトウェアツール等の整備を進め，運用実験を開始する予定である．

5. 研究成果リスト

(1) 国際会議プロシーディングス

[1] Taizo Kobayashi, Eisaku Sakane, Manabu Higashida, Hirofumi Amano, Kento Aida, Mutsumi Aoyagi, "Grid Operational Supports for Middleware Deployment and User Administration", International Symposium on Grids and Clouds (ISGC 2011), 2011. (投稿中)

(2) 国際会議発表

[1] Kento Aida, "Operational Issues in Inter-University Grid Infrastructure", 13th Teraflops Workshop, 2010.

(3) その他（特許，プレス発表，著書等）

[1] Inter-University Grid Infrastructure in Japan, SC10 Research Exhibits, 2010.