NAREGI Middleware

機能説明書

V 1.1

2011 年 3 月

国立情報学研究所

- **J2SSH Library**

  “This product includes software developed by SSHTools (http://www.sshtools.com/).”

- **openssl**

  “This product includes software developed by the OpenSSL Project for use in the OpenSSL Toolkit (http://www.openssl.org/)”

  “This product includes cryptographic software written by Eric Young (eay@cryptsoft.com)”

  “This product includes software written by Tim Hudson (tjh@cryptsoft.com)”

- **MyProxy**

  This product includes software developed by Computing Services at Carnegie Mellon University (http://www.cmu.edu/computing/).

  This product includes software developed by the NetBSD Foundation, Inc. and its contributors.

- **mod-ssl**

  “This product includes software developed by Ralf S. Engelschall <rse@engelschall.com> for use in the mod_ssl project (http://www.modssl.org/).”

- **NetBSD libnbcompat**

  “This product includes software developed by the NetBSD Foundation, Inc. and its contributors.”

- **gLite/gLite Security Utilities**

  “This product includes software developed by The EU EGEE Project (http://cern.ch/eu-egee/).”

- **VOMS/VOMS C API**

  “This product includes software developed by the EU DataGrid (http://www.eu-datagrid.org/).”

- **aica**

  “This product includes software developed by Akira Iwata Laboratory,Nagoya Institute of Technology in Japan (http://mars.elcom.nitech.ac.jp/).”

- **Xerces Java Parser , XML Security**

  “This product includes software developed by the Apache Software Foundation (http://www.apache.org/).”

はじめに

本書は NAREGI Middleware の機能説明書です。

NAREGI Middleware(以下、NAREGI ミドルウェア)は「超高速コンピュータ網形成プロジェクト(National Research Grid Initiative：通称 NAREGI)」において開発されたソフトウェアです。

NAREGI ミドルウェアは、分散した計算資源を管理するだけでなく、利用環境やプログラミング環境をも含めた、統合グリッド環境を提供いたします。NAREGI ミドルウェアを構成する多くのコンポーネントは、WSRF(Web Services Resource Framework)に準拠した Web サービスとして実装されています。NAREGI ミドルウェアは、これらコンポーネントを連携動作させ、より使い易く、より高度なグリッド環境を実現しています。

NAREGIミドルウェアについての詳細は、http://www.naregi.org/ をご覧ください。

- 本書の対象読者

  本書は、NAREGI ミドルウェアの導入を検討されている方、導入作業や運用保守を実施される方、利用される方を対象としています。
- ドキュメント体系

  NAREGI ミドルウェアのドキュメントは、次の分冊構成を取っています。
  - NAREGI Middleware 機能説明書(本書)

    NAREGI ミドルウェア全体の機能概要について説明しています。
  - NAREGI Middleware 導入手引書

    NAREGI ミドルウェア全体を通しての動作条件や、rpm パッケージを用いたインストール方法について説明しています。
  - NAREGI Middleware 使用手引書(管理者編)

    インストール後の NAREGI ミドルウェア全体を通して運用するにあたり、必要な事項について説明しています。NAREGI ミドルウェア各サービス個々の運用に関する詳細は、後述の NAREGI ミドルウェア AG および UG を参照ください。
  - NAREGI Middleware 使用手引書(利用者編)

    インストール後の NAREGI ミドルウェア各サービスを利用するにあたり、主に GUI を持つサービスについて操作方法を説明しています。NAREGI ミドルウェア各サービスの操作詳細については、後述の NAREGI ミドルウェア UG を参照ください。
  - 管理者ガイド NAREGI Middleware AG(Administrator Guide)

    NAREGI ミドルウェア各サービスの管理者向け使用手引書です。サービスを構成するコンポーネントごとに分冊構成を取っています。
  - 利用者ガイド NAREGI Middleware UG(User Guide)

NAREGI ミドルウェア各サービスの利用者向け使用手引書です。サービスを構成するコンポーネントのうち、利用者向け機能を持つコンポーネントごとに分冊構成を取っています。

- **重要**
  NAREGI ミドルウェア V1.1 の最新情報（サポートする OS バージョンなど）については、Readme を参照ください。

NAREGI ミドルウェアについてご不明な点がございましたら、NAREGI ミドルウェア FAQ も合わせて参照ください。

http://www.naregi.org/helpdesk/FAQ/index.html

目次

1. 概要

1.1 NAREGIとは

1.1.1 プロジェクト概要

NAREGI は大学・研究機関のスーパーコンピュータを連携させたサイエンスグリッドの実現を目指したグリッドミドルウェアの研究開発を目的としたプロジェクトです。サイエンスグリッドは、大学・研究機関での、スーパーコンピュータを駆使した学術研究をサポートする次世代研究環境の中核をなす技術です。NAREGI プロジェクトについての詳細は http://www.naregi.org/ を参照ください。

1.1.2 次世代研究環境

これからの研究では計算科学が不可欠なナノ・バイオの分野、計測・観測された膨大なデータの処理が不可欠な素粒子物理、天文の分野など多くの研究分野で、スーパーコンピュータを使った大規模で高速な処理が不可欠になります。そのため、スーパーコンピュータの利用者も、従来の限られた研究者から、より多くの広範囲の研究者に拡大します。また、研究はより高度で複雑になり、多くの要素が関係するために、研究者同士の連携が重要となります。このような学術研究をサポートする研究環境の必要性が増しており、NAREGI は、誰でも、どこからでも、必要な計算資源、実験設備が使えると共に、必要な研究者同士が連携して研究できる仮想組織(VO)等を実現する次世代研究環境の構築を目指しています。国立情報学研究所では次世代研究基盤としてネットワーク、認証、サイエンスグリッドからなるサイバー・サイエンス・インフラストラクチャ(CSI)の構築を進めています。

1.1.3 仮想組織（VO）

VO とは、複数の実組織にまたがって資源、データ、アプリケーションを共有し、相互に連携して活用するために形成する仮想的な組織です。

NAREGI ミドルウェア環境では、VO の構成を、当該 VO に提供された計算機資源ノード群と、当該 VO を管理するための管理用ノード群より構成します。利用者は VO の具体的な計算機資源構成を意識することなく、資源の利用、アプリケーションの共有が可能です。VO 運用に関しては、認証ポリシーが異なる VO 間でも相互に認証することで、セキュリティを確保し、相互に連携できます。利用者の VO 情報と資源使用量から課金処理することも可能です。

1.1.4 サイバー・サイエンス・インフラストラクチャ（CSI）

CSI は VO による研究コミュニティー形成、ドメインの知識蓄積、研究連携、研究資源共有により、より高度な研究を効率よく進める次世代研究基盤構築を目指しています。

CSI は学術情報ネットワーク(SINET3)をベースに、大学間連携を視野に入れたセキュリティを目指す学術認証基盤(UPKI)を SINET3 上に構築し、この上に NAREGI ミドルウェアによるサイエンスグリッド環境を構築します。

このサイエンスグリッドでは、研究者は所属する大学、研究機関などの実組織とは異なる、研究目的に応じた VO により研究コミュニティーを形成します。研究者は必要に応じて複数の VO に所属し、必要がなくなれば解散すると言うように動的に VO を運用でき、研究に応じた適切な体制を組むことができます。VO 内ではメンバーが計算資源、実験設備、データ、プログラムを共有することで、大規模シミュレーションの効率的な処理、連成解析シミュレーションなど高度な研究の促進、大規模データの効率的な利用が図れます。さらに、CSI の運用により研究者間の連携による知的資産の集積も期待されます。

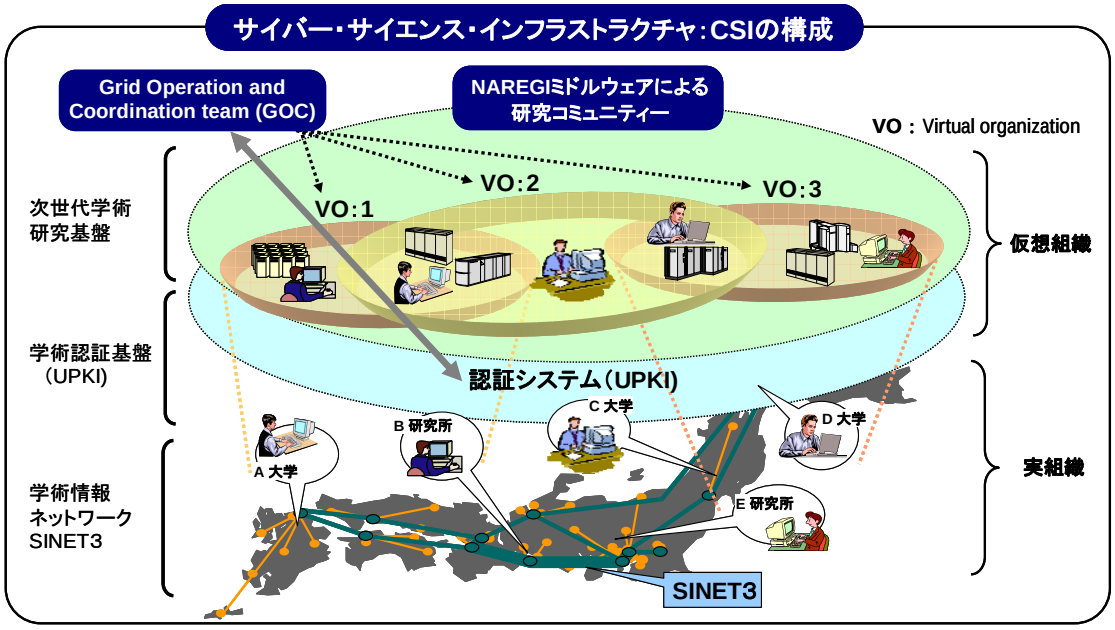


図 1.1-1 サイバー・サイエンス・インフラストラクチャ(CSI)のイメージ

1.1.5 今後の動向とサイエンスグリッドの展開

学術研究をはじめ、従来からの自動車や電機などの産業、新しいナノ、バイオなどの産業では、今後の研究開発に計算科学が不可欠です。スーパーコンピュータを利用する研究者は、従来のような特定の研究者から、より一般的な研究者に広がると考えられます。したがって、今後のスーパーコンピュータの利用は、今までのような狭い限られた環境の中だけで利用するのではなく、研究で必要とする規模、研究に合った方式の計算機を必要な時に、必要な場所から利用できることが必要です。

また、スーパーコンピュータの利用者、分野の拡大に伴い、計算コスト低減、計算のターンアラウンドタイム短縮が求められると考えられ、さらに研究の高度化に伴いマルチフィジックス、マルチスケールの計算ニーズが増加することは必須です。

サイエンスグリッドは研究に適した規模のスーパーコンピュータを選択して利用する、様々な方式のスーパーコンピュータを連携させて利用する、ポータルからスーパーコンピュータのことは意識せずに利用する、など基本的には今後のスーパーコンピュータの利用ニーズを実現できるものです。しかし、サイエンスグリッドを研究者にとって本当に有用なものとするには、様々なサポートが必要です。このようなサポート機関として Grid Operation and Coordination team（GOC）が考えられており、認証局の運用、グリッド証明書の発行、学術認証基盤との連携、さらに利用者サポート、利用者トレーニングなどのサポートを行います。サイエンスグリッドで連携する対象も大学・研究機関のスーパーコンピュータから次世代スーパーコンピュータ、研究室レベルのシステムまで広げることで、わが国のスーパーコンピュータ資源をより効率的に活用できる環境が期待されます。

これからの研究は日本だけの閉じた環境では、グローバルな研究開発の中で勝ち残ることはできず､サイエンスグリッド環境は世界とつながった環境であることが必須です。世界と繋がった研究環境を構築するために、NAREGI ミドルウェアは世界標準の仕様に合わせています。

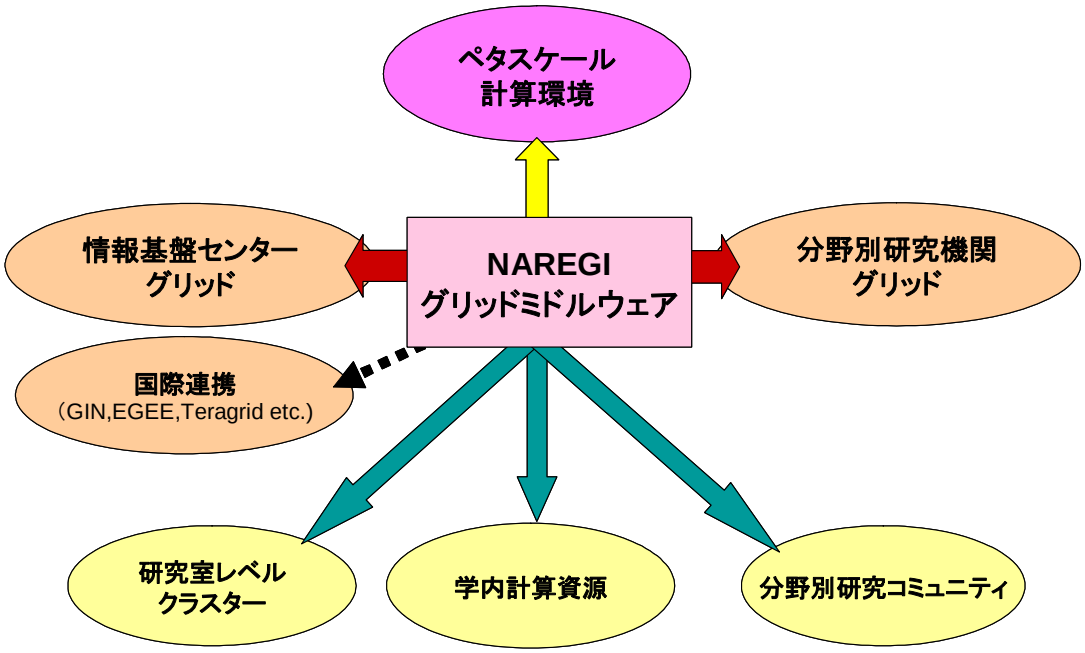


図 1.1-2 次世代研究環境としてのグリッドの展開

1.2 NAREGIミドルウェアとは

グリッドミドルウェアはネットワーク上の資源（複数かつアーキテクチャの異なる計算資源や広範に存在するデータなど）を仮想化するための要素技術から構成されており、これらの要素技術を実現したソフトウェアを組み合わせて連携させ、利用者がネットワーク上の資源を簡単かつシームレスに利用可能とすることがグリッドミドルウェアの目的の一つといえます。

一般的にグリッドミドルウェアの構成要素技術には、次のものが必要であると言われています。

- 利用者がシングルサインオンでグリッド環境を利用するための認証技術
- グリッド環境に接続された計算資源の情報を収集管理する資源管理技術
- 利用者の要求に沿った計算資源を探索し予約する技術
- グリッド上に分散したデータやプログラムを共有する技術

1.2.1 NAREGIミドルウェアを構成する要素技術

NAREGI ミドルウェアは、上記のグリッドミドルウェアの構成要素技術に加え、利用者の利便性を高めるためのアプリケーション利用環境、グリッド環境上のプログラミング環境などを実現する技術などが相互に連携し構成されている WSRF 完全準拠の Web ベースグリッドミドルウェアです。

NAREGIミドルウェアの構成要素技術（コンポーネント）には次のものがあります(図 1.2-1)。

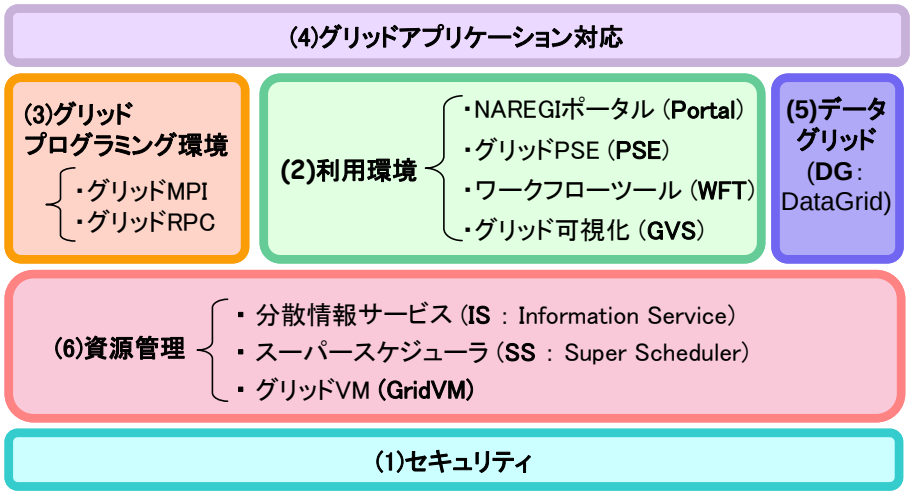


図 1.2-1 NAREGI ミドルウェアを構成するコンポーネント

(1) セキュリティ

セキュリティコンポーネントには GSI（Grid Security Infrastructure）を実装するための認証機能と認可機能があります。

認証機能は PKI（公開鍵暗号基盤）を用いて各種グリッド用証明書を発行・管理するための仕組みです。NAREGI が提供する NAREGI-CA（認証局ソフトウェア）は IGTF（International Grid Trust Federation）が定めた運用基準を満たし、国際連携が可能なグリッド用認証局の構築を可能にします。

認可機能は、実組織により異なる資源アクセスポリシーに対し、仮想組織（VO : Virtual Organization）を定義することで、組織を越えた資源の共有を可能にします。

セキュリティについての詳細は「2.1 セキュリティ」を参照ください。

(2) 利用環境

NAREGI ミドルウェアの利用環境は、NAREGI ミドルウェアの利用起点となる NAREGI ポータル、グリッド環境上でアプリケーションを簡単かつ効率的に動作させることを目的とした PSE (Grid Problem Solving Environment)、WFT (WorkFlow Tool)、および GVS (Grid Visualization System)の各コンポーネントから構成されています。

PSE は研究者が開発したアプリケーションをグリッド環境上に配置し、研究コミュニティーによる共有を支援します。WFT はジョブ実行制御を操作性の良い GUI で簡単に記述できます。GVS は計算結果をグリッド環境上で視覚化します。

利用環境についての詳細は「2.2 利用環境」を参照ください。

(3) グリッドプログラミング環境

広域に分散したアーキテクチャの異なる（異機種）計算資源が混合する環境下での並列処理プログラミングの容易化を目的に、GridRPC（Grid Remote Procedure Call）と GridMPI（Grid Message Passing Interface）を提供しています。

GridMPI はグリッド上での通信遅延を考慮した高性能かつ相互運用性の高い通信を TCP/IP レベル、MPI ライブラリレベルで実現します。GridRPC は耐障害性が高く、動的な資源利用を可能とするグリッドアプリケーションの容易な開発と高い実行効率を可能にします。

グリッドプログラミング環境についての詳細は「2.3 グリッドプログラミング環境」を参照ください。

(4) グリッドアプリケーション対応

グリッド環境における連成計算プログラムの実行支援を目的に、複数アプリケーショ

ン間での高度意味変換機能をサポートしたグリッド連成ミドルウェア（Mediator）および、複数アプリケーション間でファイルによる同期型データ転送を行う機能（SBC）を提供しています。

グリッドアプリケーション対応についての詳細は「2.4 グリッドアプリケーション対応」を参照ください。

(5) データグリッド

広域に分散するデータ資源をグリッド環境上で利用可能にすることを目的に、データグリッドを提供しています。データグリッドはデータグリッド資源管理システムとデータグリッドアクセス管理システムから構成されています。

データグリッド資源管理システムは、大規模シミュレーションなどから得られる大量のデータを共有ファイルとして格納し、グリッド上に分散した計算環境から利用可能とします。

データグリッドアクセス管理システムは、共有ファイルシステム上の大量のデータを計算環境まで効率的に転送します。この共有ファイルは WFT での設定により SS と連携して利用されます。

データグリッドについての詳細は「2.5 データグリッド」を参照ください。

(6) 資源管理

資源管理は、広域に分散された異機種混合の計算資源環境を 1 つにまとめて運用することを目的に、SS (Super Scheduler)、GridVM (Grid Virtual Machine)、分散資源情報サービス(以下、IS)のコンポーネントから構成されています。

SS は資源ブローカリング機能、ジョブワークフローエンジンなどを持ち、資源やジョブの管理を行います。GridVM は異機種の計算資源を仮想化して表現、統一されたインタフェースでの資源やジョブの管理サービスを提供します。IS はグリッド環境における計算資源、ネットワーク、ソフトウェア、アカウンティング等に関する情報の統合的な管理を行います。これら 3 つのコンポーネントの連携で、一般的なジョブ投入機能に加え、アーキテクチャの異なる計算資源へのコアロケーション機能、パラメータサーベイジョブに対応したバルクジョブ投入機能などを実現します。

資源管理についての詳細は「2.6 資源管理」を参照ください。

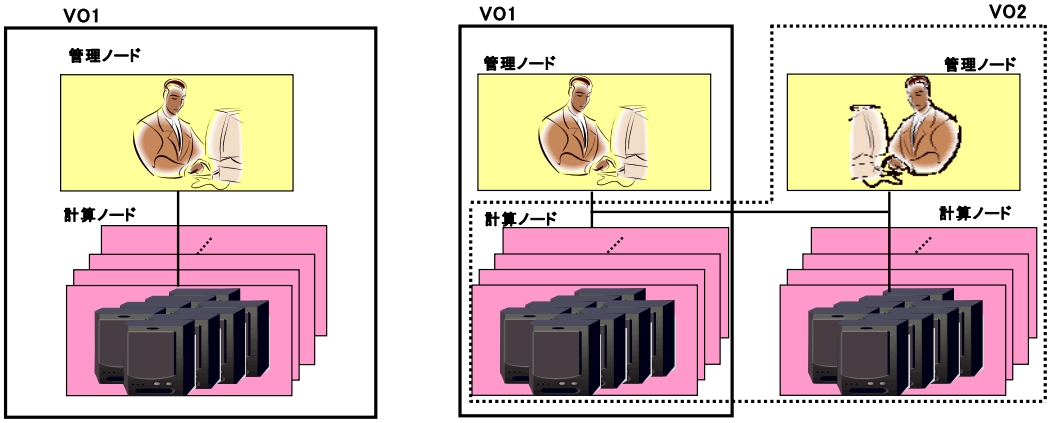
1.2.2 システム構成

ここでは、NAREGI ミドルウェアが取りうる VO の構成について説明し、次に、VO を構成する計算機（ノード）に搭載される NAREGI ミドルウェアのコンポーネントについて説明します。

(1) VOの構成

VO は人的資源と計算資源から構成されます。計算資源はさらに管理ノードと計算ノードから構成されます。管理ノードとは、NAREGI ミドルウェアの資源管理や情報サービス、データ管理などを行う計算機群を指します。計算ノードとは、ジョブ実行や資源情報の収集を行う計算資源群を指します。

NAREGIミドルウェアでは、各VOに一つの管理ノードがある構成（図 1.2-2 の(a)や(b)）を推奨しています。



(a) VO が単一の例　　(b) VO が複数の例

図 1.2-2 VO ごとに管理ノードがある構成

しかし例えばVO管理者が同一である場合など、複数のVOで管理ノードを共有することも可能です（図 1.2-3）。ただし管理ノードに負荷が集中することがあるので、管理ノードの設計には十分留意ください。

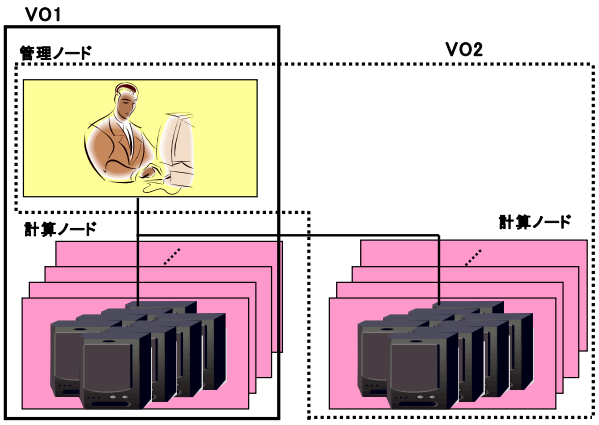


図 1.2-3 複数の VO で管理ノードを共有する構成例

管理ノードと計算ノードは、次の各ノードから構成されます。

各ノードの説明、および各ノードに搭載されるNAREGIミドルウェアのコンポーネントについては、次節「1.2.2(2) VOを構成するノード」を参照ください。

- 管理ノード

  UMS ノード、Portal ノード、SS ノード、IS-NAS ノード

- 計算ノード

  GridVM 管理ノード、GridVM 計算ノード、IS-CDAS ノード

(2) VOを構成するノード

VOを構成するノードと、各ノード上で動作するNAREGIミドルウェアのコンポーネントについて説明します。図 1.2-4は、VOを構成する計算機として管理ノード群が一つある例です（図 1.2-2 (a)に対応）。

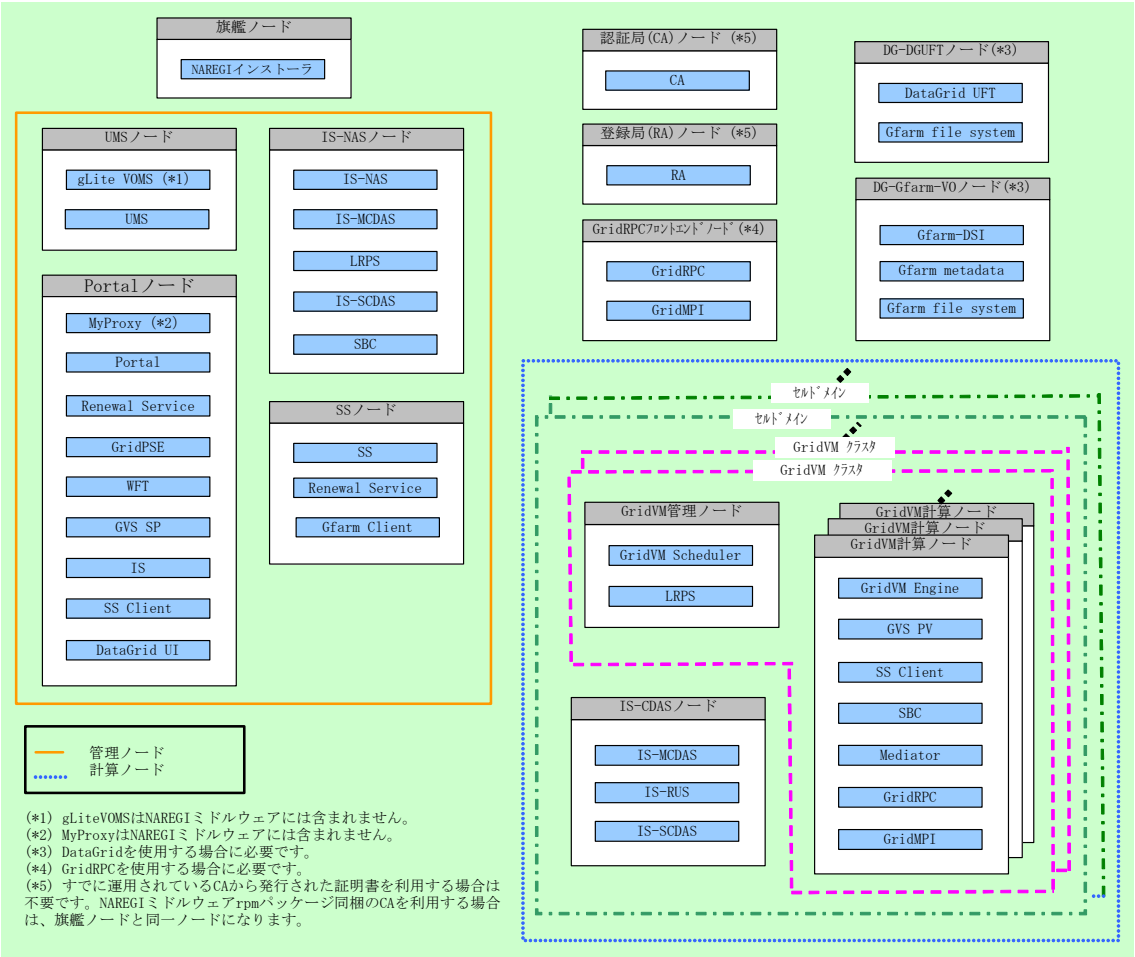


図 1.2-4 ノード構成例

図中において GridVM クラスタとは、GridVM 管理ノード 1 台が管理する、1 台以上の GridVM 計算ノードを指します。セルドメインとは IS が情報収集を行う単位であり、GridVM クラスタ 1 つ以上と IS-CDAS ノード 1 台、あるいは 1 組の IS-MCDAS ノードと IS-SCDAS ノードの構成をとります（IS-CDAS の冗長化機能を用いる場合は、同一セルドメイン内に複数の組の IS-CDAS ノードを置くことができます)。

(2-1) 旗艦ノード

NAREGI ミドルウェアを rpm パッケージを用いてインストールする場合に必要なノードです。

(2-2) CAノード、RAノード

認証局 (CA : Certificate Authority)ソフトウェア、登録局 (RA : Registration Authority) ソフトウェアが搭載されるノードです。NAREGIミドルウェアのrpmパッケージに含まれるNAREGI-CAを用いる場合は、評価・研究用の簡易的な証明書を発行することが出来ます。NAREGI-CAについては「2.1 セキュリティ」を参照ください。

(2-3) UMSノード

VO管理を行うVOMS(VO Membership Service)機能とユーザ管理を行うUMS(User Management Server)機能から成ります。VOMSはVO作成と削除、VO管理者の登録、VOユーザの登録などを行いますが、ユーザは意識する必要はなくVO管理者において管理されます。UMSはユーザ証明書の発行要求や管理などのユーザ管理機能を持っています。VOMS およびUMSについては「2.1 セキュリティ」を参照ください。

(2-4) Portalノード

利用者が Web ブラウザを用いて NAREGI ミドルウェアにサインオンする際、まず本ノード上の Web サーバにアクセスします。

利用環境コンポーネントである Portal、PSE、WFT、および GVS SP (GVS Service Provider)のほか、資源管理コンポーネントである IS や SS のクライアント部分、セキュリティコンポーネントである Renewal Service のクライアント部分が搭載されています。DataGrid の UI 部分も本ノードに搭載されます。

Globus Toolkit に含まれるMyProxyも本ノードに導入されます。My ProxyはCAから発行されたユーザ証明書の公開鍵と秘密鍵よりユーザ証明書の有効期間内で短い期間のプロキシ証明書を作成します。これによってシングルサインオンの実現や認証が必要なノ

ードへのアクセスが可能となります。利用環境については「2.2 利用環境」を、資源管理については「2.6 資源管理」を、セキュリティについては「2.1 セキュリティ」を、それぞれ参照ください。

(2-5) IS-NASノード

資源管理コンポーネント IS のサブコンポーネントである、IS-NAS (Node Aggregator Service)、IS-MCDAS (Monitoring Celldomain Aggregator Service)、IS-SCDAS および LRPS (Local Resource Provider Service)、グリッドアプリケーション環境のコンポーネントである SBC が搭載されるノードです。

IS-NAS は VO 単位で情報集約を行います。IS-MCDAS、IS-SCDAS、LRPS によって SS および PSE からの情報を受けて IS-NAS に情報を集約します。また、ジョブ監視および状態情報の収集の役割も担います。本ノード上の LRPS は、GridVM 管理ノード（後述）上の LRPS と役割が異なり、SS から登録されるジョブの状態情報を収集します。SBC は同期型データ転送機能を持ちます。

資源管理については「2.6 資源管理」を、グリッドアプリケーション環境については「2.4 グリッドアプリケーション対応」を参照ください。

(2-6) IS-CDASノード

資源管理コンポーネント IS のサブコンポーネントである IS-CDAS (IS-MCDAS および IS-SCDAS）と RUS が搭載されるノードです。

IS-MCDASサブコンポーネントは、セルドメイン単位で情報集約を行い現在参照可能な情報を扱います。一方、IS-SCDASサブコンポーネントは、現在参照可能な情報に加えて過去履歴情報の参照が可能です。RUSサブコンポーネントは、資源利用記録を扱います。資源管理については「2.6 資源管理」を参照ください。

(2-7) SSノード

資源管理コンポーネント SS のサーバ部分、およびセキュリティコンポーネントである Renewal Service が搭載されるノードです。SS は、ジョブ要求に応じた計算資源の確保、他のコンポーネント間でジョブ要求の仲介、ジョブの予約および管理、ある 1 つのジョブに対し複数のマシン資源を要求する機能などを提供します。

資源管理については「2.6 資源管理」を、セキュリティについては「2.1 セキュリティ」を、それぞれ参照ください。

(2-8) GridVMノード(GridVM管理ノード、およびGridVM計算ノード)

GridVM管理ノードには、GridVM のサブコンポーネントであるGridVM Schedulerが搭載されています。GridVM Schedulerは、GridVM計算ノード（後述）の制御を行います。また本ノードにはISのサブコンポーネントLRPSも搭載されます。本ノード上のLRPSはIS-NASノード上のLRPSと役割が異なり、計算資源（GridVM計算ノード）側のジョブの情報やCPUの情報等を収集します。資源管理については「2.6 資源管理」を参照ください。

本ノードに GridVM Engine サブコンポーネントを搭載することにより、「GridVM 管理ノード兼 GridVM 計算ノード」として動作が可能です。ただしノードの負荷は高くなりますので留意ください。また GridVM 以外のローカルスケジューラ (PBS professional や Sun Grid Engine)を GridVM と併用される場合、Sun Grid Engine 以外は NAREGI ミドルウェア導入前に導入しておく必要があります。

GridVM計算ノードには、GridVM のサブコンポーネントであるGridVM Engineの他、SSのクライアント部分、GVS PV（Parallel Viewer）、MediatorおよびSBC、GridMPIおよびGridRPCの各コンポーネントが搭載されています。本ノードはGridVM Schedulerの管理を受け、NAREGIミドルウェアの計算資源として動作します。資源管理については「2.6 資源管理」を参照ください。

本ノードには予約ジョブ（予め計算資源の利用時間を予約するジョブ）と非予約ジョブの混在投入も可能です。ただし予約ジョブの方が優先度が高いため、予約ジョブ投入時に実行中の非予約ジョブはキャンセルされます（ローカルスケジューラがPBS Professionalの場合）。非予約ジョブ実行を保証する環境を確保するには、GridVM Schedulerの設定で予約不可としたGridVM計算ノードを準備されることをお勧めします。NAREGIミドルウェアがサポートするジョブの種類については「1.2.4 ジョブの種類」を参照ください。

(2-9) DG-UFTノード、およびDG-DRMS-SVノード

それぞれ、DataGrid のサブコンポーネントである DG-DGUFT、DG-Gfarm-VO が搭載されたノードです。DataGrid は利用有無の選択が可能です。

DG-DGUFT は共有ファイルの登録・更新・削除などファイル管理を行うアクセス管理機能を持ちます。DG-Gfarm-VO は NAREGI ミドルウェア上での共有ファイルの割当て配置および使用容量を統一的に管理する資源管理機能を持ちます。

DataGridを用いると、グリッド環境上における共有空間の中で大容量のディスクの確保を行い、メタデータ（情報を付加したデータ）として登録が可能となります。また、ワークフローに基づき、共有ファイルをジョブの割当てノードにステージング（データ転送）することも出来ます。DataGridについては「2.5 データグリッド」を参照ください。

(2-10) GridRPCフロントエンドノード

NAREGI ミドルウェアから GridRPC を使用する際に必要になるノードです。GridRPC および GridMPI コンポーネントが搭載されます。

1.2.3 NAREGIミドルウェアの処理の流れ

NAREGIミドルウェア環境を利用する際の処理の流れを図 1.2-5に沿って説明します。

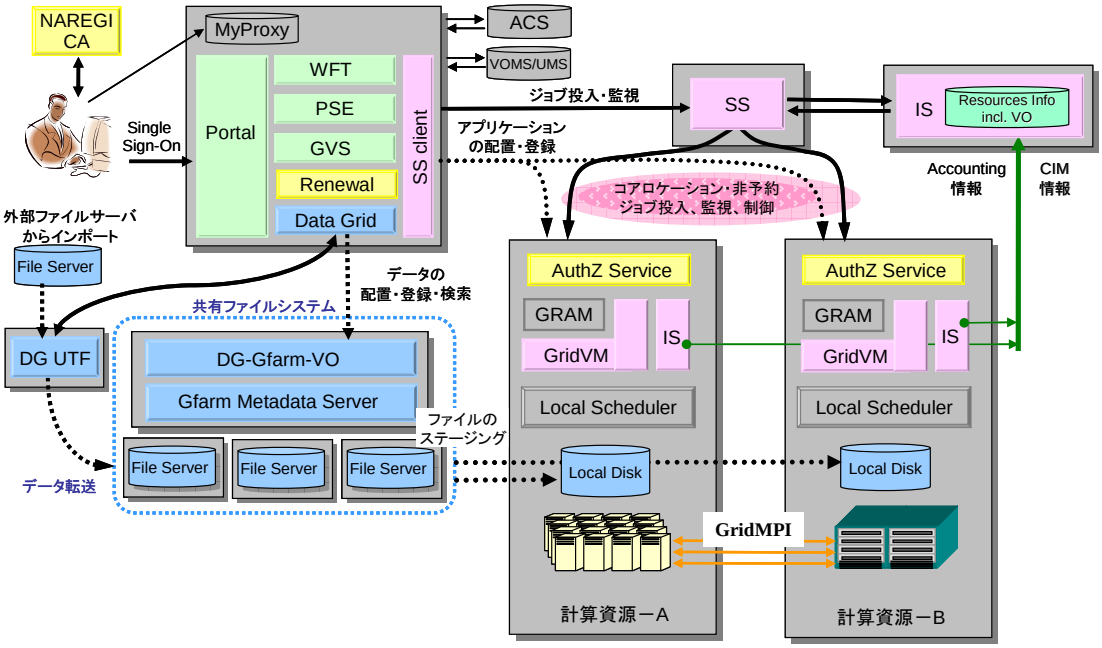


図 1.2-5 NAREGI ミドルウェアの処理の流れ

(1) ユーザ証明書取得とサインオン

NAREGI ミドルウェアを利用するには、まず Web ブラウザを用いて NAREGI Portal にアクセスしログインして、ユーザ証明書を取得します。この際、取得した証明書に所属する VO の属性を付加したプロキシ証明書が作成され、MyProxy サーバに預けられます。次に、NAREGI Portal からサインオンを行います。

ユーザ証明書の取得およびサインオンの操作方法は「NAREGI Middleware 使用手引書(利用者編)」を参照ください。

(2) アプリケーションの登録～配置

利用者は、NAREGIミドルウェア上で利用したいアプリケーションやプログラムをPSE

上のアプリケーションリポジトリへ登録、コンパイルし、利用可能な計算サイトへ配置します。この際、必要なデータやファイルを予めDataGridの共有ファイルシステム上にインポートしておけば、ジョブ実行の際に呼び出すように指定することもできます。PSEを用いたアプリケーションの登録～配置については「2.2.2 アプリケーションの登録・検索・更新 ～ 2.2.4 アプリケーションの配置（デプロイ）」も参照ください。

(3) ジョブ実行

(3-1) ワークフロー定義～ジョブ投入

WFT を用いてワークフロー（ジョブ実行手続き、入出力ファイルとの対応関係）とそれぞれのジョブ実行要件を記述します。

続いて、WFTからジョブを投入します。この際、ジョブを投入した利用者の プロキシ 証明書が、SSおよびGridVM に渡され、VO属性や利用者の認可判断の後、GridVMが管理する計算資源上で実行されます。WFTについては「2.2.7 ジョブ実行」も参照ください。

(3-2) ジョブ実行資源の探索と予約

各計算資源が属する VO の情報や CPU 数等の計算資源の条件は、あらかじめ IS に収集され、データベースに蓄積されています。

SS は投入されたジョブの実行要件と、利用者のプロキシ証明書の VO 関連情報、IS に集約された資源情報等とを照らし合わせて、実行要件に合致した計算資源を探索（資源ブローカリング）します。

コアロケーションジョブなど、資源使用時間の予約が必要なジョブ（予約ジョブ）の場合、適切な計算資源が見つかると、SS は GridVM を介して計算資源のローカルスケジューラに時刻を指定して予約をします。各計算資源上の GridVM は各資源に設定されたポリシーとジョブ要件を照らし合わせて、各資源における VO ごとの許可資源量を判断します。

なお、資源プロパティの設定により、予約を必要としないジョブ（非予約ジョブ）の場合はローカルスケジューラに直接投入されるジョブとの共存も可能です。NAREGIミドルウェアがサポートするジョブの種類については「1.2.4 ジョブの種類」を参照ください。

(3-3) ジョブ実行

予約ジョブの場合、予約した時刻になると、GridVM に予約しておいたジョブが起動

されます。非予約ジョブの場合、ジョブがローカルスケジューラのキューに投入され、ローカルスケジューラのスケジューリングポリシーにしたがってジョブが起動されます。

(4) 計算結果の可視化

計算途中の状況、結果は GVS で可視化し、確認できます。

(5) アカウンティング情報の収集

ジョブ実行に利用した計算資源の情報や利用量などのアカウンティング情報は、GridVM から IS に登録されます。この登録情報は IS の GUI から参照することができます。

1.2.4 ジョブの種類

NAREGI ミドルウェアがサポートするジョブには、次のものがあります。

- 予約ジョブ

  NAREGI ミドルウェアでは、MPI ジョブ、コアロケーションジョブを予約ジョブとして扱います。なおコアロケーションを必要としない MPI ジョブに関しては今後、非予約ジョブに変更予定です。コアロケーションジョブとは、複数のアプリケーションを連携させて解を求めるために複数の計算資源を同時使用するジョブを指します。予約ジョブは、その実行を確実なものとするために、非予約ジョブよりも優先度を高く設定しており、予約ジョブ実行時は、NAREGI ミドルウェアが該当の計算資源（ノード）を占有します。予約時間は SS が自動的に決定します。

- 非予約ジョブ

  利用時間を予約せず、運用ポリシーに応じて（First Come Fair Share等）処理するジョブです。MPIジョブ、コアロケーションが必要なジョブ以外は、既定で非予約ジョブとして投入されます。

  計算資源のプロパティを設定することにより、予約ジョブや、ローカルスケジューラに直接投入されるジョブとの共存が可能です。ただし予約ジョブの方が優先度が高いため、予約ジョブの投入時または実行時（ローカルスケジューラに依存）に実行中である非予約ジョブはキャンセルされます。ローカルスケジューラがPBS Professionalの場合、予約ジョブの投入時に実行中の非予約ジョブはキャンセルされます。このため、非予約ジョブの実行を保証する環境を確保するには、予約不可の計算資源を準備されることをお勧めします（表 1.2-1）。計算資源のプロパティは、GridVM SchedulerにおいてGridVMクラスタ単位で設定可能です。

表 1.2-1 ジョブの種類と計算資源のプロパティ

計算資源のプロパティ	NAREGI ミドルウェア経由で投入		ローカルスケジューラ直接投入ジョブ
	予約ジョブ	非予約ジョブ	
予約可	○（ノード占有）	Δ（ノード非占有）	×
予約不可	×	○（ノード非占有）	○

○はジョブ投入可、Δはジョブ投入可だがキャンセルされることがある、×はジョブ投入不可であることを表します。

1.2.5 運用モデル

VOによる研究コミュニティーの運用モデルについて 図 1.2-6に沿って説明します。

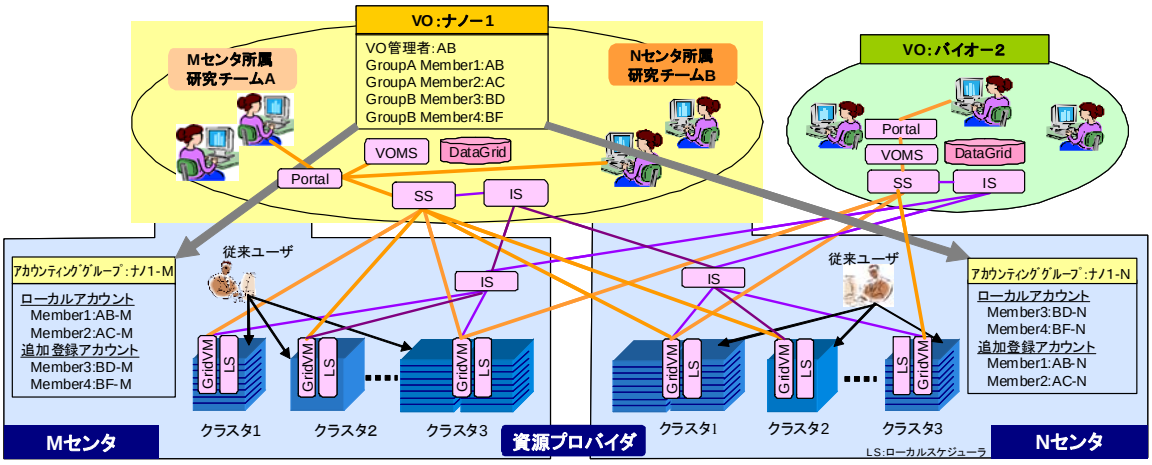


図 1.2-6 VO の作成例

図 1.2-6の例では、実機関 (RO：Real Organization)としてMセンタとNセンタがあり、VOとして「ナノー1」と「バイオー2」があります。「ナノー1」VOのメンバにはMセンタの研究チームAとNセンタの研究チームBが属しています。

以下に NAREGI ミドルウェアで構築した VO 環境での運用モデルとして、「ナノー1」VO の新規メンバが新規に M センタおよび N センタの計算資源を使用する場合について説明します。

(1) 事前準備

「ナノー1」VO の管理者（以下、VO 管理者）は VOMS に新規 VO メンバを登録します。この際、各新規ユーザのユーザ証明書が必要となり、この証明書は IGTF 公認の認証局から取得したものである必要があります（1 ユーザに付き 1 証明書。UPKI 運用後はその証明書を利用）。

(2) ユーザアカウントの準備

VO 管理者は、利用するセンタ（ここでは M センタおよび N センタ）の管理者に依頼し、新規 VO メンバ（以下、利用者）のためのユーザアカウント（計算資源上のアカウント）を準備します。

(3) grid-mapfileの準備

VO 管理者は準備したユーザアカウントを grid-mapfile に記述し、センタの管理者へ提供します。grid-mapfile とは証明書上のユーザ名と計算資源上のユーザ名を記述したファイルで、ユーザのマッピングを行うためのものです。センタの管理者は grid-mapfile を自センタの計算資源に配置します。

(4) NAREGIミドルウェアの利用

利用者はWebブラウザを用いてNAREGI Portalにアクセスし、所属するVOを選択してサインオン、NAREGIミドルウェアの各種機能を利用します。NAREGIミドルウェアの利用の流れは「1.2.3 NAREGIミドルウェアの処理の流れ」を参照ください。

2. NARAGIグリッドミドルウェアの機能

本章では、NAREGI ミドルウェア各サービスの持つ機能について説明します。

2.1 セキュリティ

NAREGI ミドルウェアにおけるセキュリティおよび認証機能は、認証局ソフトウェアである NAREGI-CA、VO メンバの管理システムである VOMS、ユーザ管理サーバ（以下、UMS）から構成されており、Globus で用いられている GSI（X.509 ユーザ証明書とプロキシ証明書を用いた権限委譲によるセキュリティシステム）を基盤としています。これらにより、利用者は NAREGI ミドルウェアをシングルサインオンで利用可能となります。

また、VO 間における連携のため、ジョブ実行時に SS と連携してプロキシ証明書の自動更新を行う Renewal Service 機能も持っています。

2.1.1 NAREGI-CA

NAREGI-CA は NAREGI ミドルウェアの認証局ソフトウェアであり、この NAREGI-CA で構築された認証局を NAREGI CA と呼んでいます。NAREGI ミドルウェア環境を使用する際は NAREGI CA が発行した証明書が必須となります。

NAREGI ミドルウェアの利用に際しては、証明書は正式に運用されている認証局より取得ください。ただし評価・研究用として試用される場合は、NAREGI ミドルウェアの rpm パッケージに含まれる NAREGI-CA を用いて証明書を発行することも可能です。CA と RA は、別ノード、同一ノードに導入することの両方が可能です。

(1) NAREGI-CAの主な機能

NAREGI-CA の主な機能には、次のものがあります。

表 2.1-1 NAREGI-CA の主な機能

機能名	概要
証明書要求作成	鍵ペアの生成と証明書要求(CSR)の作成を行います。
証明書要求への発行	証明書要求に署名を行い、証明書を発行します。
証明書要求登録と確認発行	証明書要求(CSR)を CA の CSR キューに溜め込み、キューの中にある CSR から証明書を発行することができます。リモートマシンからの証明書要求に対して即時に発行せず、一度 CA 運用者が確認した上で証明書の発行操作が可能です。
証明書の更新	証明書のシリアルナンバや公開鍵をそのままにして証明書の有効期限や拡張情報のみ変更できます。
証明書の一括発行	指定したフォーマットに従った CSV ファイルを使用することで、証明書の一括発行を行うことができます。
証明書の失効	証明書を発行した利用者が秘密鍵を紛失したり鍵の漏洩が起きた場合などは証明書の失効を行います。
証明書の失効解除	既に失効済みの証明書の状態から正常(失効を取り消した)の状態にします。

CRL の発行	証明書失効リスト(CRL: CertificateRevocationList)を発行します。CRL を発行することで、ユーザ証明書や署名の検証時に破棄状態のチェックを行えます。
証明書と秘密鍵のエクスポート	CA には証明書と秘密鍵が保管されており、シリアル番号を指定することでユーザ証明書またはユーザ秘密鍵のエクスポートが行えます。また、CA の CSR キューに保管されている証明書要求のエクスポートも行えます。
秘密鍵のインポート	CA にはユーザ秘密鍵を保管する機能があり、発行した証明書と秘密鍵のペアがそろっている場合にユーザ秘密鍵のインポートが行えます。
秘密鍵の削除	CA には発行した証明書とユーザ秘密鍵を保管する機能があります。このうちユーザ秘密鍵は、CA の鍵ストアから削除することが可能です。また、CA の CSR キューに保管されている証明書要求も削除することが可能です。
プロファイル設定の表示	CA が保持しているプロファイル設定の表示を行います。このコマンドを実行することで、CA の Issuer と Subject、指定したプロファイルのカレントシリアルナンバ、有効日数、拡張情報を表示します。
プロファイル設定の更新	CA が保持しているプロファイル設定の更新を行います。このコマンドを実行することで、指定したファイルのカレントシリアルナンバ、有効日数の設定を行うことができます。
プロファイル拡張情報の更新	証明書拡張情報はプロファイルごとに設定がおこなえます。証明書を発行するときに、指定したプロファイルに設定してある拡張情報を証明書に追加するため、ここで設定した情報がそのまま証明書発行時に反映されます。プロファイルの拡張情報の設定は、スクリーンに表示される設定情報に従い、ステップバイステップで行うことができます。
プロファイルの追加、削除と名称変更	１つの CA で複数のプロファイルを保持することができ、証明書を発行するグループ毎にプロファイルを用意したり、必要のなくなったプロファイルを削除することができます。また、プロファイル名の変更を行うことができます。
オペレータの追加と削除	CA をローカルマシンにて起動し、直接操作を行う場合は CA のマスタパスワードを入力することで CA の操作が可能ですが、リモートコンソールから CA サーバに接続して使用する場合は、必ず利用者(オペレータ)の認証が必要です。この CA サーバに接続可能なオペレータ証明書の発行、ID/Password やアクセス権限の設定が可能です。

(2) 証明書の登録および認証処理の流れ

下図は、NAREGI-CA で構築された CA(認証局)と RA(登録局)を用いた証明書の登録および認証処理の流れです。

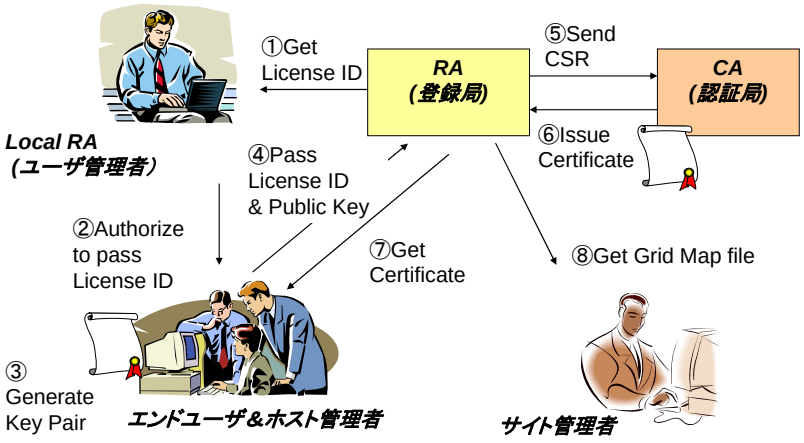


図 2.1-1 証明書の登録および認証処理の流れ

以下、図中の流れに沿って説明します。

① ユーザ管理者(Local RA)は、RA(登録局)よりエンドユーザのライセンス ID を取得します。
② ユーザ管理者(Local RA)は、エンドユーザの本人性を審査・確認し、OK なら①で取得したライセンス ID を認可します。
③ エンドユーザ(もしくはホスト管理者)は、ユーザ管理者(Local RA)から認可されたライセンス ID を取得します。
④ エンドユーザ(もしくはホスト管理者)はペアキー(秘密鍵、公開鍵)を作成し、ライセンス ID と一緒に RA(登録局)に渡します。
⑤ RA(登録局)は CA(認証局)に対し証明書を要求します。
⑥ CA(認証局)は証明書の発行を行い、管理します。
⑦ エンドユーザ(もしくはホスト管理者)は発行された証明書を RA(登録局)より取得します。
⑧ サイト管理者は RA(登録局)より grid-mapfile を取得し、管理、運用を行います。

2.1.2 VOMS

NAREGI ミドルウェアでは、EGEE で開発されている gLite VOMS をベースに VO 情報を証明書の中に入れ込むことで VO 間のセキュリティをはかっています。

VOMS は VO 新規作成や追加／削除、VO 管理者の登録、VO へのユーザ登録／削除、VO に登録したユーザの認可属性(Group や Role、Capability)管理などの機能を持っています。また、VO に登録したユーザに VO 属性付きプロキシ証明書を発行します。VO 属性付きプロキシ証明書は、ユーザが属する VO の情報(認可属性)を持たせた証明書です。ユーザがジョブを実行する際、計算資源はこの認可属性をもとに認可を行います。

UMS（後述）を用いる場合、各ユーザごとの設定ファイル(vomses ファイル)を UMS に設定します。vomses ファイルは、VO がどの VOMS サーバで管理されているかなどの情報を記述しているもので、プロキシ証明書のライフタイムや、Portal の自動タイムアウト時間間隔、NAREGI ミドルウェアのセッションタイムアウト時間間隔の設定が可能です。

2.1.3 UMS(User Management Server)

UMS は、ユーザ証明書の発行要求や証明書の保管、VOMS から VO 属性付きプロキシ証明書の取得、プロキシ証明書の証明書リポジトリ（MyProxy サーバ）への格納（預け入れ）など、証明書の一元管理を行います。これにより、利用者自身がユーザ証明書・秘密鍵の管理を行う必要がなくなります。

2.2 利用環境

利用者がNAREGIミドルウェアを使用するためのインタフェースとして、NAREGI Portalを提供しています。また、NAREGIミドルウェアへのアプリケーション登録からジョブ実行までを、直感的なGUIで支援するグリッドアプリケーション利用環境（機能群）、ISに集約された各種情報表示の機能、データグリッド起動画面も提供しています。

2.2.1 NAREGI Portal

NAREGI Portalは、利用者とNAREGIミドルウェアとのインタフェースとなる機能です。利用者はPortalノードのwebサーバ上のportal画面からNAREGIミドルウェアにシングルサインオンすることで、さまざまな機能（グリッドツール）を呼び出してシームレスに利用することができます。シングルサインオンするために必要なユーザ証明書の取得やプロキシ証明書の登録もPortal画面から行います。

以下にNAREGI Portalの提供する画面と機能について説明します。

(1) サインオン／サインアウト画面

利用者がNAREGIミドルウェアへシングルサインオン／サインアウトするための画面です。

(2) グリッドツール起動画面

NAREGIミドルウェアのさまざまな機能（Grid Tools）を起動する起点となる画面です。規定ではGridアプリケーション環境のPSE、WFT、GVSのほか、資源管理コンポーネントのIS、データグリッドが設定してあります。

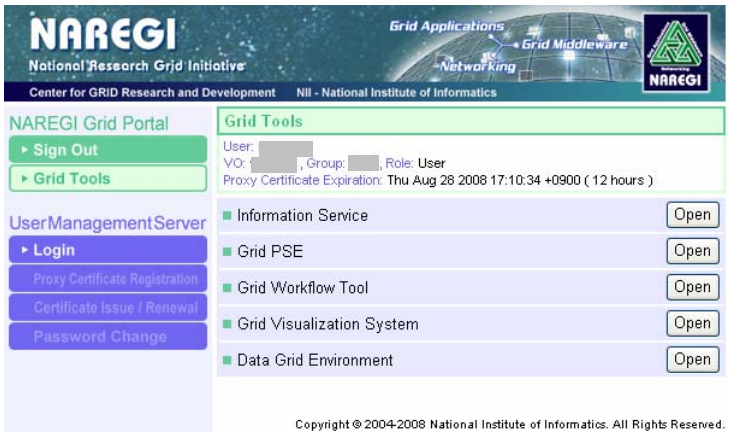


図 2.2-1 Grid Tools 画面

(3) ログイン／ログアウト画面

利用者がユーザ証明書を取得するためのUMS（2.1.3参照）へログイン、またはログアウトするための画面です。

(4) 証明書取得／登録画面

利用者が NAREGI ミドルウェアの機能をシームレスに利用するには、利用者を判別するためのユーザ証明書や、VO をまたがって使用するためのプロキシ証明書が必要になります。portal はこれら証明書の取得や登録を行うための画面を提供しています。UMS サーバ上のユーザパスワードを画面上から変更することもできます。

(5) カスタマイズ

Portal 画面は、設定ファイルの編集により VO ごとに画面構成をカスタマイズすることが出来ます。カスタマイズ可能な項目には次のものがあります。

・Grid Tools	利用できる Grid Tools の追加や削除、オープン方法を指定します。
・サービスの設定	利用できるサービスの情報として、次の項目を指定します。詳細は Portal の管理者向け操作手引書、および Renewal Service の管理者向け操作手引書を参照ください。 ➢ SS ノードのホスト名と port 番号（既定では http://<SS ノードの FQDN>:8080） ➢ Renewal Service が稼働するノード（SS ノード）のホスト名と port 番号 ➢ ・IS ノードの URL（既定では https://<IS ノードの FQDN>:8443/wsrf/services/org/naregi/infoservice/aggregator/node/factory/NAFS）
・タイトル画像	Grid Tools 起動画面やサインアウト画面のタイトル画像を指定します。
・表示色	Grid Tools 起動画面やサインアウト画面の表示色を指定します。

2.2.2 アプリケーションの登録・検索・更新

利用者が作成したアプリケーションは、NAREGI ミドルウェア上に登録したり、削除することができます。また、他利用者が登録したアプリケーションを検索することもできます。

(1) アプリケーションの登録

利用者は、利用環境を構成するコンポーネントのうち PSE を用いて、NAREGI ミドルウェア上のアプリケーションリポジトリ（ACS(Application Contents Service)）へアプリケーションを登録することができます。

登録の際は、プログラム名やその実行に必要な入力ファイル等の他、アプリケーションの説明、アプリケーションの実行に必要なリソース要件、登録するユーザ名や所属 VO 名などが入力できます。

なお、リソース要件は資源情報管理の IS との連携により、予め IS が収集した計算資源の情報を参照して登録先を絞り込むことが可能です。登録するアプリケーションがリソースに条件（OS、プロセッサ種別、メモリ量等）を持つ場合に有効です。

その他、アプリケーションバイナリを計算機上に配置や配置後に実行するスクリプトを登録したり、あとで検索が可能なようにアプリケーション検索のためのキーワードを登録することもできます。

入力された情報は実行時に必要な JSDL(Job Submission Description Language)に反映されて PSE 上に保存されます。必要に応じてコンパイルやテスト実行のためのスクリプトも併せて登録します。

なおアプリケーションの登録は NAREGI Portal との連携により、クライアント PC 上のコマンドラインツールから行うこともできます。

(1-1) 登録するアプリケーションの形式

入力データやパラメタデータと共に ZIP ファイル形式とし、PSE の稼動するノード（ノード構成例図の場合 Portal ノード）にアップロードします。この際、クライアント PC(ローカル環境)からのアップロードか、リモート環境からのアップロードか選択が可能です。登録可能なアプリケーションはソースファイル、実行モジュール(バイナリ形式)、インストール済みアプリケーション(Pre-Installed Application Binary)の三種類です。

(1-2) WFTとの連携

アプリケーション実行時の入出力ファイルや実行に必要なパラメータの流れを記述した手順（ワークフロー）を予め WFT で作成しておき、PSE に登録することが可能です。アプリケーションの開発者と実行者が異なる場合でも、容易にアプリケーションを実行可能とすることを目的とした機能です。

(1-3) DataGridとの連携

アプリケーション登録時に、アプリケーション実行時に必要なデータファイルとの関連付けを、アプリケーション情報の一部として登録することが可能です。この際、DataGrid との連携により、関連付けるデータファイルを DataGrid 上に登録すると共に、アプリケーション情報で関連付けることが可能です。

(1-4) アプリケーションの複製登録

予め登録されたアプリケーション情報を雛形として、新たなアプリケーションを登録することが可能です。これにより、共通な項目を入力する手間を省くことができます。

雛形となるアプリケーションの選択する際は、アプリケーション検索インタフェースを利用することで、利用者の属する VO に応じた適切なアプリケーションを選択することが可能です。

(1-5) 複数アプリケーションの一括登録

一つのソースツリーから複数の実行モジュールが生成される場合、一回の登録で複数のアプリケーションとして登録できます。なおこの際、コンパイルは一括して行われます(「2.2.3 アプリケーションのコンパイル」参照)。

(1-6) 個人利用／VO内共有アプリケーションの登録

個人利用アプリケーションとは、あらかじめ登録したユーザだけが参照・更新・コンパイル・デプロイ・削除できるアプリケーションです。これによりアプリケーション開発者個人だけが利用するアプリケーションの登録が可能です。

VO 内共有アプリケーションとは、同一 VO に所属する利用者が参照・コンパイル・デプロイ・アンデプロイできるアプリケーションです。

(2) アプリケーションの検索

検索項目として、「アプリケーション名」、「ユーザ名」、「グループ名」、「概要」等で検索が行える他、検索キーワードの指定より、アプリケーション候補を検索することができます。この時、検索キーワードはカンマ区切りで複数指定することもでき、複数キーワードに対しては AND 条件なのか OR 条件なのかを指定します。また、リソース要件に応じて利用可能な計算リソースの参照が可能です。

なおアプリケーションの検索も、登録と同様に、クライアント PC 上の Portal コマンドラインインタフェースから行うこともできます。

(3) アプリケーションの更新

あらかじめ登録されたアプリケーション情報を更新する際に、アプリケーションを構成するソースファイルやバイナリファイルの更新を指定することができます。更新対象となるアプリケーションの選択に際しては、アプリケーション検索インタフェースを利

用し、利用者の属する VO に応じた適切なアプリケーションが選択できます。

2.2.3 アプリケーションのコンパイル

登録したアプリケーションから、リソース要件（OS、プロセッサ種別等）にあわせたバイナリを生成します。また、実行するコンパイルスクリプトの登録が可能であり、動作確認の実行を行うことができます。

(1) 複数アプリケーションのコンパイル

１つのアプリをコンパイルする場合、それに紐付く他の全ての一括登録アプリも一括コンパイルします。この時、コンパイルする順番を決めます。

(2) 関連アプリケーションの自動コンパイル

アプリケーション登録時に、予め他のアプリケーションとの参照関係（例えば他アプリケーションを呼び出して実行するなど）を登録することができます(図 2.2-2)。これにより、アプリケーションをコンパイルする利用者が意識する必要なく、参照先のアプリケーションを自動的に配置して実行します。アプリケーションを登録した利用者とコンパイルする利用者が異なる場合などに有効です。

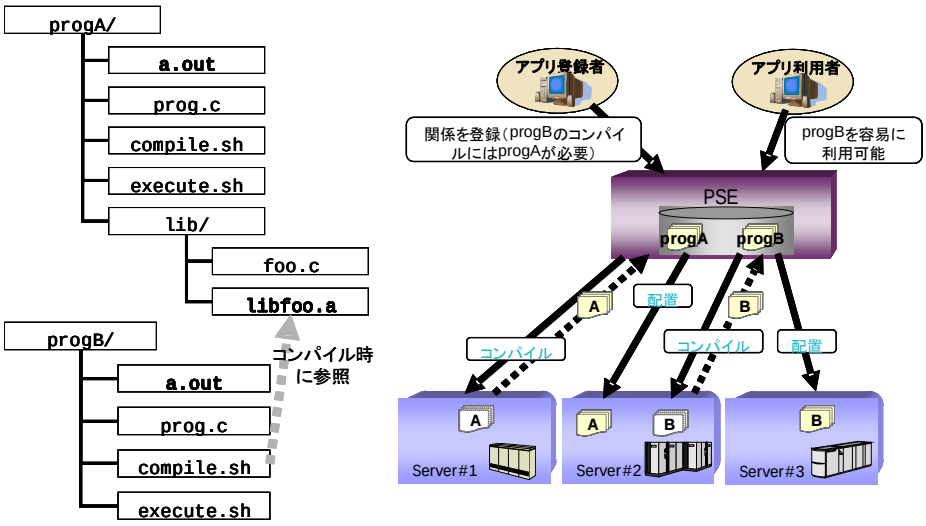


図 2.2-2 関連アプリケーションの自動コンパイル

(3) テスト実行

アプリケーションの正常動作や、環境に特化した処理実行の動作確認テストを事前に

実施する機能です。

2.2.4 アプリケーションの配置（デプロイ）

アプリケーションリポジトリに登録されたアプリケーションバイナリを計算機上に配置(デプロイ)します。

配置には、ユーザのホームディレクトリ配下にデプロイする「非共有デプロイ」と VO 内共有ディレクトリにデプロイする「共有デプロイ」があります。ただし、「共有デプロイ」は VO 毎のポリシー設定と計算機毎のポリシー設定により許可された場合のみ利用可能です（下図)。VO 内共有ディレクトリに共有デプロイしたアプリケーションバイナリは、同一 VO に所属する利用者がジョブ実行に利用することができます。

	共有設定ホスト	非共有設定ホスト
共有設定VO	①共有ディレクトリにデプロイ	②非共有ディレクトリにデプロイ(※)
非共有設定VO	③非共有ディレクトリにデプロイ(※)	③非共有ディレクトリにデプロイ(※)

※非共有ディレクトリは、デプロイユーザのHomeディレクトリ内に設置。

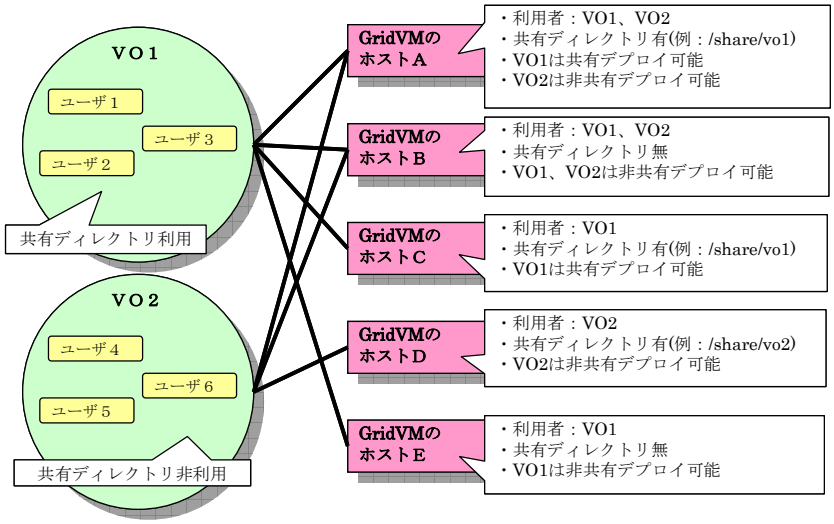


図 2.2-3 デプロイしたアプリケーションの VO 内共有

またコンパイル時と同様、他アプリケーションとの参照関係が登録されているアプリケーションは、利用者が意識する必要なく、参照先のアプリケーションが自動的に配置されます。

(1) 複数リポジトリによる配置

計算機の台数やネットワーク上の配置に応じて、複数のアプリケーションリポジトリ(ACS)を設けることができます。これにより、サイトをまたいだアプリケーション配置においても、多数の計算機にアプリケーションを効率良く配置することができます。計算機が複数のサイトに分かれて配置された場合、ACS サービスとアプリケーション配置サービスを各サイトに設置することで、サイト間で最小限のファイル転送回数で効率よくアプリケーションを配置できます。

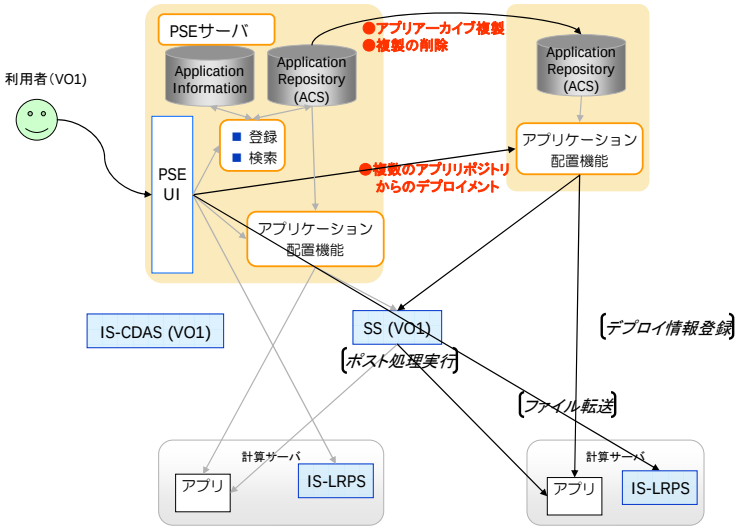


図 2.2-4複数リポジトリからのデプロイメント

(2) テスト実行

コンパイル時と同様に、アプリケーションの正常動作や、環境に特化した処理実行の動作確認テストを事前に実施する機能です。

2.2.5 ワークフローのインポート機能

WFT と連携し、アプリケーションのワークフロー(後述)をインポートする機能です。

2.2.6 状態取得

アプリケーション登録やコンパイル、デプロイ中のステータスを画面上から確認することができます。

2.2.7 ジョブ実行

ジョブ実行には、プログラムやデータ、I/O 関係をアイコンとして GUI 表現し実行する方法と、コマンドラインツールからコマンドベースで実行する方法があります。以下、それぞれの方法について説明します。

(1) ワークフローからのジョブ実行

プログラムやデータ、それらの I/O 関係をアイコンとして GUI 表現し、実行することが出来ます。さらにこれらの実行順序や関連（ワークフロー）もアイコンとして表現することが出来ます。これらにより、直感的なプログラムの実行が可能となります。

アイコンは、Grid MPI、コアロケーション、連成ジョブ用アイコンなど、一つのアイコンで同時に複数のジョブ実行ができる大規模サイエンス向けのものが予め用意されています。

作成したワークフローを NAREGI ミドルウェア上からジョブ投入（submission）すると、あとはワークフローの GUI 上で実行状態の監視を行います。初期（initial）状態では黄色、計算が始まると青になり、計算が終了すると緑、異常終了すると赤という具合に実行状態のモニタリングができます。

(1-1) アイコンの種類

NAREGI ミドルウェアで用意しているアイコンには次の種類があります。ワークフローの定義には、これら作成したアイコンを使用します。

アイコンの種類	説明
ワークフローアイコン	複数のプログラムやデータをプログラムアイコンやデータアイコンとして関連付け、複数のシステム上で連携して実行するように定義したワークフローをアイコン化したものです。ワークフローフォルダに保存することができます。
プログラムアイコン	実行するプログラムごとに実行要件や入出力情報を定義したものです。複数計算アイコンにより、パラメータスタディなど、パラメタを変えて同時に実行したいプログラムの定義を支援します。
データアイコン	計算機上のデータファイルをアイコンで表現したものです。
アレイジョブアイコン	本アイコンを使用したワークフローにより、ローカルスケジューラのアレイジョブ機能を用いたバルクジョブ（複数同時に実行したいプログラム）の実行を支援します。
GridMPI アイコン	GridMPI を使用したプログラム群の定義に使用します。ワークフローアイコンや Co-Allocation アイコンから呼び出すことができます。本アイコンにより、パラメータスタディなど、パラメタを変えて同時に実行したい場合の複数実行を支

アイコンの種類	説明
	援します。
Co-Allocation アイコン	ワークフローを構成するプログラム群を、同時に資源割り当てを行い実行開始する場合に使用します。つまり本アイコンで定義された各プログラムアイコンは、対応する全プログラムが同時に実行開始されます。 ワークフローアイコンから呼び出すことができ、他のアイコンと同様にワークフロー定義で使用することができます。
フロー制御アイコン	条件分岐アイコン・ループ実行アイコンの 2 つがあります。それぞれ、通常のプログラムにおける条件文や、ループ文のように、実行するプログラム群のフローを制御することができます。

(1-2) ファイル操作

ワークフローの定義時に、クライアント PC と管理ノード計算機間でファイルの転送や編集ができます。

ファイル操作	説明
ファイルのアップロード／ダウンロード	クライアントマシン上で作成したデータファイル等を、管理ノード計算機上にアップロード／ダウンロードします。アップロードしたファイルはワークフロー作成時にデータアイコンとして定義し利用可能することができます。
管理ノード計算機のデータファイルの編集	ワークフロー編集画面や Co-Allocation ジョブ編集画面上のデータアイコンに定義されている、管理ノード上のデータファイルを編集することができます。

(2) ジョブの実行

ジョブの実行は、ワークフローを用いる方法に加えて、クライアント PC に導入したコマンドラインツールからも可能です。この場合、NAREGI ミドルウェアへのサインオンはサインオンツールを用いて行われます。

投入したジョブの状態はジョブ一覧画面から確認することができます。また、ジョブを構成するアイコンを用いた実行状態とログを Monitor 画面から確認することができます。これらの画面は定期的に更新されます。

(3) デバッグ機能

WFT が提供するデバッグ機能は次のものがあります。

ファイル操作	説明
ステップ実行	デバッグ実行画面上のプログラムアイコンを 1 つ選択し、1 つずつ実行します。選択されたプログラムアイコンの実行終了後、実行が停止し、後続のアイコンは実行されません。
ネクスト実行	デバッグ実行画面上のワークフローアイコンまたは制御アイ

ファイル操作	説明
	コンを１つ選択し、入れ子ワークフロー内をまとめて実行する。選択されたアイコンの実行終了後、実行が停止し、後続のアイコンは実行されない。入れ子ワークフロー内に設定されているブレークポイントは無視する。
ブレークポイント設定／解除	デバッグ実行画面上のアイコンにブレークポイントを設定する。または、ブレークポイントが設定されているアイコンのブレークポイントを解除する。
部分実行	現在の実行停止位置から次のブレークポイントまで実行する。ワークフローアイコンの入れ子ワークフローにブレークポイントがある場合、入れ子ワークフロー内のブレークポイントで停止する。
コンティニュー実行	現在の実行停止位置より後続のアイコンを全て実行する。後続のアイコンに設定されているブレークポイントは無視する。
デバック実行キャンセル	デバッグ実行中のジョブをキャンセルする。
デバッグ実行保存	デバッグ実行画面のジョブを保存する。デバッグ実行画面を閉じる際に、保存確認ダイアログが表示され、[Yes]ボタンを押下した場合にデバッグ実行中のジョブを保存し、デバッグ実行画面を閉じる。[No]ボタンを押下した場合はデバッグ実行キャンセルを実行し、デバッグ実行画面を閉じる。[Cancel]ボタンを押下した場合は保存、キャンセルは行わず、デバッグ実行画面も閉じない。保存したジョブはジョブ一覧画面の[Status]に表示されている実行状態の文字列の後に“(Debug)”が付加されているジョブを選択することで再度呼び出すことが可能。
レスキュー実行	実行状態画面において、通常のジョブ実行(デバッグ実行ではない)が異常終了したワークフローを開き、異常終了したアイコンのプログラム、または入力データを修正後、異常終了したアイコンから再実行する。実行は簡易 WF エンジンではなく、通常実行を行う実行管理部で行われる。

(4) ワークフロー環境の退避と復元

利用者が Main 画面の Folder フォルダ以下に作った一群のワークフローアイコンを、クライアントマシンに退避したり、クライアントマシン上に退避した利用者環境ファイルを WFT 上に復元することができます。

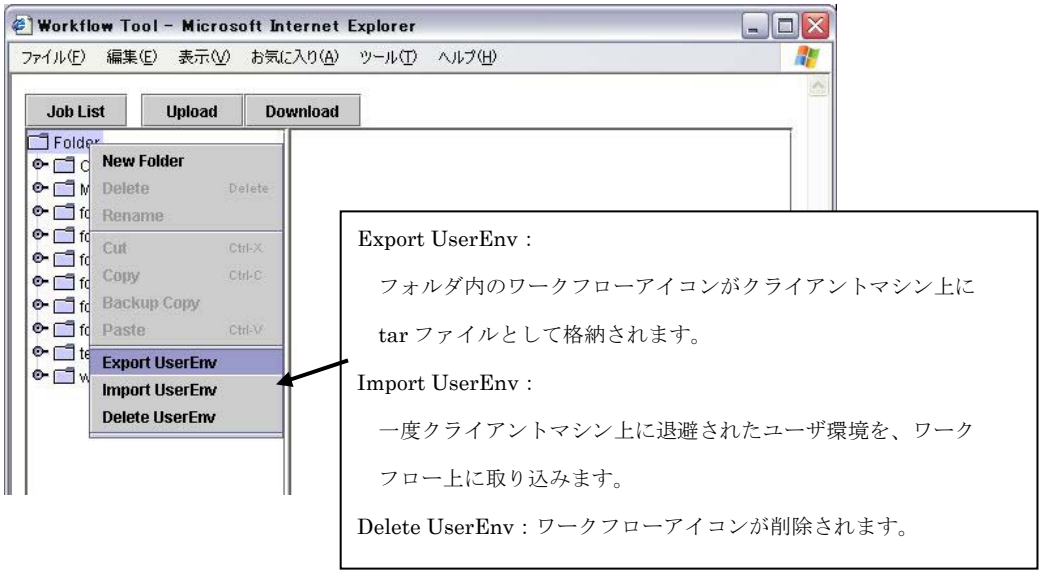


図 2.2-5 ワークフロー環境のフォルダによる操作

2.2.8 コマンドラインツール

これまで述べてきた利用者環境の機能のうちジョブ操作関連のいくつかは、クライアント PC 上に導入したコマンドラインツール（CLT：Command Line Tool）から実行することができます。コマンドラインツールから実行可能な主な操作は次のとおりです。

- クライアント PC 上のアプリケーションやワークフローの登録
- 登録済アプリケーションやワークフローの検索および取得
- アプリケーション配置可能な計算資源の検索
- 登録済アプリケーションの計算資源への配置
- ジョブの実行要件を定義し、ジョブ実行
- WFT 管理下にあるジョブ一覧、およびジョブ実行状態の表示
- 実行ジョブのキャンセル

2.2.9 GVS (Grid Visualization System)

GVS についての実行イメージを説明します。Client は利用者 PC 上で起動される GUI、Visualizer は計算サーバ上で起動される可視化処理本体です。Service Provider は利用者の PC 上からのアクセスポイントであり、Visualizer との間の通信を中継するほか、別の Service Provider と連結して複数の Visualizer を並列協調動作させる機能ももちます。

複数サイトに跨る連成計算や並列計算においては、計算結果データは複数サイトに分散して保存されており、それぞれ独立に計算サーバ上で可視化され、複数の可視化画像

は圧縮送信され、1 枚の画像に矛盾なく合成されて利用者の PC 上に表示されます。

利用者 PC とサイトの間のみならず、サイト間においても巨大な計算結果データの通信は一切不要であり、通信されるのは可視化パラメータと圧縮画像といった小サイズのデータのみとなっています。

GVS の特徴は、計算サーバ上での並列可視化により、メモリ容量、処理速度の限界を打破することができ、計算ノード数の許す限りどんな大規模な計算結果データでも可視化が可能となります。

(1) 可視化対象データ動的切替

可視化セッションの途中で、並列可視化モジュールが組みこまれたリモート可視化プロセスを起動しなおすことなく、GUI により別の計算結果データに可視化対象を切り替えられます。また、その他の特徴として以下の点が挙げられます。可視化対象の計算結果データの指定及びその変更を可視化セッション開始時でなく可視化セッション途中で任意回数行えます。

連成シミュレーション可視化のために、複数の計算結果データのファイルパスを指定できます。

上記複数の計算結果データそれぞれに対し、可視化処理に割り当てる並列プロセス数を指定でき、並列プロセス最大数は可視化セッション開始時に指定する必要はありますが、割り当て並列プロセス数の合計がこれを超えた場合、利用者に修正を促します。

(2) 動画生成シナリオエディタ

動画生成シナリオを GUI によるキーフレーム(物体の形や位置の変化ポイントが指定されたフレーム)指定とタイミング指定により作成できます。

(3) 表示機能

主な表示機能については、以下の一覧(表 2.2-1 **GVSにおける表示機能一覧**)の通りです。

表 2.2-1 GVS における表示機能一覧

機能名	概要
分子モデル表示機能	並列可視化モジュールでは 3 次元物理空間に定義された分子データに基づいて､分子モデルを計算サーバ上でリアルタイムに描画する以下の機能を実装しています。 ・ 陰影付けのされた分子モデルを描画します。 ・ 空間充填モデル、スティックモデル、ボール＆スティックモデルの表示切り替えを随時行えます。
半透明等値面表示機能	並列可視化モジュールでは 3 次元物理空間に定義されたスカラデータに基づ

機能名	概要
	いて、半透明な等値面を計算サーバ上でリアルタイムに描画する以下の機能を実装します。 ・ 3 次元物理空間内に定義されたスカラ物理量に基づいて、陰影付けのされた半透明な等値面を描画します。 ・ 複数枚の等値面を同時に描画でき、等値面の透明度が指定できます。
コンタ平面表示機能	任意に配置した平行四辺形平面上に 3 次元スカラ場の分布をカラーマップで色分け表示します。以下、この平面をコンタ平面と呼びます。本機能においては以下の操作が行えます。 ・ コンタ平面を複数設定し、同時に重ねて表示します。 ・ 各コンタ平面の平行四辺形平面の位置・向き・形状を任意に設定できます。
コンタ等値面表示機能	3 次元スカラ場の等値面上に、これとは別の 3 次元スカラ場の分布をカラーマップで色分け表示します。本機能においては以下の操作が行えます。 ・ コンタ等値面を複数設定し、通常等値面も含め、同時に重ねて表示します。 ・ 各コンタ等値面の色分けに用いるカラーマップを任意に設定が可能です。
チューブモデル表示機能	タンパク質等の生体高分子の構造データ(原子座標および階層構造情報)に基づいて、その主鎖をチューブにより表示します。本機能においては以下の操作が行えます。 ・ 複数の主鎖それぞれに対しチューブモデルの適用を選択でき、複数のチューブモデルを同時に表示します。 ・ 各チューブモデルの半径と色の設定が可能です。
リボンモデル表示機能	タンパク質等の生体高分子の構造データ(原子座標および階層構造情報)に基づいて、その 2 次構造をリボンにより表示します。本機能においては以下の操作が行えます。 ・ 複数の 2 次構造それぞれに対しリボンモデルの適用を選択でき、複数のリボンモデルを同時に表示します。 ・ 各リボンモデルの幅、厚さ、色の設定が可能です。

(4) 可視化対象更新の一時停止機能

クライアントから一時停止のメッセージを受信したとき、並列可視化モジュールは可視化機能を停止する一時停止モードに入ります。時系列の可視化の場合には、時刻ステップが止まった状態となり、クライアントから可視化パラメータの更新があったときに限り、可視化画像の更新を行います。

(5) 等値面生成用構造格子の自動設定機能

等値面表示時のパラメータ設定に伴う利用者負担を軽減するために、物理量格子データの座標値(直方体の位置と大きさ)を分子の形状を踏まえて自動生成する機能を実装します。

(6) 時系列データ表示機能

時系列データに含まれる任意の時刻ステップのデータを可視化し画像を表示します。詳細機能については、以下に説明します。

(7) ImageRecording操作機能

動画用の画像を保存するディレクトリと開始ステップ番号・終了ステップ番号・増分ステップ数を指定し Start ボタンを押すことで、クライアントは画像記録モードに移行し、動画表示用の画像生成を開始します。指定したフォルダが存在しない場合は、その親フォルダが存在する場合に限り新規作成します。このモードの間は、メインウィンドウに生成途中の最新画像を刻々と表示し、ステータスバーには画像生成中である旨の表示を行い、画像枚数の情報も表示します。ただし、セッションの終了・クライアントの終了は画像生成が完了するまで行えません。途中で画像生成を強制中止したい場合は Stop ボタンを押すことで中止できます。これらのメニューは、時系列データの場合に有効で、Step メニューにより指定したステップが、実際に Visualizer セッションでの可視化処理と一致しない場合はエラーを表示します。

(8) Control Step操作機能

セッション開始後、可視化対象が時系列データである場合、途中で可視化対象の時刻ステップを変更することができ、継続して可視化操作を行えます。このメニューは、時系列データでない場合は無効となります。

(9) Dynamic Data画面(可視化対象データ動的切替機能)

可視化セッションの途中で、並列可視化モジュールが組み込まれたリモート可視化プロセスを起動しなおすことなく、GUI により別の計算結果データに可視化対象を切り替えることができます。

(10) Image Viewer画面(時系列画像再生機能)

時系列の可視化データを連続的に再生する際に、1 コマの時刻ステップ増分、再生速度などの設定が可能です。

2.3 グリッドプログラミング環境

グリッド環境で大規模な応用プログラムを開発するためには、地理的に分散した計算資源上のプログラムの起動と、実行時のプログラム同士のデータ転送および同期のためのプログラミング手法の研究・開発が必要になります。

NAREGI ミドルウェアでは、これらのプログラミング手法として GridMPI と GridRPC を提供しています。

2.3.1 GridMPI

GridMPI は、NAREGI ミドルウェアの環境上で MPI アプリケーションを高性能に実行するため、通信遅延を考慮した通信および相互運用性の高い通信を実現する MPI 通信ライブラリです。

GridMPI の特徴には次のものがあります。

- 標準規格の準拠

  MPI-1.2 規格および MPI-2.0 規格に準拠しています。既存の MPI アプリケーションを変更することなく、グリッド環境上で実行することが可能です。また、クラスタ間通信として、IMPI（Interoperable MPI）規格を利用します。

- ヘテロ環境のサポート

  グリッド環境での動作を目的に、計算機アーキテクチャの混在環境（Linux/IA32 および 64、AIX/POWER、Solaris/SPARC、SX-8）へ対応しています。

- クラスタ間通信性能の向上

  高バンド幅広域ネットワークに適した集団通信アルゴリズムにより、クラスタ間の通信の往復を減らしてバンド幅の活用を図っています。また、TCP/IP プロトコルの輻輳時性能を改善するペーシング・モジュール PSPacer(*)を開発、提供しています。

  (*) PSPacer : http://www.gridmpi.org/pspacer-2.1/index.ja.jsp

図 2.3-1にGridMPI実行の概要を示します。互いのクラスタ情報を交換し、クラスタ間の通信（TCP/IP接続）を確立するために、IMPIサーバプロセス（IMPIサーバ）が起動します。続いてクラスタごとにmpirunコマンドが実行され、GridMPIプロセスは、IMPIサーバを介して他のクラスタ上のプロセスとの通信路を確立します。そしてGridMPIプロセスのランク番号が全体で一意になるように再割り当てされ、ジョブの実行が始まります。なおクラスタ内通信として、独自プロトコル（YAMPI）またはベンダー提供MPIを利用します。

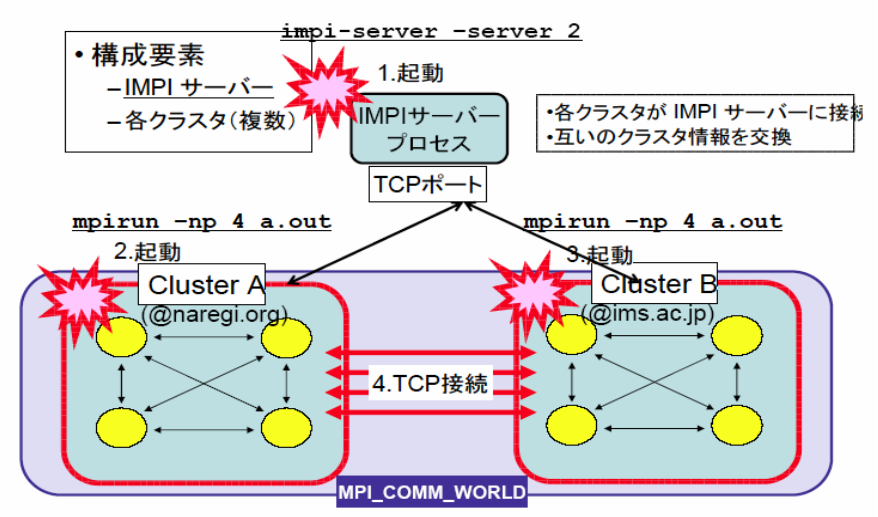


図 2.3-1 GridMPI 実行の概要

2.3.2 GridRPC

GridRPC は、遠隔計算機上でライブラリ関数を呼び出すモデルに基づき、数十から数百 CPU のグリッドアプリケーションを容易に開発し、高い実行効率を可能とする機能です。特定の計算部分を遠隔の複数の計算資源で実行するもので、RPC モデルに適合した、比較的単純な並列性を記述する場合に有効です。

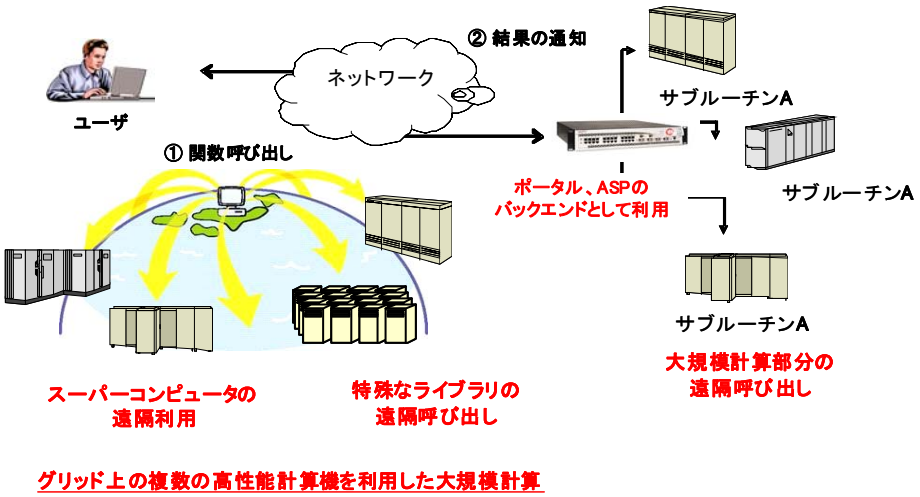


図 2.3-2 GridRPC

GridRPC の特徴として、以下の点が挙げられます。

- グリッド上の特定のコンピュータ資源を利用できます。

- 広域の計算サーバに offloading することで、プログラムの実行時間の短縮ができます。
- グリッド上の複数のサーバでパラメータスィープ（Parameter Sweep）を実行します。パラメータスィープとは、パラメータの一部を使い、並列に複数のサーバで計算を実行するものです。このとき各サーバは別々のパラメータを使って独立に動作します。
- タスク並列プログラムを簡単に作ることができます。
- 複数のクライアント・サーバのやりとりが混ざった、様々なタスク並列の同期をサポートする API が利用可能です。
- 数学の計算が書きやすく、動的に変化するグリッドの規模にあわせて処理を実行します。

2.4 グリッドアプリケーション対応

NAREGI ミドルウェアは、グリッド上のアプリケーションとして、グリッド上の連成ミドルウェア Mediator を提供しています。

連成シミュレーションにおける、異なる物理モデルや計算手法のシミュレーションプログラムにおける同一の物理量を異なった表現を用いての計算、広域グリッド上の疎結合やローカルグリッド上の密結合の連成シミュレーションの対応などの課題に対し、Mediator ではデータ交換機能を提供することで解決をはかっています。

主なデータ交換機能にはプロセス管理機能、異種データのセマンティック変換機能、同期型のデータ転送機能があり、共通ライブラリ（API）として提供しています。これら API を連成シミュレーションプログラムに組込むことにより、プログラムの改変を最小限に抑えた効率的な連結を実現します。

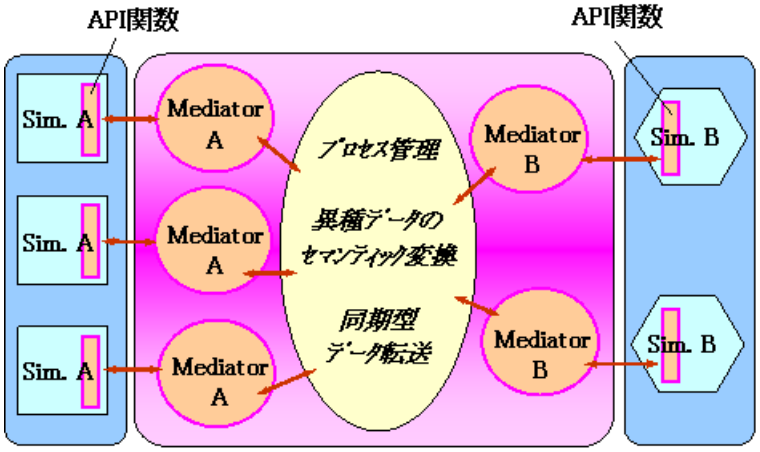


図 2.4-1 Mediator の基本構成と機能

以下、Mediator の持つ主なデータ交換機能について説明します。

2.4.1 プロセス管理

シミュレーションプロセスや Mediator プロセスのグループ分割、シミュレーションプロセスごとのデータ交換を担う Mediator プロセスの対応付けなどのプロセス管理を行います。また、複数シミュレーション間のファイル転送を同期させるためのプロセス管理を行います。

2.4.2 異種データセマンティック変換

分子動力学法や分子軌道法で設定した原子座標、有限差分法や有限要素法で設定したメッシュ点などの異種離散点間で、データ交換を必要とする離散点の位置関係を高速に

探索します。このように、データ交換の行われる異種離散点の位置関係を、異種離散点の相関関係として定義します。例えば、原子座標とメッシュ点の相関関係を最近接と定義すれば、ある原子座標と最近接に位置したメッシュ点の間でデータ交換を行います。

相関関係を有する離散点上のデータは、ビルドイン関数、または、利用者の定義した変換関数を用いてデータ変換（セマンティック変換）を行います。

2.4.3 同期型データ転送（SBC）

データ転送には、MPI ライブラリを用いたメッセージパシング方式と GridFTP などを用いた同期型のファイル転送方式があります。この同期型のファイル転送機能を Synchronous File Transfer library（SBC）機能と呼んでいます。連成シミュレーションの結合度に応じて柔軟に選択可能です。

2.5 データグリッド

グリッドで扱うデータはテラバイトからペタバイトへと研究データが増え、単一のファイルサーバにまとめて利用することが限界に近づきつつあります。そのため、グリッドの長所を活かすためには、分散されたデータ資源をそのまま大きな仮想的なファイルシステムとして使うことができる拡張可能な共有ファイルシステム機能を実現しています。NAREGI ミドルウェアが提供するデータグリッドコンポーネントは、計算・シミュレーションの高度化、高速化などを実現するだけではなく、広域に分散するコンピュータ資源を結び付け、大規模な計算処理や大量なデータを保存、利用を可能にします。

データグリッドはデータグリッド資源管理システムおよびデータグリッドアクセス管理システムから構成されます。データグリッド資源管理システムは、共有ファイル空間を実現し、その中にどういうファイルが存在しているか、データ資源を管理します。データグリッドアクセス管理システムは、共有ファイルの登録・更新・削除などのファイル管理機能を提供します。
以下、データグリッドの各システムについて説明します。

2.5.1 データグリッド資源管理システム

データグリッド資源管理システム（DGRMS）は、国際的な GGF の表示に従い、WSRF（Web Service Resource Framework）をベースに拡張可能な共有ファイルの割当て配置と使用容量を統一的に管理します。

主な機能として以下の機能があります。

- ファイルシステムノードごとのファイル数、使用容量、利用者毎のファイル数、使用容量、について、全体、特定ユーザ、特定ノード、特定ディレクトリ以下の情報を表形式で表示することができます（図 2.5-2）。
- 検索条件として、ファイル名、ファイルサイズ(min or max)、ファイル所有者、更新日の指定が行えます。

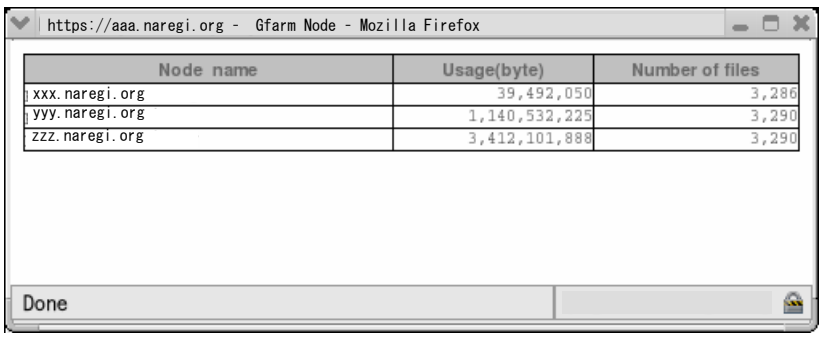


図 2.5-1 ノードごとのファイル数表示

2.5.2 データグリッドアクセス管理システム

データグリッドアクセス管理システム（DGAMS）は、共有ファイルの登録・更新・削除などのファイル管理を行います。

主な機能として以下の機能があります。

- ローカルホストのファイルを Gfarm へインポートすることができます。
- ファイル転送ジョブの一覧を表示します。
- 転送ジョブの削除が可能です。
- データアクセスの際に使用するプロキシ証明書として、VO 属性付のものを使用できます。gLite では、VO 情報はプロキシ証明書の拡張情報部分に保存されているので、そのようなプロキシ証明書でアクセスが可能です。
- WFT へファイル URL 所有者などのファイル属性情報を渡すことが可能です。
- Gfarm、GridFTP、SRM に配置されているファイルに対する操作（インポート、エクスポート、ファイル・ディレクトリの削除、ディレクトリの作成等）が行えます。
- ファイル転送時の一次中断、再開が可能です。
- ファイルシステムノード毎のファイル数、使用容量、利用者毎のファイル数、使用容量を表示することができます。
- ファイルの分割情報と保存先を表示することができます。
- Gfarm 上のファイルにコメント情報を付加することができます。
- 検索条件として、ファイル名、ファイルサイズ(min or max)、ファイル所有者、更新日、コメントの指定が行えます。

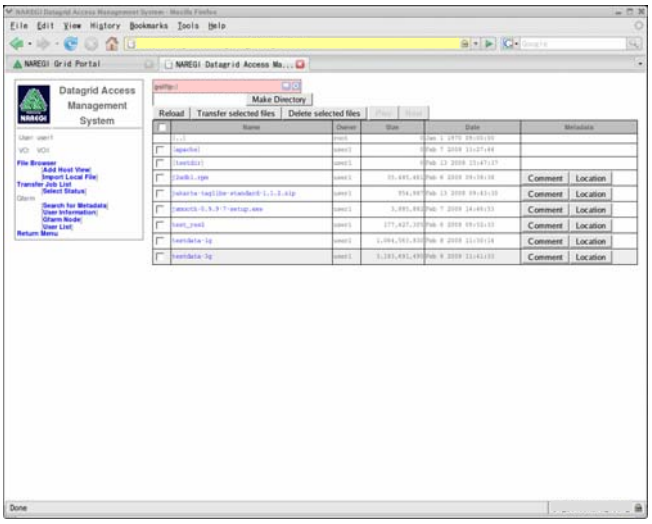


図 2.5-2 データグリッドアクセス管理システム トップ画面

2.6 資源管理

NAREGI ミドルウェアの資源管理機能は、グリッドコンピューティング環境上に広域に分散された､異機種が混合した計算資源環境を 1 つにまとめて運用することを目的に、SS 、GridVM、IS の 3 つのコンポーネントから構成されます。

SS は、資源ブローカリング機能、ジョブワークフローエンジンなどを持ち資源やジョブの管理を行います。GridVM はローカルスケジューラの違いを吸収し、利用者がスケジューラの違いを意識せずにジョブ投入することを可能にします。IS は、グリッド環境における計算資源、ネットワーク、ソフトウェア、アカウンティングなどに関する情報の統合的管理を行います。これら三者の連携で、一般的なジョブ投入機能のみならず、アーキテクチャの異なる計算資源のコアロケーション機能、パラメータサーベイジョブに対応したバルクジョブ投入機能などの機能を実現します。

以下、資源管理の持つ機能について説明します。

2.6.1 資源情報の収集と蓄積

グリッド環境上に分散した複数の計算センタにまたがる資源情報、様々な VO に対して動的な計算環境を提供し、いつ、どこで、誰が、何のジョブを何回実行したかという資源利用記録など、VO のポリシーに基づく管理に必要な情報に対して、検索や収集、登録、通知を行います。

情報は階層的に収集され、CIM（Common Information Model）をベースにしたNAREGI Schemaに従い蓄積・管理されます（図 2.6-1)。

収集された資源情報は、運用管理者による構成管理やユーザ管理、資源利用記録（アカウンティング情報）管理、障害管理業務の支援を目的に、GUI上に表示する機能も有しています（2.6.6 情報表示機能)。

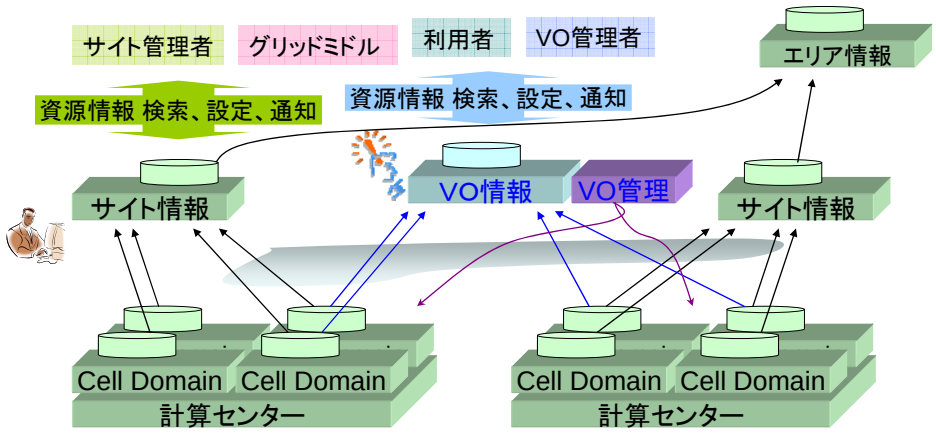


図 2.6-1 階層的な情報収集のイメージ

2.6.2 計算資源の探索と確保

ある 1 つのジョブに対して、VO の資源制約やポリシーとの整合性を確認し、利用可能な計算資源を探索して発見、選択し、確保して投入する機能を持っています。この際、複数の計算資源を同時に確保することもできます。計算資源の探索には、上記の資源情報収集・蓄積機能が用いられます。また、他のコンポーネントからのジョブ要求（サブミット、状態取得、削除）の仲介を行い、ジョブ管理や予約サービスを行います。

2.6.3 ジョブの予約と管理

予約ジョブについては、開始時刻、期間、ノード数で資源の仮予約をおこない、実際に予約が成功した際は予約を確定します。非予約ジョブの場合は資源予約を行わず、ローカルスケジューラへジョブを投入します。

予約ジョブは、いったん予約が確定すれば、指定時刻にジョブが開始され、指定された期間必要な資源を使用でき、後続の予約ジョブや非予約ジョブによる影響は受けません。非予約ジョブは、すでに確定されたジョブの空き資源、空き時間を使って実行されるようにスケジューリングされます。

非予約ジョブがスケジューリング（使用資源と開始時間が確定）されていない段階で予約ジョブが投入された場合、予約ジョブが優先して実行されます。なお非予約ジョブの場合でも、状態取得（実行中かどうか）やアカウンティング情報の収集と蓄積、表示が可能です。非予約ジョブ・予約ジョブに関わらず、指定期間を超えて実行しようとした場合は自動的にジョブを強制終了します。NAREGIミドルウェアがサポートするジョブの種類については「1.2.4 ジョブの種類」を参照ください。

2.6.4 アクセス制御

NAREGI ミドルウェアの資源管理機能は、違法な資源アクセスからローカルな資源を保護するために、アクセス制御機能を提供しています。

(1) ファイルアクセス保護

UNIX ベースのシステムでは、資源はファイルとして扱われるため、ファイルへのアクセス制御を行うことにより資源の保護を行います。

ファイルへのアクセスは、ジョブが実行するシステムコールを介して行われるので、ポリシーに基づいてシステムコールを監視・制御します。システム管理者は、自身が管理する計算資源に対して、誰に(subject)、どの資源を(target)、どうのようなアクセス(action)を許すか許さないか(permit/deny)を XML 形式のアクセス保護ポリシーで指定し

ます。ここで、subject はローカルアカウントでなく、グローバルユーザ ID や VO 名です。ポリシーに基づいて、アクセスしようとしている subject(あるいはその属性)、アクセスしようとしているファイル(およびディレクトリ)、アクセス操作 action およびその他の条件を判定し、トラップしたシステムコールの実行を許可するか、拒否するかを判定します。

(2) 資源利用量制御

ポリシーとして指定された資源利用条件に基づき、自律的にジョブを制御し、資源利用量の超過防止など資源の保護を行います。

サイトポリシーは、サイト管理者（計算資源所有者）がユーザや VO に対して提供可能なジョブ毎の資源利用量を定義します。ポリシーファイルは、XML 形式であり、主体(subject)、監視対象(metric)、制御(control)のセットを記述します。資源利用量の監視対象として、実行時間、CPU 時間、ディスク使用量が指定可能です。また VO ごとの資源利用量を予め設定することで、制限超過時のジョブ制御（ジョブ削除）が可能です。

2.6.5 ジョブ監視

グリッド環境を構成する計算資源や、計算資源上で実行されるジョブの管理のため、ジョブの監視情報をシステム管理者や利用者が利用できるように提供しています。

(1) 資源情報プロバイダ

各サイトの保有する資源情報を NAREGI ミドルウェアの資源管理機能に登録するための機能です。

各サイトのシステム管理者は、グリッド環境に提供する各サイトの資源をポリシーなどにより自律的に設定することが可能です。つまり、全ての物理資源をグリッド環境に使わせるのではなく利用可能な資源量を制限し、また、利用可能な利用者やジョブを限定することが可能となります。

(2) ジョブトラッキング

各ジョブがどの計算資源上で実行されるかを把握することは、システム管理者や利用者にとって障害場所の特定などのために有用です。また、アプリケーション開発においても、ブローカリング後サイト間で通信相手となるジョブを動的に特定してファイル転送先を決めるなどを行う上で、有用なユーティリティ機能となります。

ジョブ実行時に決定された計算資源群は NAREGI ミドルウェアの資源管理機能に登録されます。この登録された情報は、運用管理ツールやアプリケーションからジョブの

ID をキーとして参照できます。

(3) アカウンティング情報レポート

ジョブの終了時、そのジョブが消費した資源を情報サービスに登録する機能です。アカウンティング情報の登録は、OGF（Open Grid Forum）で標準化が行なわれている Usage Record(UR)および Resource Usage Services(RUS)に従って登録されます。UR は、登録情報を定める XML スキーマであり、登録情報は XML データとして登録されます。

2.6.6 情報表示機能

資源管理機能は NAREGI ミドルウェアの各種稼働状態を GUI で表示することが可能です。ここでは主な情報表示機能について説明します。

(1) エントランス画面

NAREGI Portal の Grid Tools 起動画面において、Information Service を選択すると起動します。ユーザのプロキシ証明書から取得した識別名、VO 名、プロキシ証明書の残り有効時間(秒)等が表示されます。Next ボタンを押すとメイン画面に遷移します。

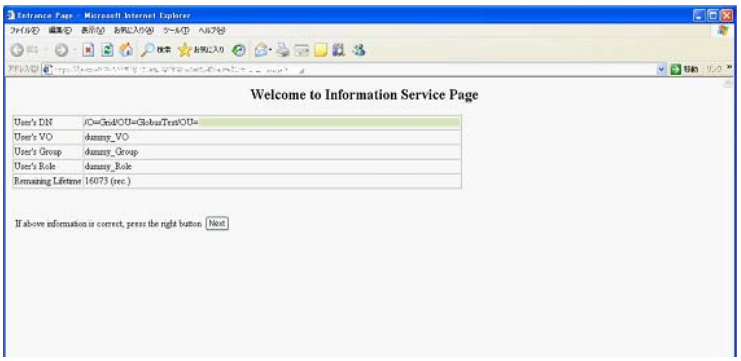


図 2.6-2 エントランス画面

(2) メイン画面

画面上部にエントランス画面より引き継がれた内容、検索のために接続する分散資源情報サービスの URI、選択したセルドメイン名が表示されます。また、プルダウンメニューでは選択可能な機能、画面下部には接続する分散資源情報サービスから検索可能な範囲のセルドメインのツリーを表示します。

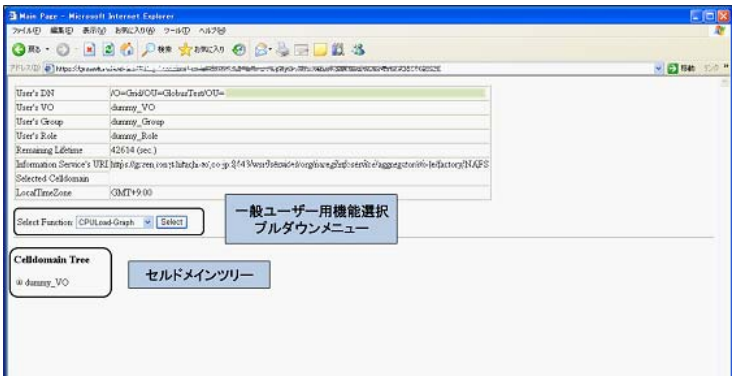


図 2.6-3 メイン画面

メイン画面において選択可能な機能は次のものがあります。

- CPULoad Graph 機能
- UsageRecordUser 機能
- JobMonitorUser 機能
- ジョブ実行結果集計画面
- VomsQueueUser 機能
- Resource 機能
- LogBrowser 機能
- UsageReordAdmin 機能
- JobMonitorAdmin 機能
- インターオペレーション共通属性表示画面
- サービス動作状態一覧画面

以降、各機能/画面の概要について説明します。各機能や画面の詳細は「利用者ガイド NAREGI Middleware IS(Distributed Information Service)」を参照ください。

(3) CPULoad Graph機能

GUI 上からホスト名を指定することにより、対象ホストの CPU 状況を確認することができます。グラフの右端が現在を表し、現在までの時系列の値を表示しています。縦軸は CPULoad をパーセント表示します。画面は自動的に 1 分間隔で更新されます。

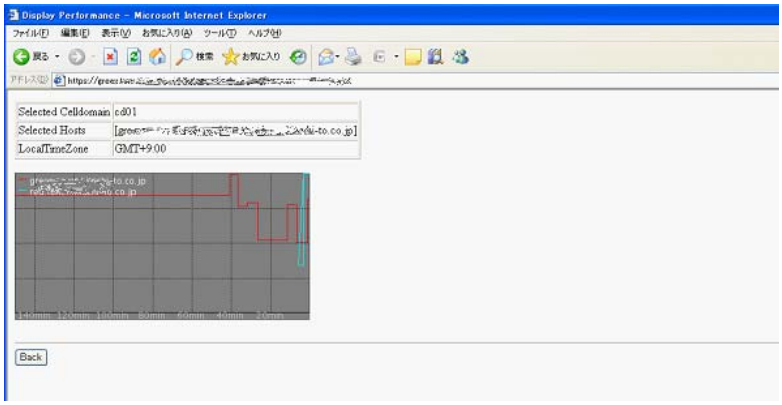


図 2.6-4 CPULoadGraph 機能画面

(4) UsageRecordUser機能

ユーザが利用した資源利用記録の概要をグローバルジョブ ID、ジョブを実行したキュー名、ジョブの開始時間のみの概要表として表示します。

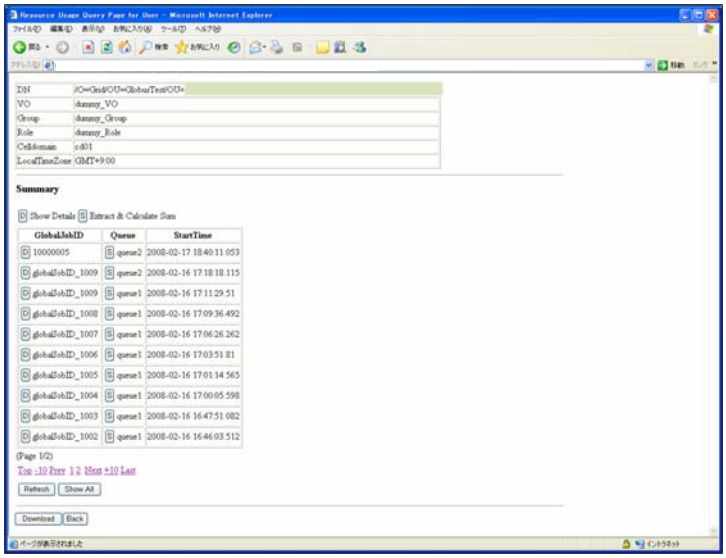


図 2.6-5 UsageRecordUser 機能画面

(5) JobMonitorUser機能

接続したノード配下のセルドメインに記録されている、ワークフロー情報、ジョブ実行状態情報及び資源予約情報の概要をグローバルジョブ ID、ユーザ識別名、実行開始時刻、状態の一覧表として表示します。

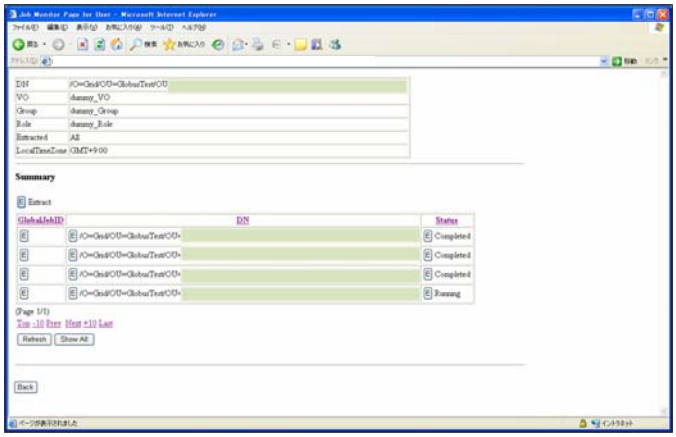


図 2.6-6 JobMonitorUser 機能画面

(6) ジョブ実行結果集計画面

ある期間における成功したジョブ数及び失敗したジョブ数を、VO 毎、資源プロバイダ毎、グリッド全体で集計して表示します。

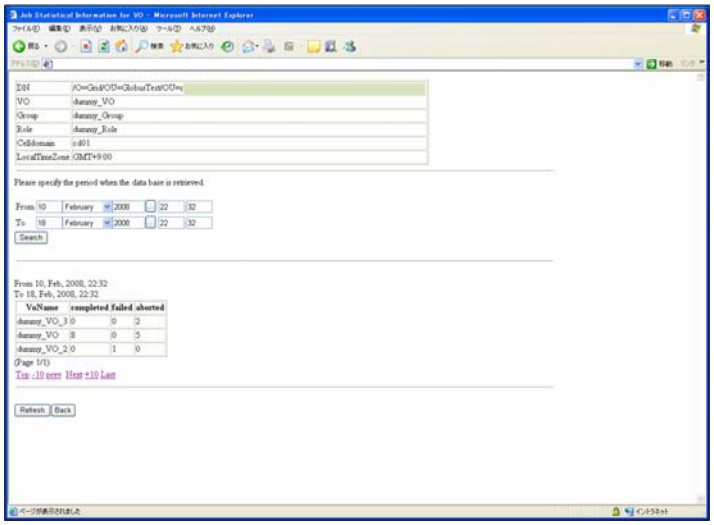


図 2.6-7 VomsQueueUser 機能画面（VO ごとに集計時）

(7) VomsQueueUser機能

ユーザが利用可能なキューに関する情報を表示します。

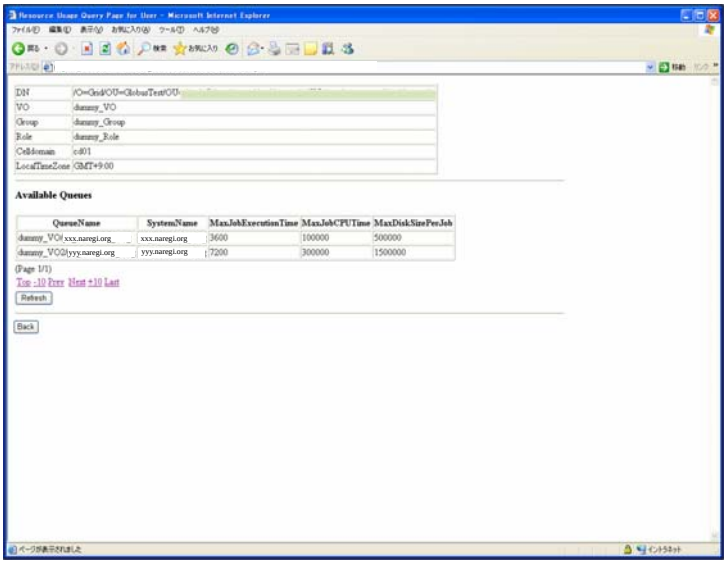


図 2.6-8 VomsQueueUser 機能画面

(8) Resource機能

検索対象 CIM クラスを選択し、対象セルドメインに関する当該クラスの持つインスタンス情報を表示します。

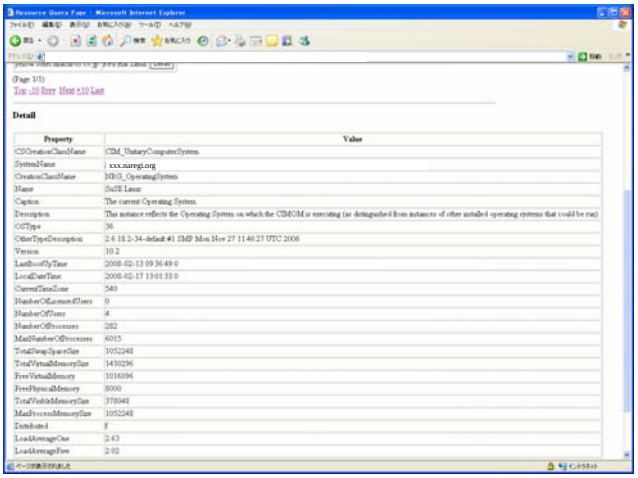


図 2.6-9 Resource 機能画面（詳細表示時）

(9) LogBrowser機能

ログ情報の検索をルートノード「Logs」より開始し、ツリー状に階層を下がるにつれ

て対象データの範囲を狭めていくことで最終的に必要としているログ情報を取得します。

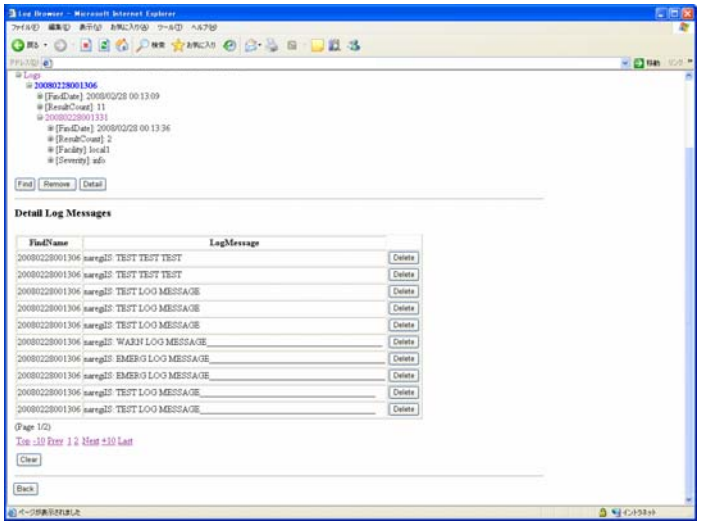


図 2.6-10 LogBrowser 機能画面(詳細表示時)

(10) UsageReordAdmin機能

管理対象のセルドメインの資源利用記録を表示します。

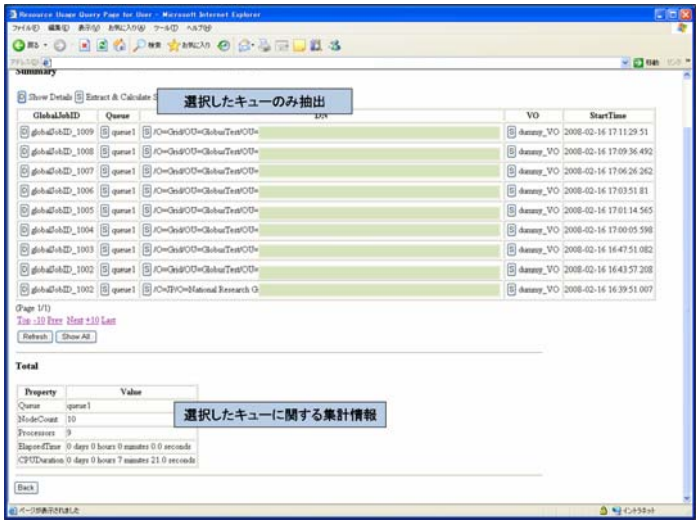


図 2.6-11 UsageRecordAdmin 機能画面(キュー利用記録表示時)

(11) JobMonitorAdmin機能

接続したノード配下のセルドメインに記録されている、ワークフロー情報、ジョブ実

行状態情報及び資源予約情報の概要がグローバルジョブ ID、ユーザ識別名、状態の一覧表として表示します。

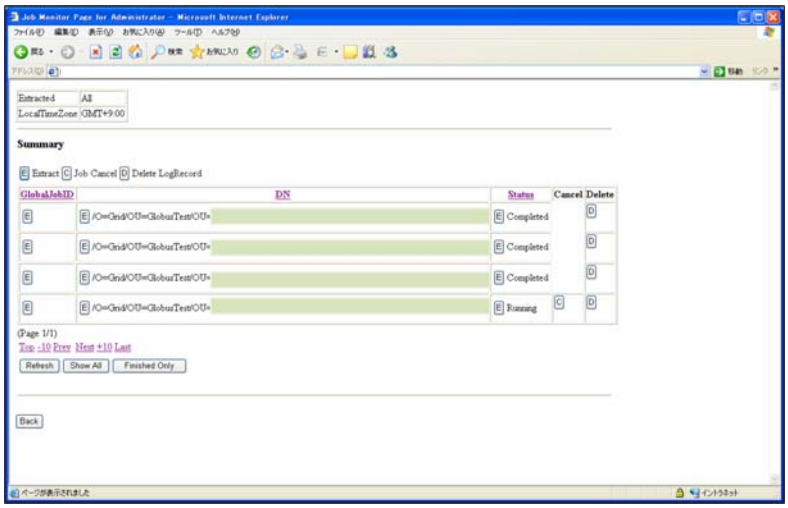


図 2.6-12 JobMonitorAdmin 機能画面

(12) インターオペレーション共通属性表示機能

接検索対象クラスを選択し、対象セルドメインに関する当該クラスの持つインスタンス情報を表示します。

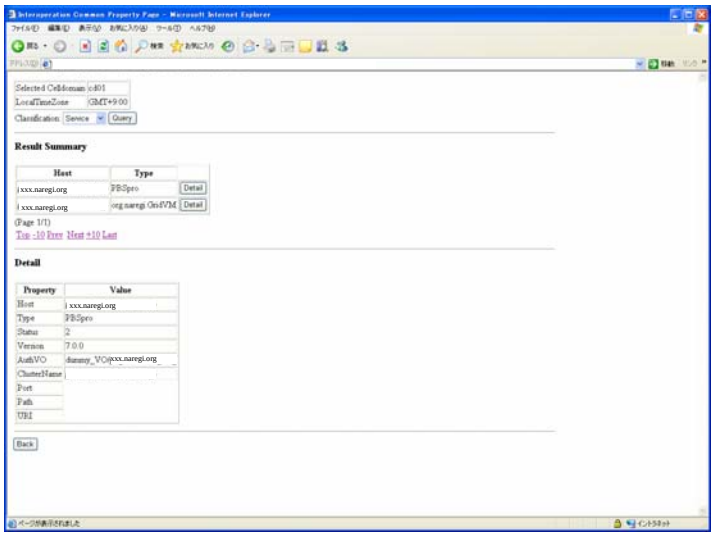


図 2.6-13 インターオペレーション共通属性表示画面(詳細表示時)

(13) サービス動作状態一覧画面

選択したセルドメインに記録されているサービス動作状態が、サービス名(ServiceName)、サービス URI(URI)、サービス動作状態(Status)の一覧表として表示します。一覧表中のサービス動作状態情報は動作状態により色分けされて表示します。

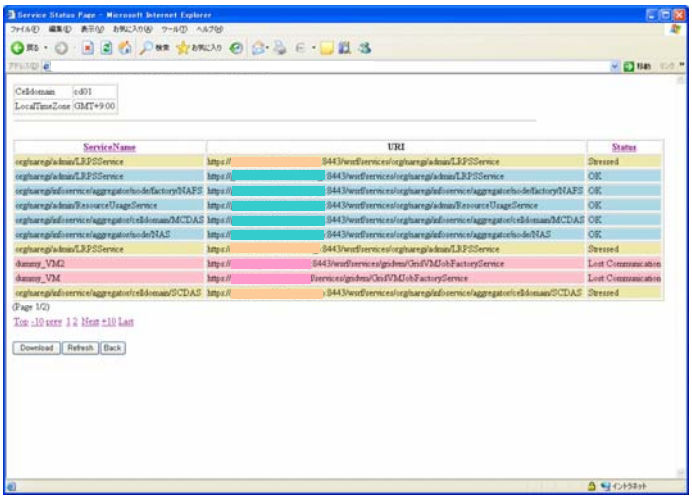


図 2.6-14 サービス動作状態一覧画面