異種の複数スケジューラで管理される資源を事前同時予約するグリッド高性能計算の実行環境

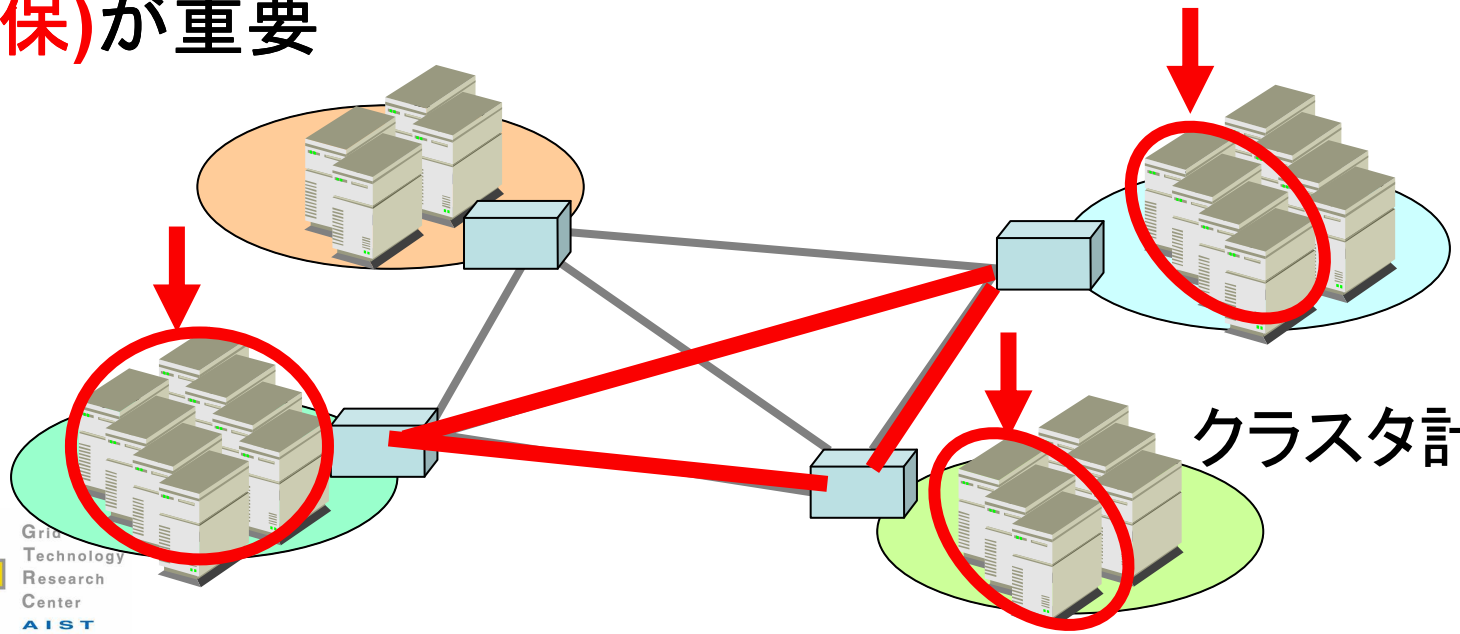
竹房あつ子, 中田秀基, 武宮博, 松田元彦,
工藤知宏, 田中良夫, 関口智嗣

産業技術総合研究所　グリッド研究センター

Grid Technology Research Center AIST

グリッドにおけるメタコンピューティング

- 異なる組織から提供される分散資源を同時に利用
  - 大規模科学技術計算が可能
- 並列アプリケーションでは資源の性能・負荷が実効性能に影響
- 性能を保証し，多様な資源のコアロケーション(同時確保)が重要



グリッド資源のコアロケーションの課題1

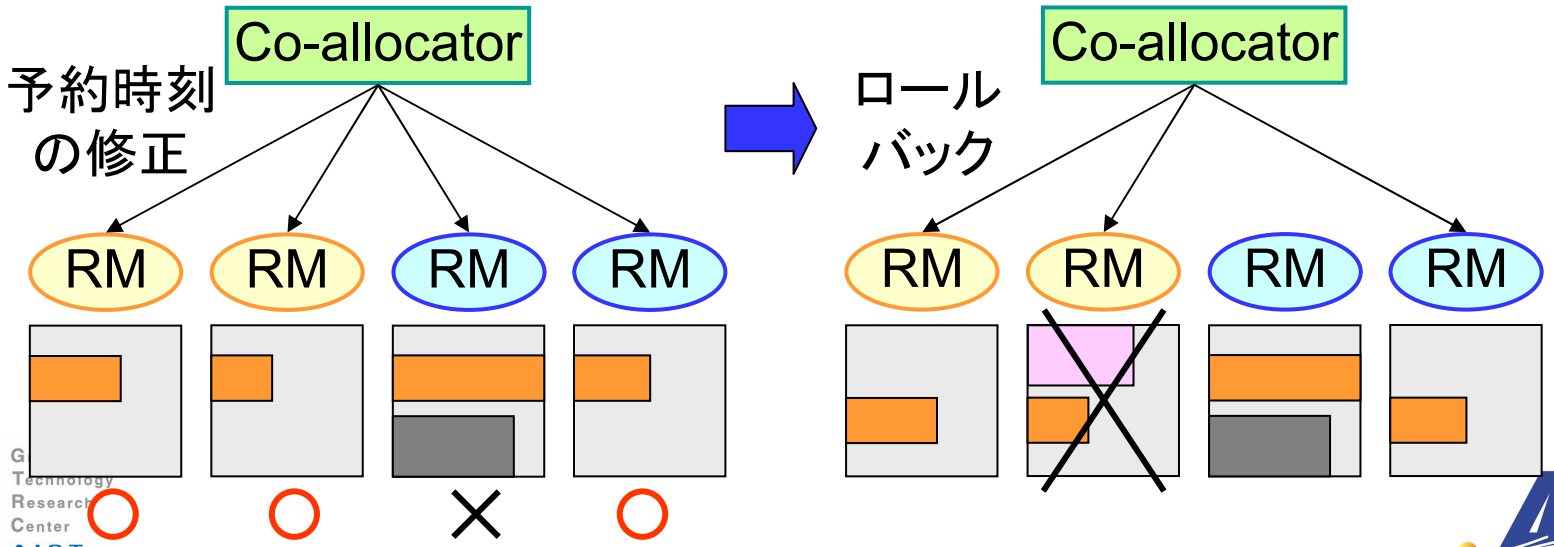
- 既存キューイングスケジューラとの連携
  - クラスタ計算機等の資源は有効利用のため多様なスケジューラで管理
  - サイト(組織)ごとにスケジューリングポリシも異なる
- 事前予約
  - キューイングスケジューラではジョブ投入から実行開始までの時間が一定でない
  - 他の資源と同時確保するには事前予約機能が必要
- **WSRF(Web Services Resource Framework), GSI**
  - セキュアで標準的なインタフェース

→ グリッドコアロケーションシステムを開発 (**GridARS**: **Grid Advance Reservation-based System framework) [SACSIS2006]**

グリッド資源のコアロケーションの課題2（1/2）

- 多様な資源への対応
  - 異なる資源のスケジューラとも容易に連携
- 2相コミット
  - 分散資源の同時確保のため，トランザクションのサポートが必須

1相コミットの場合



グリッド資源のコアロケーションの課題2（2/2）

- 異種の複数スケジューラで管理される分散資源上での＊並列計算＊が容易に実行可能
  - MPI等で実装されるような並列アプリケーションは資源の性能変動に弱い
  - 紳士協定＋SSHでは資源の有効利用や性能保障はできない！

本研究の成果

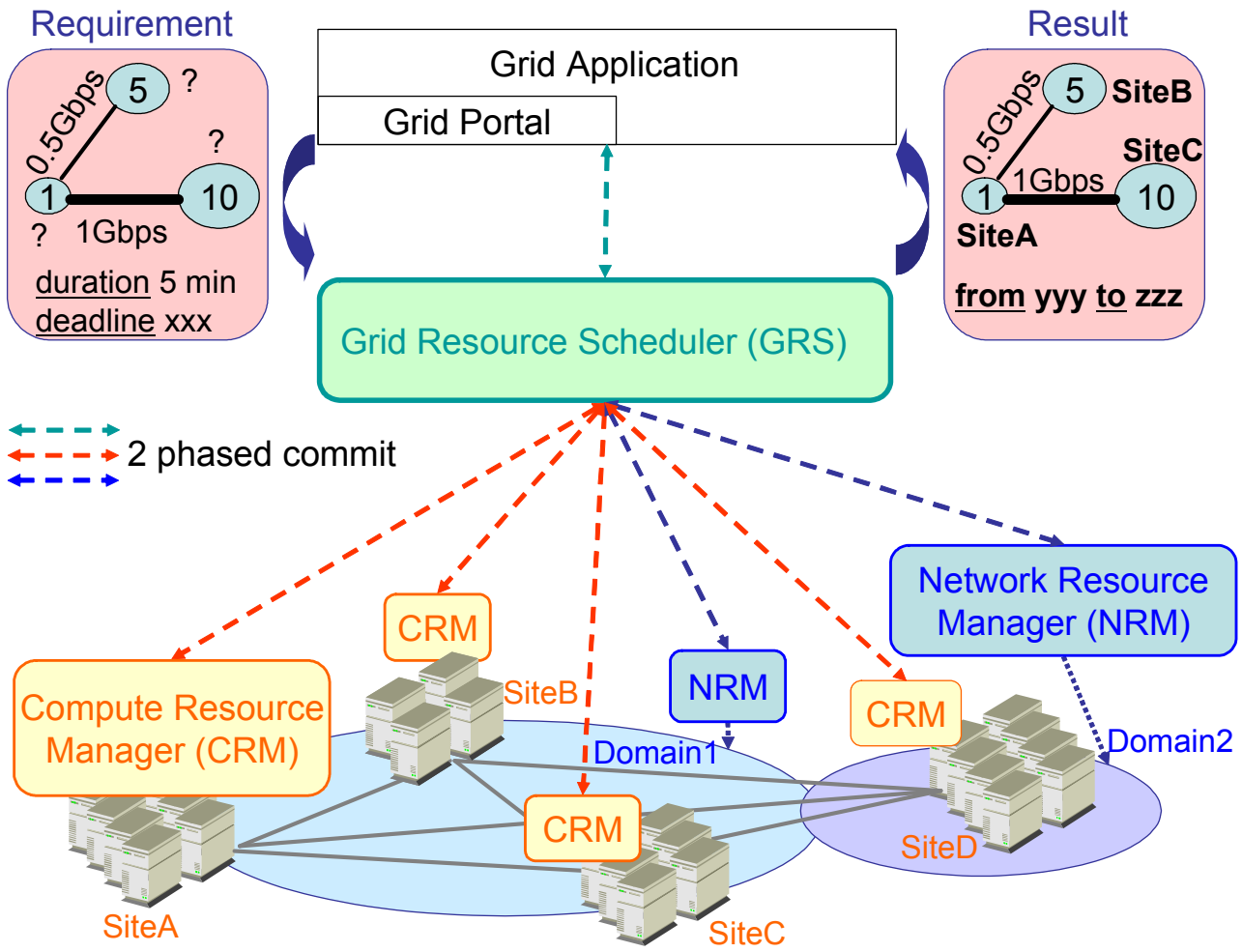
- **GridARSを改良し，階層的な2相コミットでの事前予約手続きを実現（→GridARS-WSRF I/Fモジュール）**
  - GridARS-WSRFはグローバルスケジューラ(Co-allocator)，多様なローカル資源マネージャで利用可能
- **改良したGridARSを用いてポータルを構築**
  - GridMPIで実装された科学シミュレーション
  - 日米間に跨る複数スケジューラで管理される資源上で容易に並列計算が実行できることを実証

発表内容

- **GridARSコアロケーションシステム**
  - GridARS-WSRFによる2相コミット
  - クライアントインタフェース
- **GridARSを用いたグリッド高性能計算ポータル構築事例**
  - QM/MD連成シミュレーション (GridMPIで実装)
  - ポータルの構成と実装

GridARSコアロケーションシステムの概要

- **GRS**(グローバルスケジューラ), **RM**(資源マネージャ)で構成
- **GRSが複数のRMと連携し，要求された資源を割り当てる**
- **ユーザ-GRS間，GRS-RM間は階層的な2相コミット→GRS自身もRMの1つとなり得る**

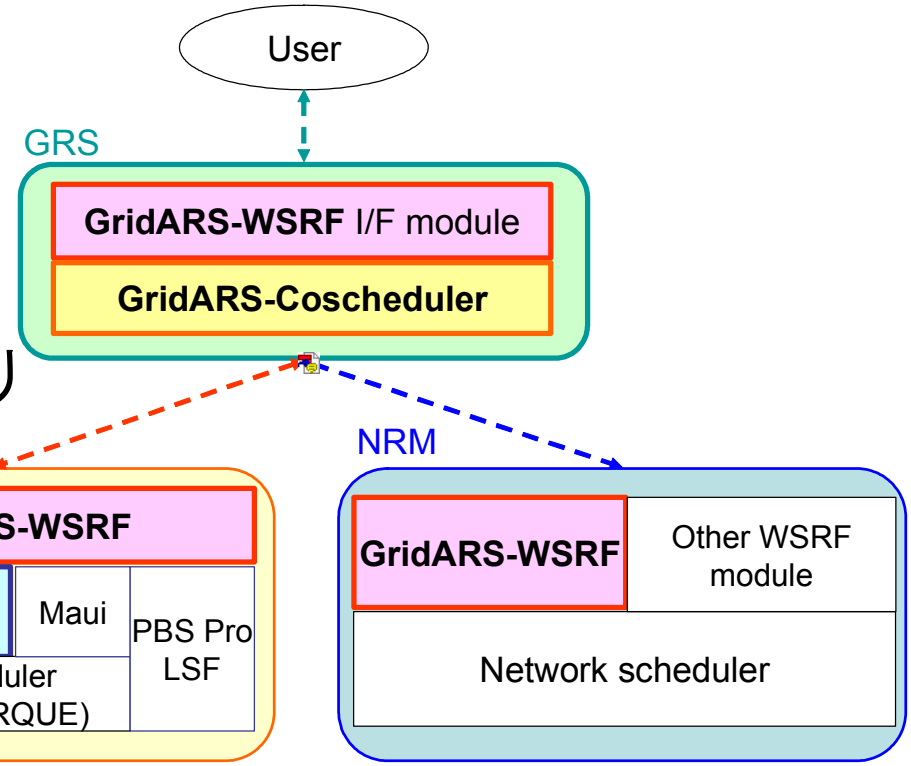


Grid Technology Research Center AIST

AIST

GridARSのシステムアーキテクチャ

- **GRS**
  - WSRFのインタフェースモジュールGridARS-WSRF
  - 事前同時予約を行うGridARS-Coscheduler
- **CRM(計算資源)，NRM(ネットワーク)**
  - GridARS-WSRF
  - ローカルな事前予約機能付き資源スケジューラ
    - PluS/MAUI + SGE/TORQUE
  - GridARSが採用するWSDLで実装されていれば，互換性あり
    - NRMではGNS-WSI (v.2)を採用
    - G-lambda プロジェクトでネットワークに関するサービスI/Fを規定



Grid Technology Research Center AIST

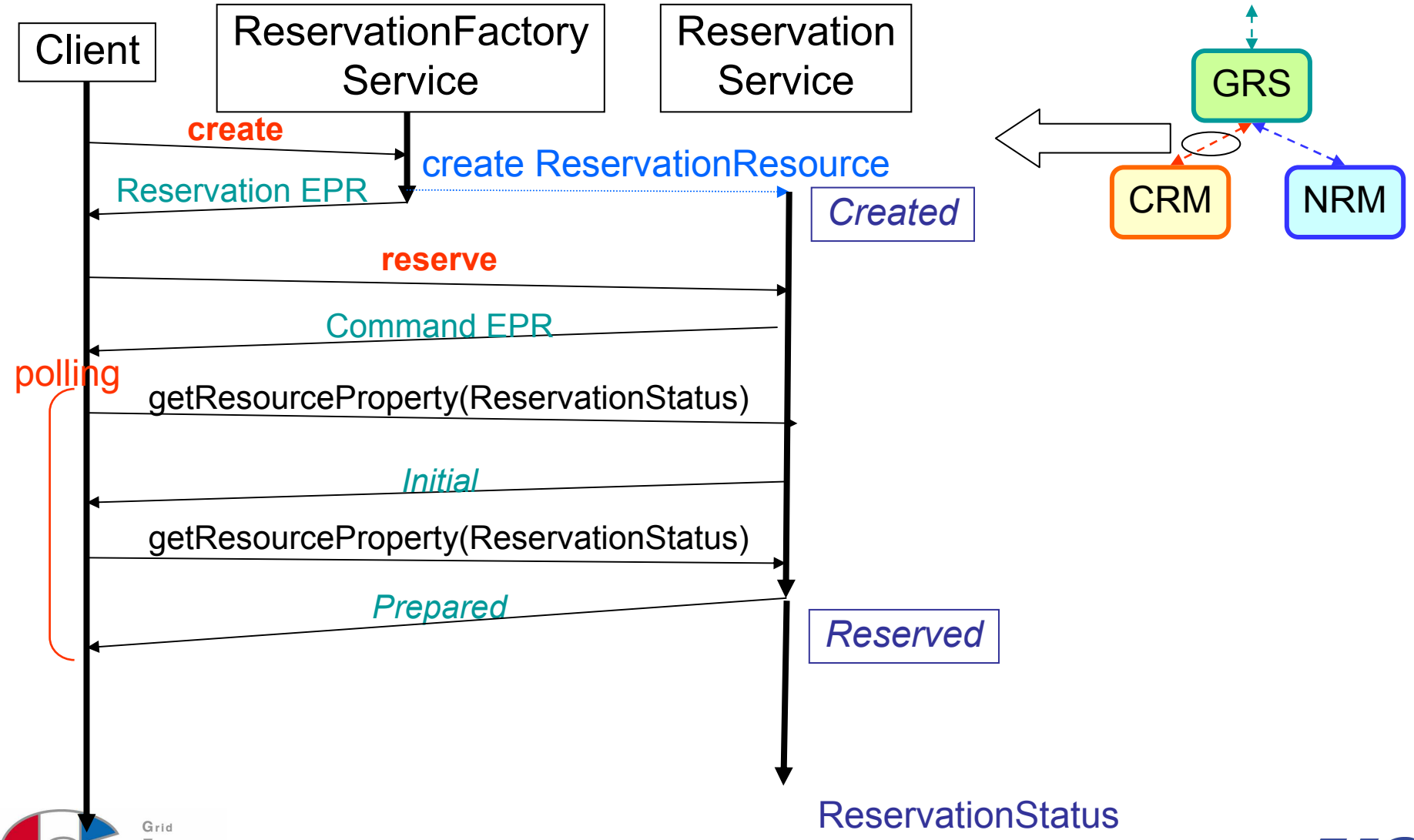
G-lambda http://www.g-lambda.net/

AIST

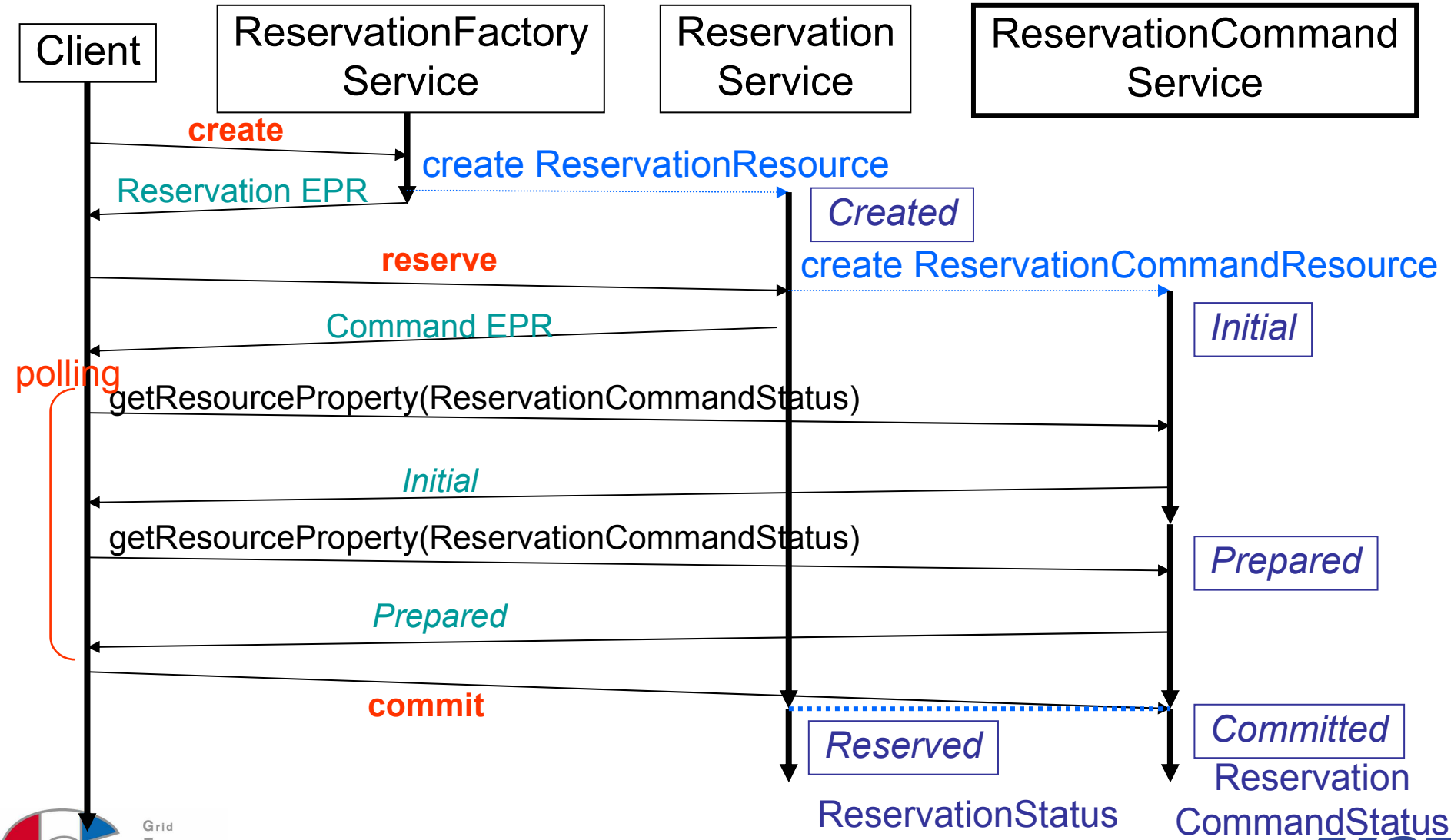
GridARS-WSRF

- **WSRFで2相コミット事前予約をするためのモジュールを提供**
- **ポーリングベース**
- **Globus Toolkit 4 (GT4) [ANL] で実装**
  - GSIとgridmapfileによる認証・認可
- **構成モジュール**
  - WSDLラッパ
    - 多様な資源のWSDL(資源パラメータ)の差異を吸収
    - WSDLは資源ごとに定義
  - メインモジュール
    - 共通する事前予約・修正・解放手続きを実現
    - 各オペレーションはノンブロッキング
  - 資源マネージャラッパ
    - コスケジューラや実資源スケジューラに対し，Java APIを提供

1相コミット版事前予約シーケンス[SACSIS06]



GridARS-WSRFによる事前予約2相コミット



GridARSクライアントインタフェース

- **JavaクライアントAPI, コマンドラインI/F，シェルインタフェースを提供**
- **WS-Addressing仕様で定義されたEPR(Endpoint Reference)[OASIS]を用いて，各予約インスタンスを識別**
  - サービスへのアクセスポイント(URI)
  - 予約識別番号
- **予約資源の表現では，計算資源にはJSDL[GGF]，ネットワーク資源にはGNS-WSI[G-lambda]をそれぞれ拡張したものを利用**

Java APIを用いた事前予約手続き

事前予約手続き

```
// 予約インスタンスを生成
EPR rsvEPR = GrsClient.create(GRS_FACTORY_URI);
// 資源予約要求を送信
EPR cmdEPR = GrsClient.reserve(rsvEPR, REQUIREMENTS);
// 資源の仮予約手続きが完了するまでポーリング
// 予約手続きを完了させる
GrsClient.commit(cmdEPR);  // abortの場合，予約を破棄する
```

- EPR : EndpointReferenceType
- GRS_FACTORY_URI : GridARS GRSのFactoryサービスのURL
- REQUIREMENTS : 要求する資源の情報

Java APIを用いた予約資源の修正・解放手続き

予約資源の修正手続き

```
// 予約資源修正要求を送信
EPR cmdEPR = GrsClient.modify(rsvEPR, REQUIREMENTS);
// 仮修正手続きが完了するまでポーリング
...
// 修正手続きを完了させる
GrsClient.commit(cmdEPR);
```

予約資源の解放手続き

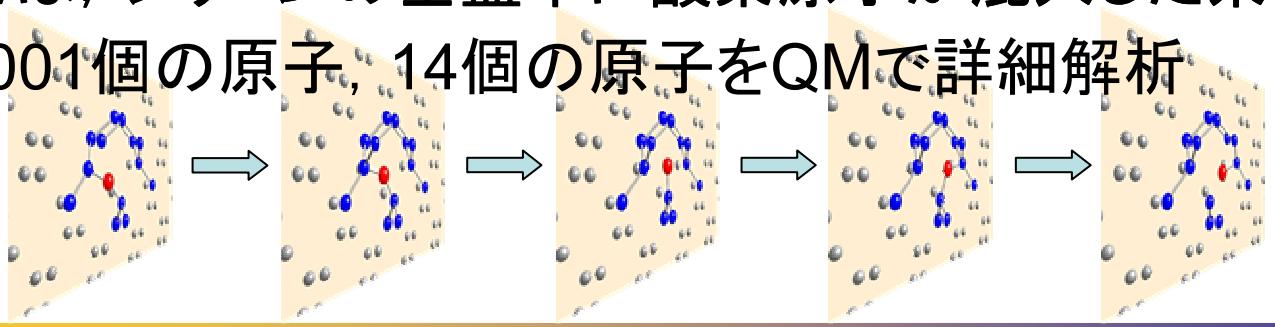
```
// 予約資源解放要求を送信
EPR cmdEPR = GrsClient.release(rsvEPR);
// 仮解放手続きが完了するまでポーリング
...
// 解放手続きを完了させる
GrsClient.commit(cmdEPR);
```

グリッド高性能計算のポータル構築事例

- 改良したGridARSを用いてポータルを構築
  - GridMPIで実装された科学シミュレーション
  - 日米間に跨る複数スケジューラで管理される資源上で容易に並列計算が実行できることを実証
- ポータルの構成
  - QM/MD連成シミュレーション
  - GridARSで資源事前予約，GT4 WS GRAMでジョブ起動
  - 計算資源スケジューラにはPluSと既存事前予約キューイングスケジューラを利用
  - ネットワーク資源スケジューラにはキャリアの提供するスケジューラを利用

QM/MD連成シミュレーション

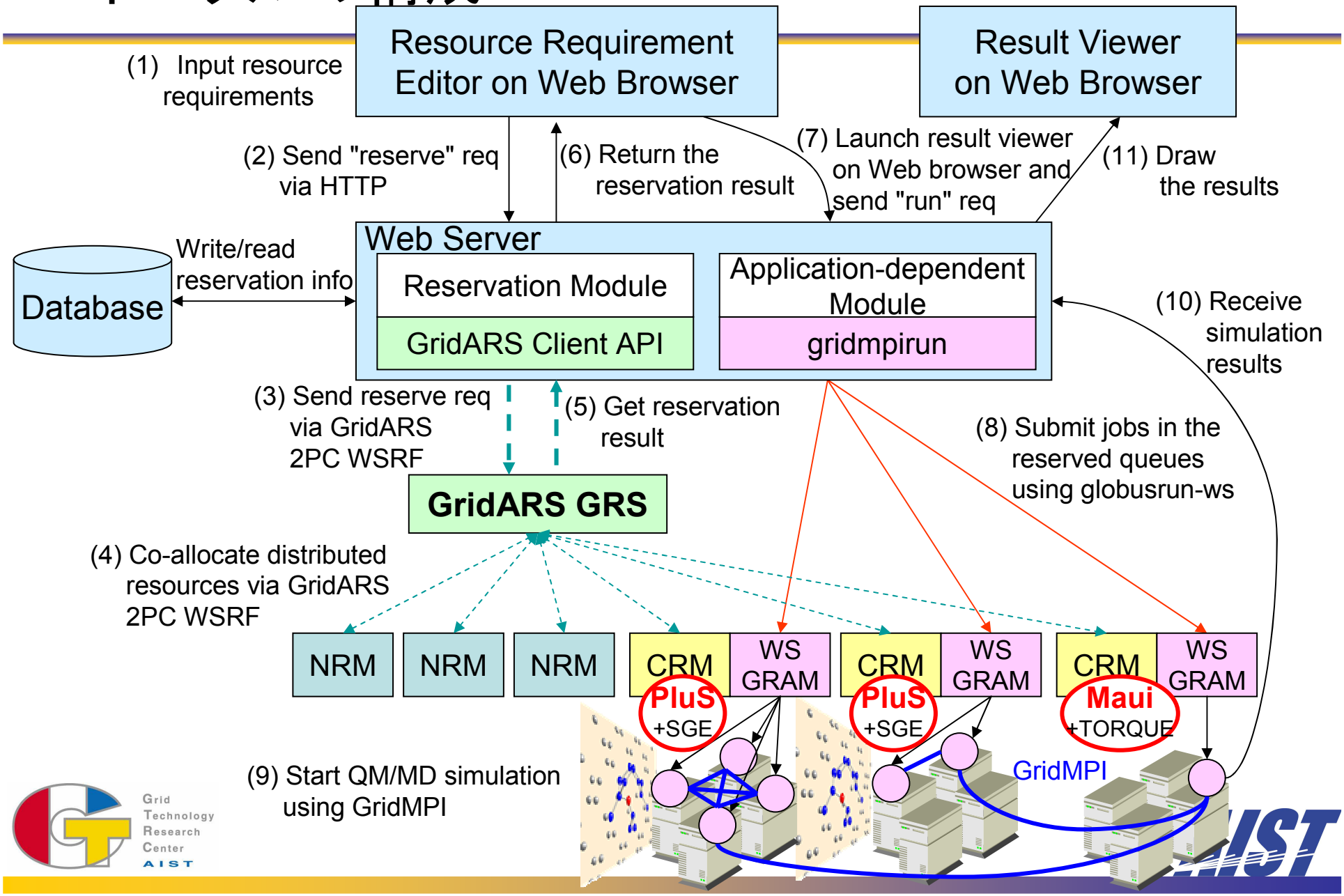
- 量子力学/分子動力学(QM/MD)連成シミュレーションコード[名工大 尾形ら]を機能拡張し，NEB (Nudged Elastic Band) 法に基づく化学反応をシミュレート
  - 反応の開始・終了時の系の状態(点)から化学反応経路を推測
  - 反応過程の各点は並列(→MPI)に求められる
- コードはMD部分，QM部分，コントローラからなる
  - コントローラは反応中の各点での系の原子分布を推測
  - QM/MD部分は連携して与えられた系のエネルギー計算
- 実験では，シリコンの基盤中に酸素原子が混入した系
  - 64001個の原子，14個の原子をQMで詳細解析



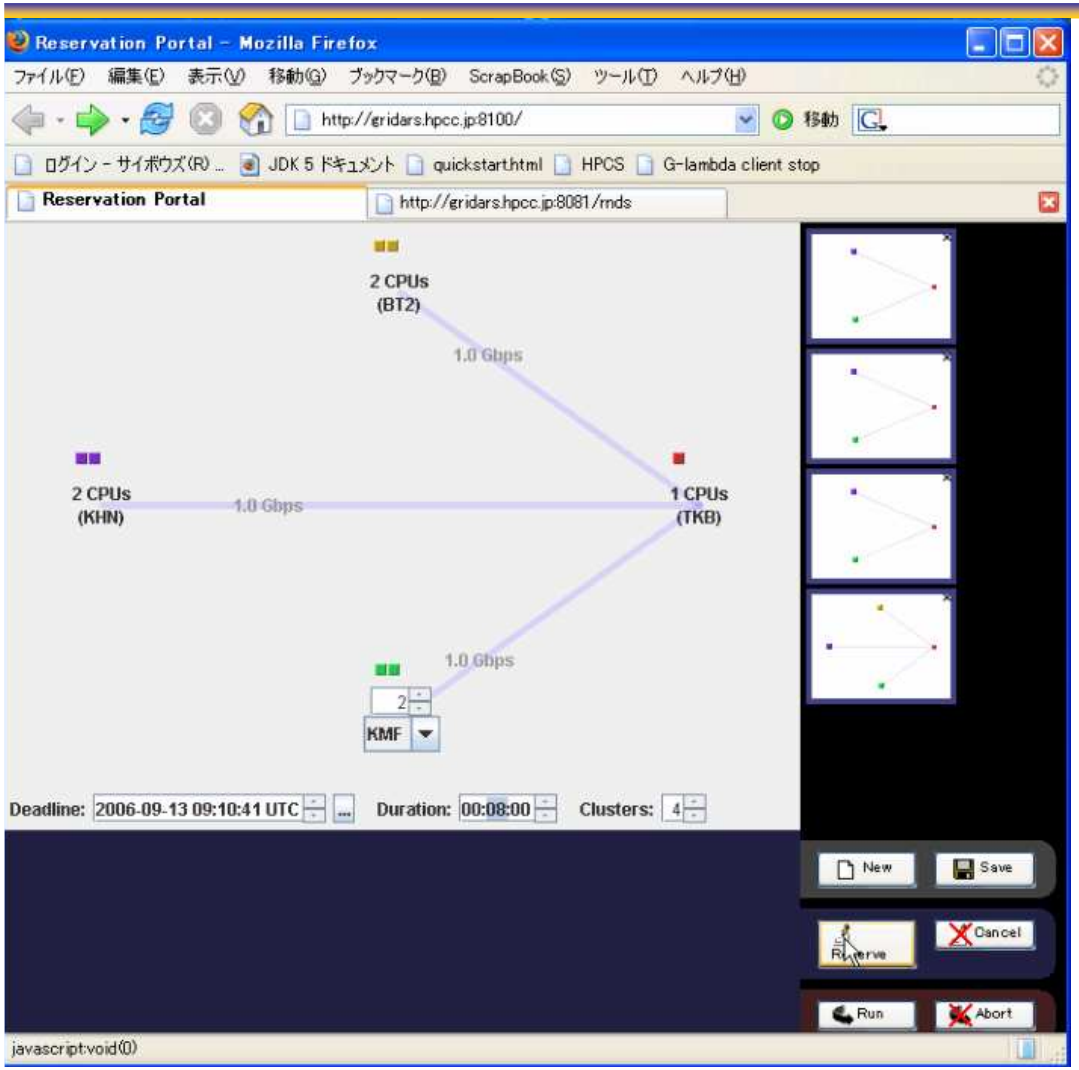
GridMPI [産総研]

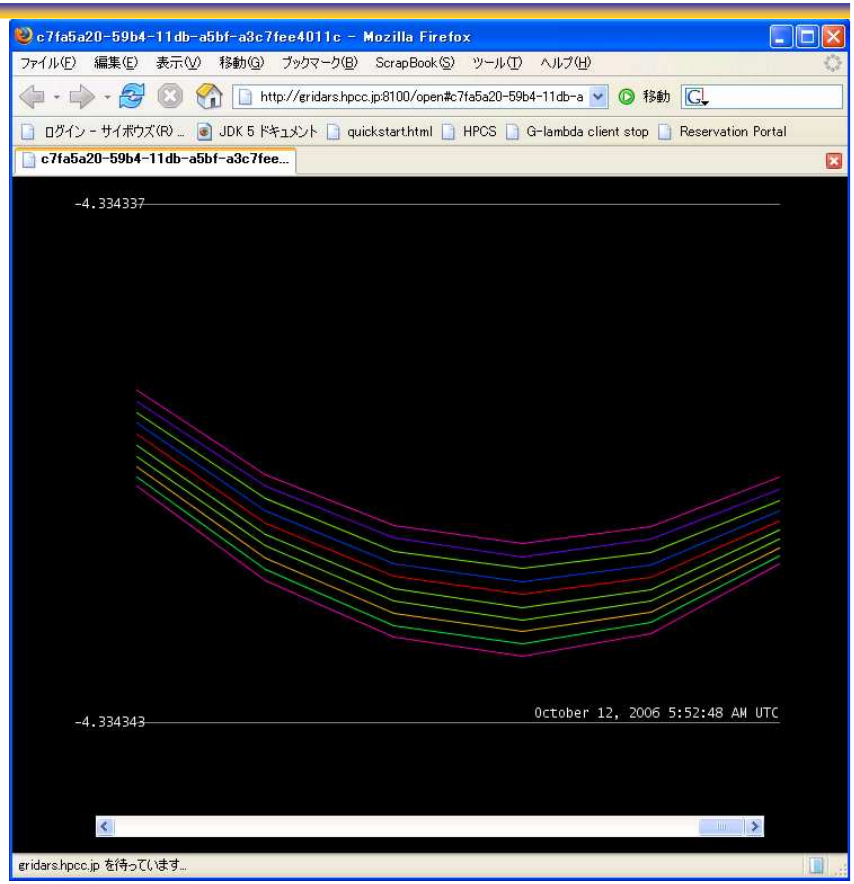
- グリッド環境用に開発されたMPI実装
  - 複数アーキテクチャの混在する環境への対応
  - TCP/IP通信性能
- Interoperable MPI (IMPI)標準で定義される通信プロトコル拡張
  - クラスタ内は独自プロトコル
  - クラスタ間はIMPIプロトコル
    - IMPIサーバを介して各クラスタのアドレス・ポート情報の交換
- gridmpirunコマンドでIMPIサーバ，MPIプロセスの起動をサポート
  - GT2, GT4, ssh経由でのプロセス起動
  - GT4版では，globusrun-wsでプロセス起動

ポータルの構成


Resource Requirement Editor on Web Browser
Result Viewer on Web Browser
(1) Input resource requirements
(2) Send "reserve" req via HTTP
(6) Return the reservation result
(7) Launch result viewer on Web browser and send "run" req
(11) Draw the results
Web Server
Reservation Module
GridARS Client API
Application-dependent Module
gridmpirun
Database
Write/read reservation info
(10) Receive simulation results
(3) Send reserve req via GridARS 2PC WSRF
(5) Get reservation result
(8) Submit jobs in the reserved queues using globusrun-ws
GridARS GRS
(4) Co-allocate distributed resources via GridARS 2PC WSRF
NRM
NRM
NRM
CRM
WS GRAM
CRM
WS GRAM
CRM
WS GRAM
PluS +SGE
PluS +SGE
Maui +TORQUE
GridMPI
(9) Start QM/MD simulation using GridMPI
Grid Technology Research Center AIST
AIST

ポータル画面(資源予約入力，結果出力)





- ポータル画面からの簡単な操作で日米間に跨る資源での並列計算が可能

Grid Technology Research Center AIST

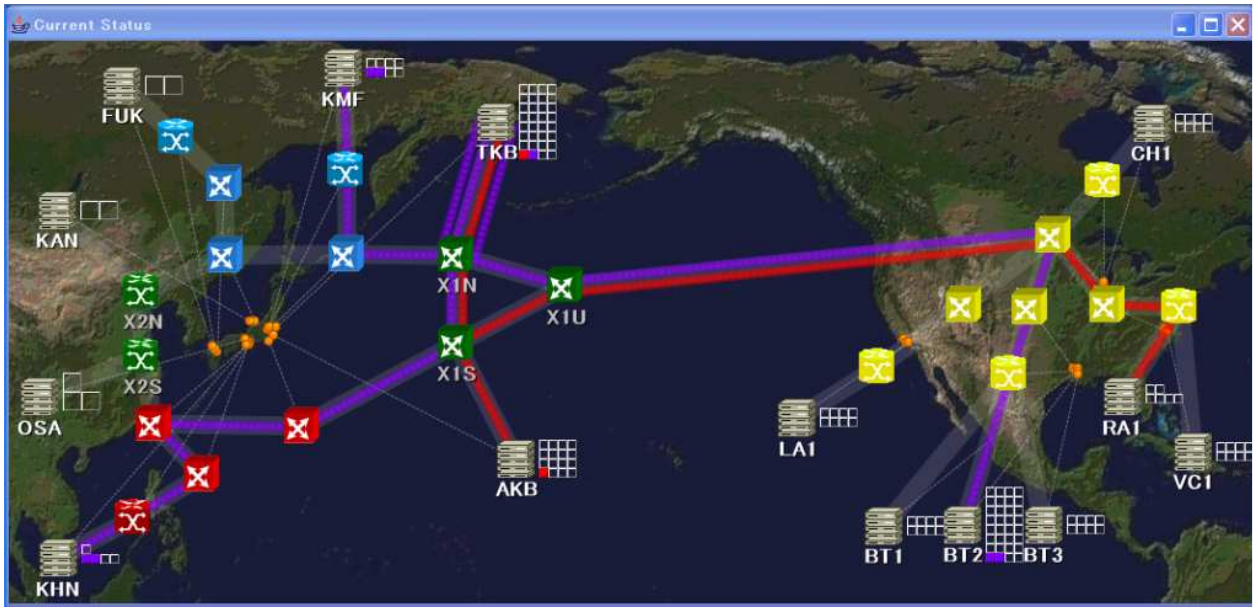
AIST

ポータルの実装

- 事前予約HTTPインタフェース
  - 通常のHTTPリクエスト(GET/POST)
    - サービスURL/reserve, /cancel, /list, /load, /save
- ウェブサーバ
  - 軽量なJava組み込みHTTPサーバ OOWeb
    - Javaオブジェクトをウェブページとしてマッピング
    - GridARS Java APIで資源予約
- セキュリティ
  - MyProxyによりユーザ証明書を取得
  - GridARSでは証明書を委譲
  - ジョブ投入も同様に行える→シングルサインオン

ポータルを用いた実証実験

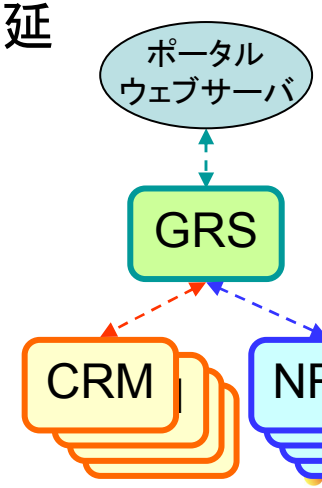
- **2006年9月11-13日のG-lambdaプロジェクトと米国EnLIGHTtenedプロジェクトとの共同実証実験**
  - 日米間に跨る複数資源上で並列計算
- **実験環境**
  - サイト数：9
    (国内7，米国2)
  - CRMの構成：
    - GridARS-WSRF
    - PluS,SGE/
      Maui/TORQUE
    - 計算ノード：
      Linux系OS，X86



  - NRM：GNS-WSI2インタフェースでサービスを提供する
    4つのNRMと連携
    - KDDI研，NTT，EnLIGHTened，AIST

ポータルを用いた実証実験結果

- 他のユーザがいる環境で，分散する資源を10分間事前予約で確保し，予約時刻にQM/MD連成シミュレーションを開始するデモを繰り返し行った
  - GridARS GRSが4つのCRM，4つのNRMに同時にreserve/commitを実行→安定して動作
- 事前予約所要時間　*reserve(P→G)*　*reserve(G→RMs)*　*status取得(G→RMs)*
  - $RsvTotal=a+(a+a\times n+b\times(n-1))+a+c$
    $=3a+(a+b)n-b+c$　*ポーリング間隔*　*commit(P→G)*
    - ***a***: WSRF+GSIのオーバヘッド(0.6s)+通信遅延
    - ***b***: 状態を取得する際のポーリング間隔(1s)
    - ***n***: ポーリング回数(n=4)
    - ***c***: create処理を含むGRSでのその他の処理
    - 実験ではRsvTotal = 8 秒程度
  - RsvTotalはRMの台数には依存しない



関連研究

- **Moab Grid Suite [Cluster Resources]**
  - Moab LS+ヘテロキューイングスケジューラ(QS)
  - 商用でモニタリング･レポーティング等のツールが豊富
- **CSF (Community Scheduler Framework) [Platform]**
  - 事前予約はLSFでのみ有効
  - GridARS同様，WSRF/GSIインタフェース
- **GUR [SDSC]**
  - Catalina+ヘテロQS
  - GSI-enabled SSHでLSの予約コマンドを実行
- **2相コミット，ネットワーク資源との連携を実現しているのは GridARSのみ**

まとめ

- **GridARSを改良し，階層的な2相コミットでの事前予約手続きを実現**
  - 多様な資源のためのGridARS-WSRF I/Fモジュールの開発
  - 安全な分散トランザクション
- **改良したGridARSを用いてポータルを構築**
  - GridMPIで実装されたQM/MD連成シミュレーション
  - 日米間に跨る複数スケジューラで管理される資源上で容易に並列計算が実行できることを実証

今後の課題

- **GridARS**
  - 課金，SLA
  - グローバルコスケジューリングアルゴリズム
  - WS-Notification
    - 資源の状態変化(故障等)の通知
  - 事前予約インタフェースの標準化
- 資源スケジューラへの要求
  - 2相コミット，前/後処理のサポート→PluSでは実現
- ポートレットによるモニタリング等の関連ツールの統合

謝辞

- **G-lambdaプロジェクトの皆様に感謝いたします**
  - http://www.g-lambda.net/
- 本研究の一部は，文部科学省科学技術振興調整費「グリッド技術による光パス網提供方式の開発」による