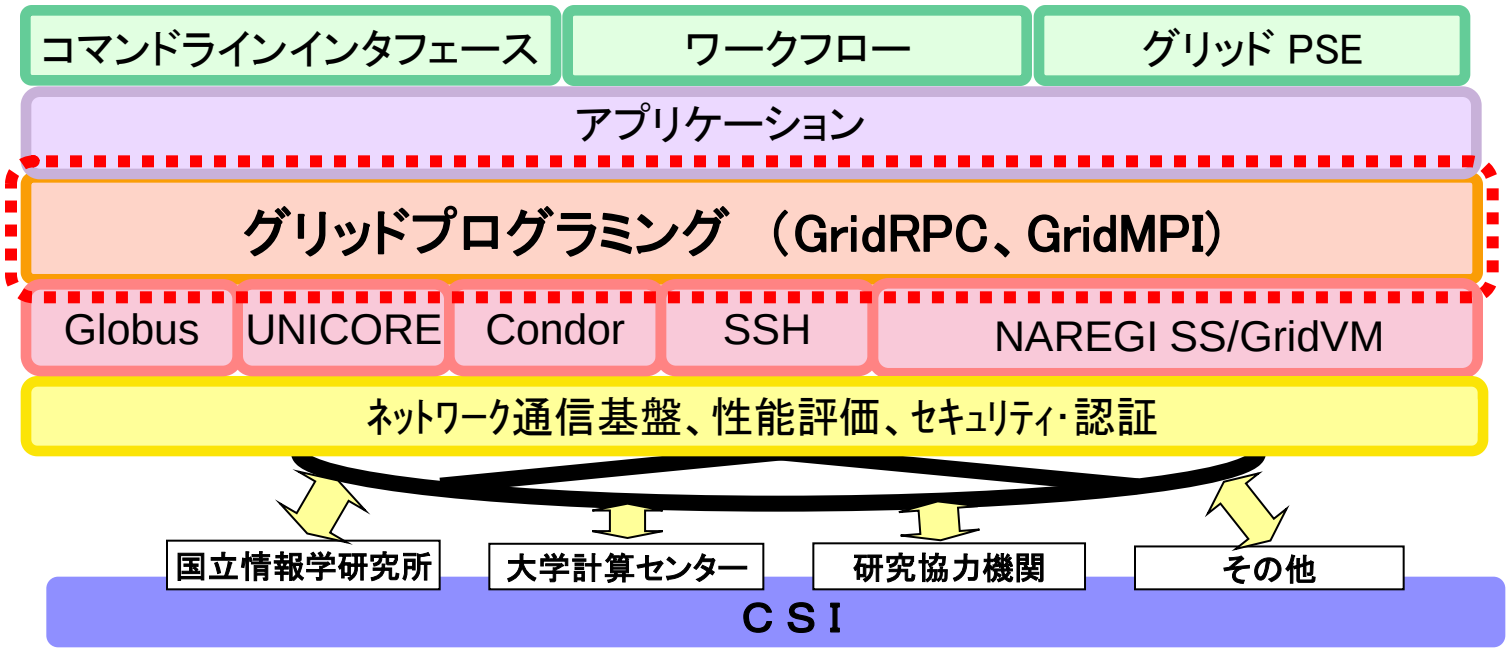
CSIにおけるグリッドプログラミングミドルウェアの高度化

独立行政法人
産業技術総合研究所

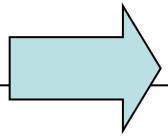
委託事業の目的および内容

複数の計算資源が高速ネットワークで接続されているCyber Science Infrastructure(CSI)において、単体のスーパーコンピュータからグリッドに至る様々な計算基盤上で高い性能を発揮し、長時間安定して動作するグリッドアプリケーションの開発、実行を支援するプログラミング環境を構築・提供する、Grid Message Passing Interface (Grid MPI)およびGrid Remote Procedure Call (GridPRC)に基づくプログラミングミドルウェアについて、機能拡張や性能改善などの高度化を容易に実現するために、内部仕様書等のドキュメント整備を行う。



グリッドプログラミング環境の目的と指針

- 目的
  - （資源）遠隔地に設置され、高速なネットワークで接続された複数のコンピュータ（クラスタやSMP機）
  - （環境）標準的なグリッド機能(セキュリティ、情報サービス、実行環境）
  - VO/VCにおけるプログラム/ソフトウエア開発環境を提供
- 設計指針
  - 従来の並列プログラミングとの親和性がよいこと
    - 大規模計算ユーザのプログラム移行性
  - グリッド環境をユーザが意識しないこと
    - 独特のコマンド等を覚える必要がない
  - 小規模から大規模まで安定して拡張できること
    - 高いスケーラビリティ（100TFlops規模のグリッドでも動作する）



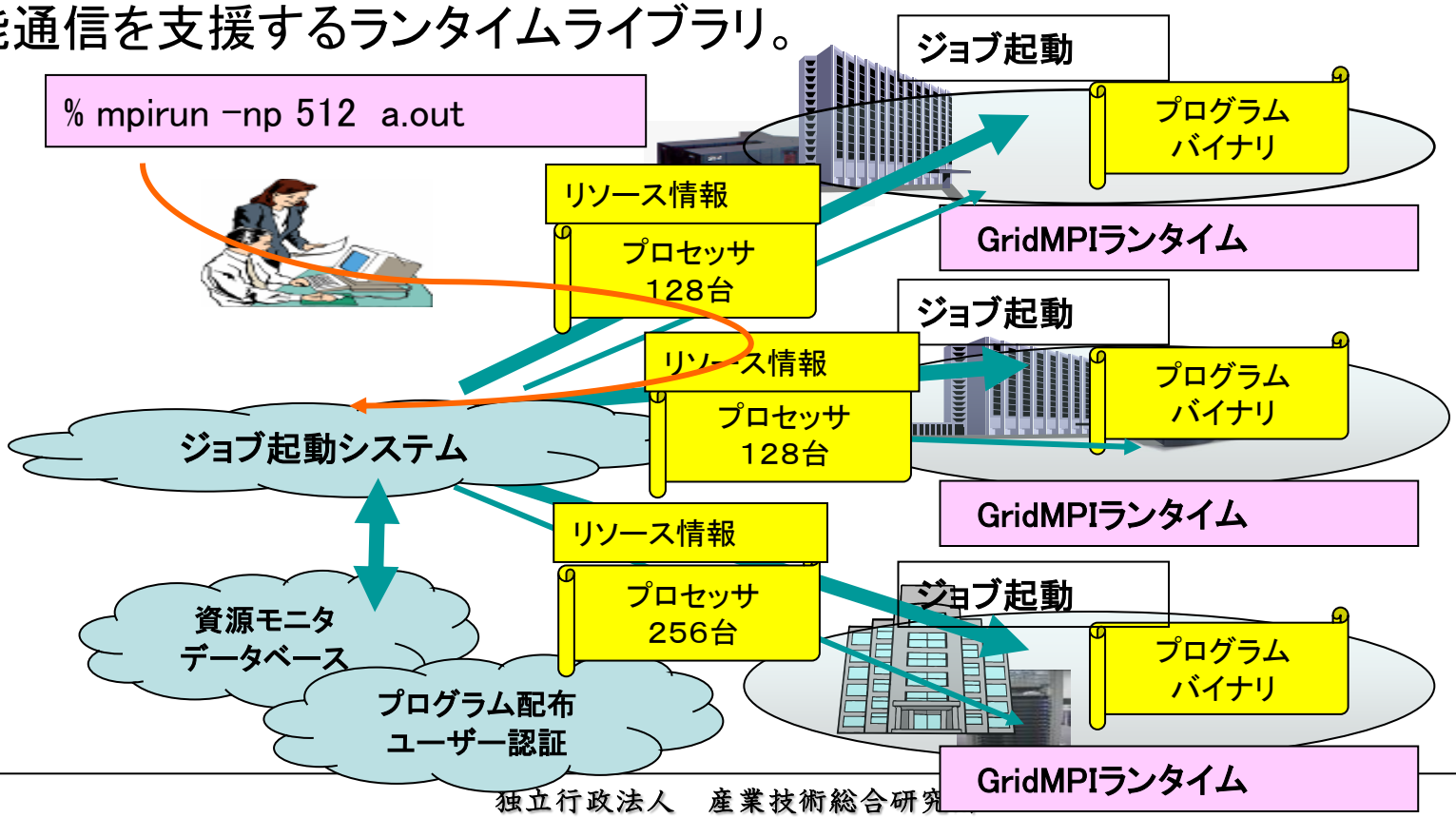
Grid RPC / Grid MPI を採用

Grid MPI と Grid RPC

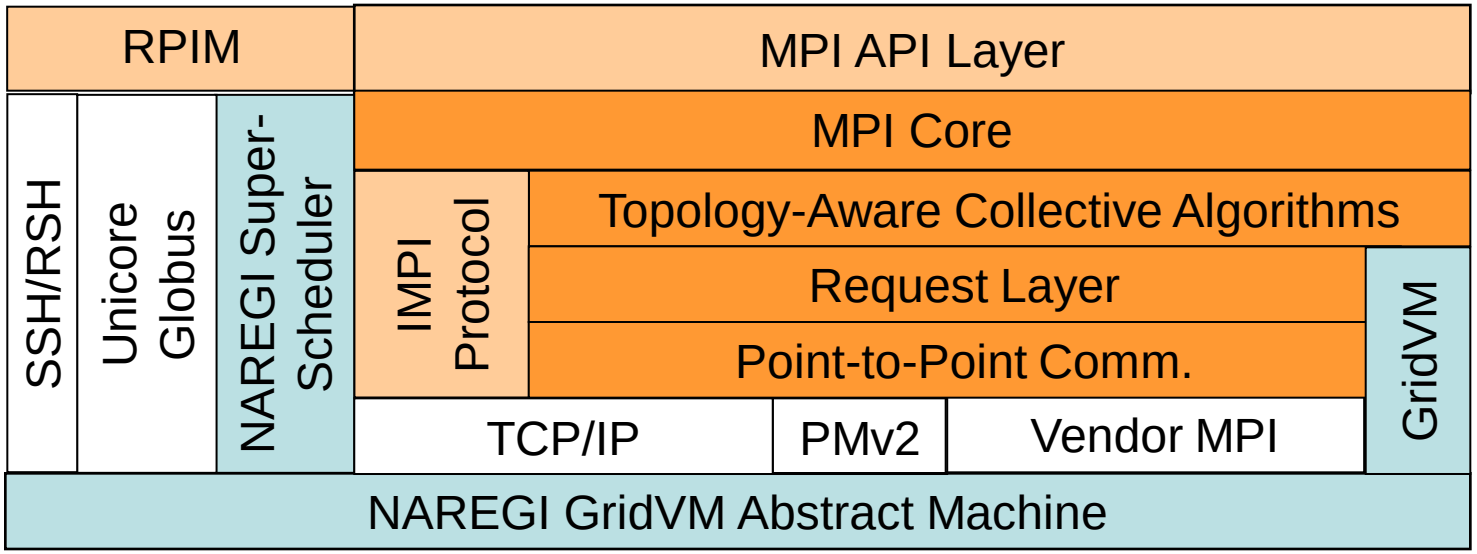
	Grid MPI	Grid RPC
特徴	並列プログラムの構造を理解する必要があり、若干困難であるが、記述能力は高い。並列性が複雑に絡み合う場合に有効。	特定の計算部分を遠隔実行するだけなので、VO/VCにおけるプログラムが容易。比較的単純な並列性を記述する場合に有効。
	MPI プログラムがそのまま動作可能	RPC呼出手続きの記述を若干追加
GGF 関連 WG	IMPI/OpenMPI	Grid RPC API (OGF standard)
主な実装	MPICH-G/2, PACX-MPI, STAMPI, Grid MPI, OpenMPI	Ninf, NetSolve

GridMPIの概要とアーキテクチャ

- MPI-2に準拠した通信ライブラリを提供（最新版は2.1.4）
- グリッド上での通信遅延を考慮した高性能通信およびグリッド上でのインターオペラブル通信の実現。
- TCP/IPレベル、MPIライブラリレベルでの通信ライブラリの開発、バイナリAPIレベルおよび通信プロトコルの規格化および実装を行なう。
- ジョブ管理モジュール（NAREGI SSなど）によってプロセスが生成された後の高性能通信を支援するランタイムライブラリ。



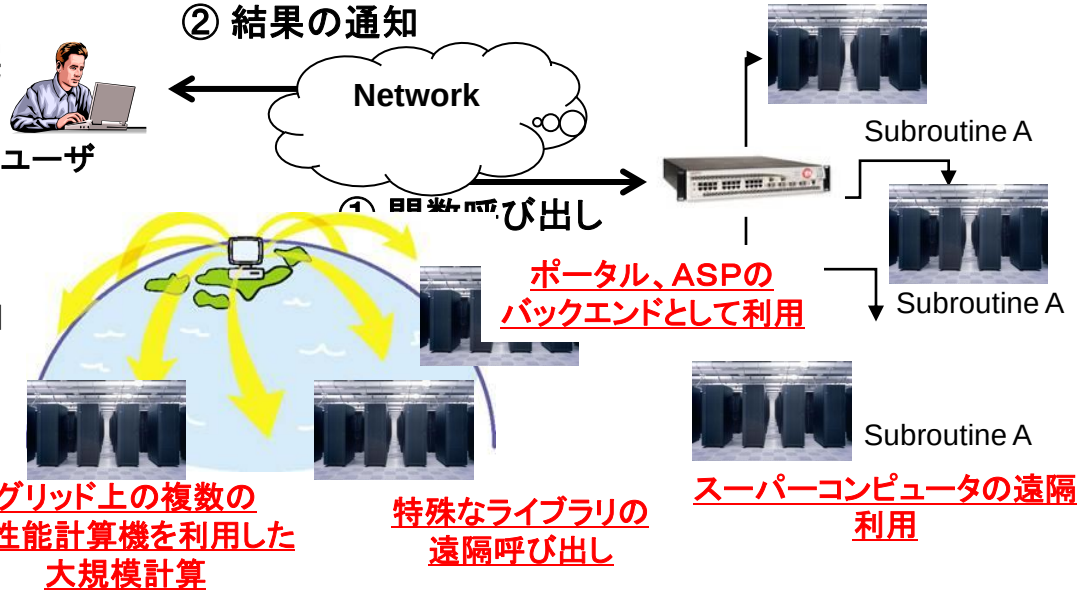
GridMPI ソフトウェアアーキテクチャ

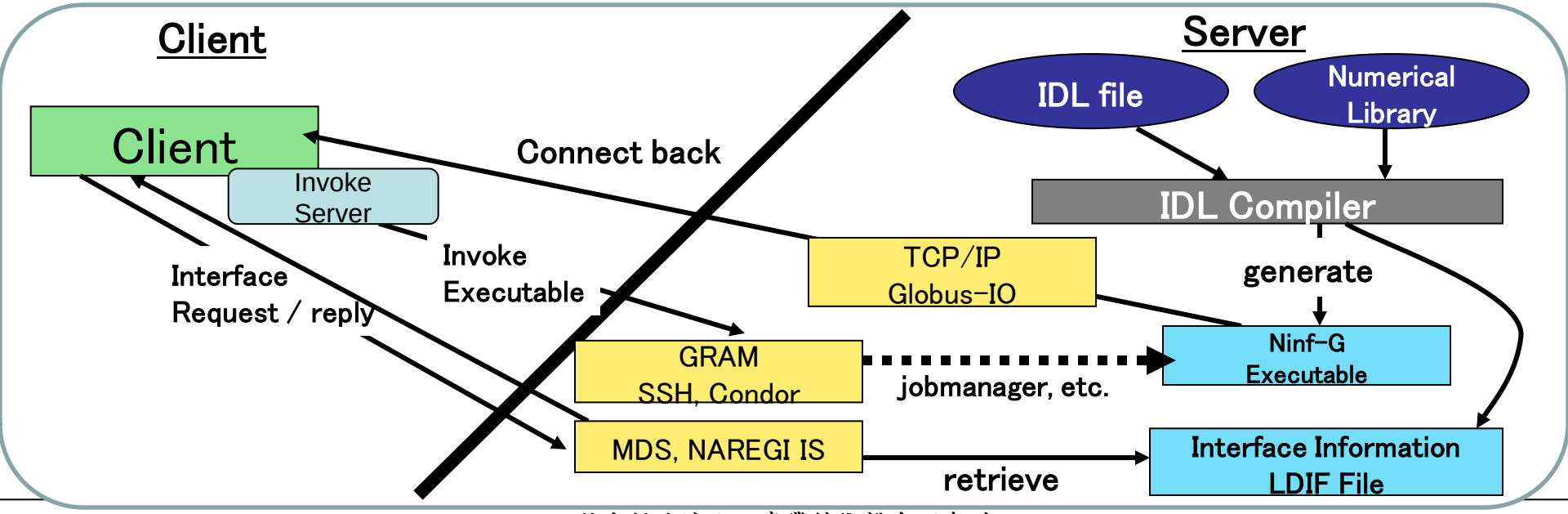


- MPI Core & MPI API Layer
  - コミュニケータ、グループ、その他を実現し、トポロジ層を利用して通信を実現
  - MPI Core はYAMPII (東大開発、オープンソース) を使用
- RPIM: リモートプロセス生成機構
- Topology-Aware Collective Communication Algorithms
  - 通信遅延＆ネットワークトポロジ透過（トランスペアレンシ）を提供
- Point-to-Point Communication
  - プロセス対プロセス通信

GridRPCの概要とNinf-Gのアーキテクチャ

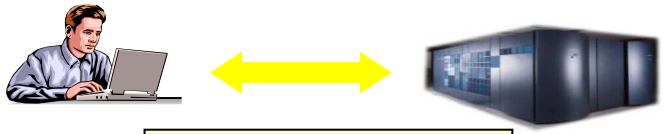
- GridRPCによるプログラム開発および実行を支援するソフトウェアパッケージ。
- NAREGI SS, GRAM, Condorなど様々なミドルウェアを介したRPCが可能。
- タイムアウトなど障害を検知するための豊富な機能により、頑健なアプリケーションの開発、実行を支援。
- 記述力の豊富なコンフィグレーションファイルにより、サーバ側の細かなレベルの非均質性に対応。



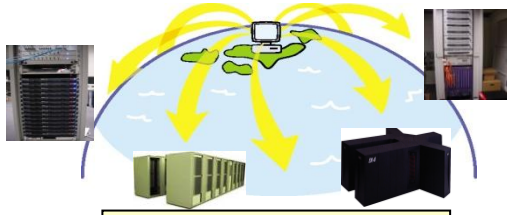


従来技術との比較

- ネットワークの高速化に伴い、ネットワークで接続されたスーパーコンピュータを束ねて同時に利用して大規模な科学技術計算を行うグリッドコンピューティングへの期待が高まっている。



Desktop Supercomputing



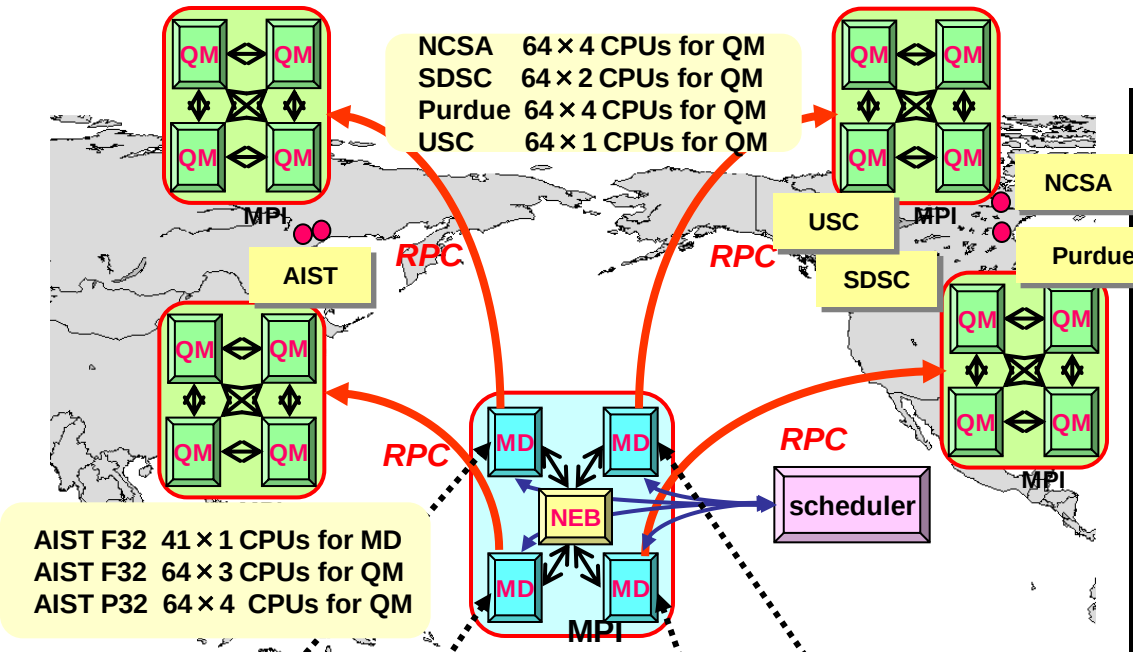
Task parallel processing

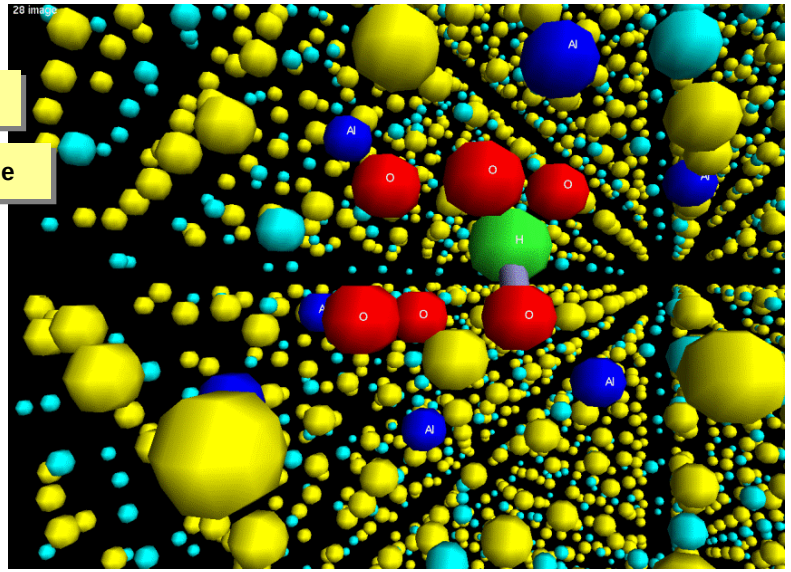
- グリッドには単体のシステムにはない特有の性質があり、プログラミングも既存の手法をそのまま適用できない。
  - セキュリティ、非均質性、計算機の可用性、耐障害性、スケーラビリティなどに対応する必要がある。
- 安定して効率の良いプログラムを開発する方法が確立されておらず、アプリケーションの開発が困難であった。

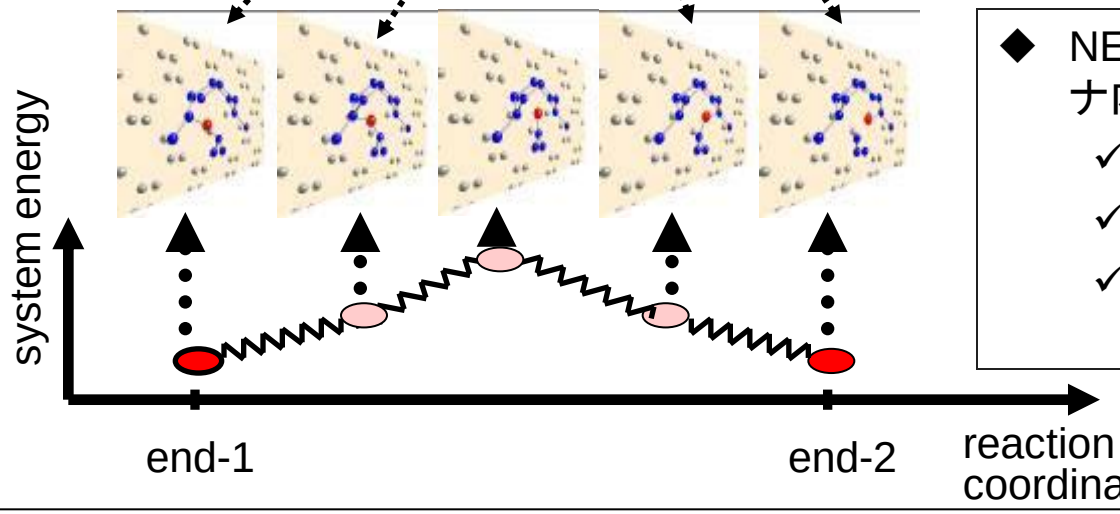
従来技術との比較

	本研究	遠隔手続き呼び出し	CORBA	MPI	他のグリッドミドルウェア
科学技術計算向け	○	×	×	○	○
プログラミングインタフェースの提供	○	○	○	○	×
グリッド固有の性質への対応	○	×	×	×	○
分散環境への対応	○	○	○	×	○
大規模並列処理	○	×	×	○	○

実績：日米グリッド上での大規模実証実験（2007年）

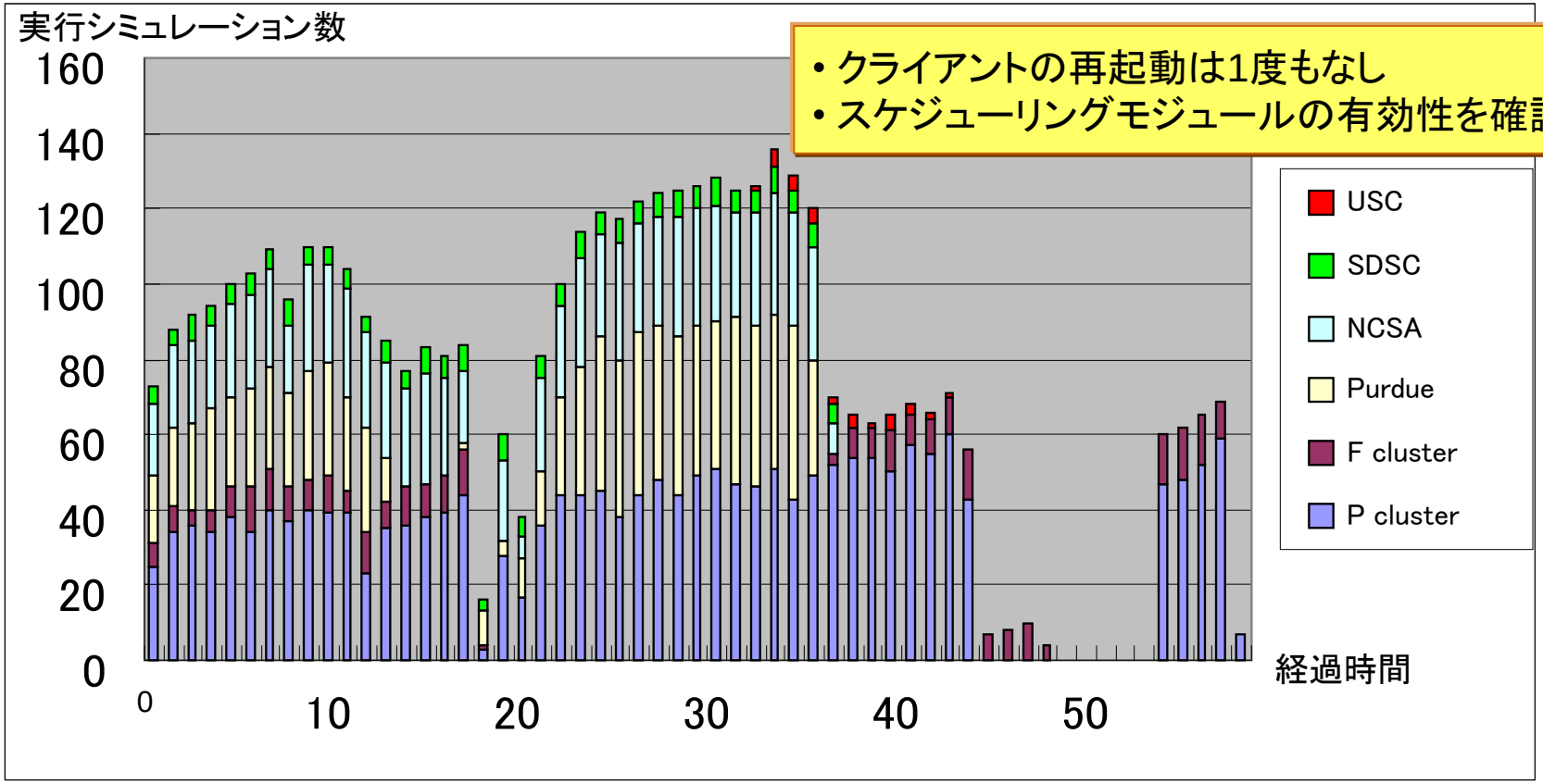




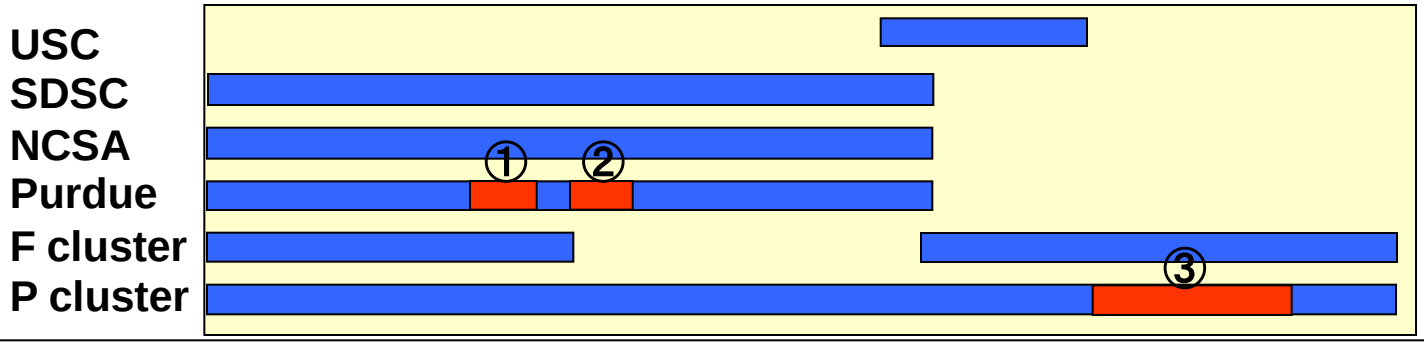


- NEB+Hybrid QM/MD法によるガンマアルミナ内での水素原子拡散シミュレーション
  - ✓ GridRPC + MPI によるプログラミング
  - ✓ 日米グリッド上で実行(1193 CPUs)
  - ✓ 約58時間にわたり、5000 QM シミュレーションを実行

実行履歴



- クライアントの再起動は1度もなし
- スケジューリングモジュールの有効性を確認



① NFSトラブル
② PBSトラブル
③ NFSトラブル

21年度の業務：内部仕様書の作成（1／2）

- **コンポーネントを構成する全てのパッケージおよびクラスについて、下記表に示す情報を記載した。**

項目	内容
パッケージ一覧	各パッケージ一覧を記述する。内容は次の通り。 パッケージの名称 パッケージの役割
クラスまたは関数一覧	各クラスまたは関数の一覧を記述する。内容は次の通り。 クラスまたは関数の名称 クラスまたは関数の役割
クラス階層	各クラスまたは関数の階層構造をブロック図を用いて記述する。
クラス仕様	クラスまたは関数毎に次の内容を記述する。 クラスまたは関数名称：クラスまたは関数の名称 クラスまたは関数の機能概要：クラスまたは関数の機能概要を記述する。 フィールドまたは変数：フィールド変数または変数の説明を記述する。 コンストラクタまたは引数：コンストラクタの引数または引数の説明を記述する。 メソッドまたは関数：メソッドまたは関数の機能概要、処理概要を説明する。また、引数、戻り値、例外について説明する。

21年度の業務：内部仕様書の作成（2／2）

- コンポーネントが、NAREGIミドルウェアの他のコンポーネント（外部コンポーネント）から利用可能なサービスを含む場合、該当するサービスのインタフェースについて、下記表に示す情報を記載した。

項目	内容
サービス名	対象とするコンポーネントが外部コンポーネントに提供するサービス名を記述する。
サービス概要	上記サービスの内容を記述する。
メソッド	外部コンポーネントが上記サービスを利用するためのメソッド名、および本メソッドの引数、戻り値、例外にについて説明する。

- コンポーネントを構成するクラスのうち、プラットフォーム（OSやバッチスケジューラ等）に依存する実装がある場合は、該当するクラスの名前および実装依存部分の解説を記載した。

21年度の成果

- GridMPI、GridRPCのそれぞれについて、内部仕様書を作成。
- GridMPI
  - Version 2.1.4内部仕様書（130ページ）
- GridRPC
  - Version 4.2.5内部仕様書（464ページ）
  - Ninf-G Invoke Server仕様書（13ページ）
  - Ninf-G プロトコル仕様書（42ページ）

まとめと今後の展開

- まとめ
  - GridMPIおよびGridRPCの内部仕様書を作成した。
- 今後の展開
  - CSI上での実証利用
  - Peta～Exa Flops級の大規模アプリケーションの実装・実行に向けてのフィージビリティスタディ
  - ユーザサポート
  - HPC Cloudにおけるプログラミング環境