広域分散アプリケーション特論

2005前期　月曜　3時限 21世紀交流プラザ2F講義室

担当　青柳　睦
aoyagi@cc.kyushu-u.ac.jp

4月25日 (月)

講義の内容, 成績評価方針(再々掲)

グリッドのサイエンス分野での利用

高エネルギー物理, 天文, 気象, バイオ, ‥
サイエンス分野におけるグリッド技術の課題

講義の内容

- グリッドの概要
  - Gridコンピューティングとは
  - ビジネス分野での利用
  - サイエンス分野での利用
- 計算科学の概論
  - 主要なシミュレーション手法
- サイエンスGrid　NAREGI
  - Globus, Unicoreの現状
  - NAREGIミドルウエア概要
  - 連成計算とその類型化
- テーマ　考え中・・・

講義資料はWebで公開

server-500.cc.kyushu-u.ac.jp

レポートの提出先

- aoyagi@cc.kyushu-u.ac.jp
- Subject:　**広域分散アプリ＃**
  ＃は課題番号
- **本文：学籍番号，氏名，専攻**
  レポート(PS, PDF, Word,**添付可**)

メールSubject**には必ず**
「広域分散アプリ(＃課題番号)」と
記入してメールしてください．

成績評価

- 出席点　6割
- レポート　4割
- 前期末試験なし

Gridのサイエンス分野への応用

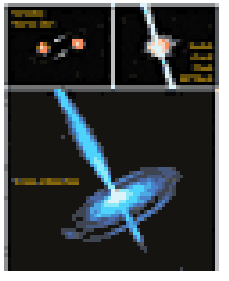
- GriPhyN project http://www.griphyn.org/
  Sloan Digital Sky Survey **等**
- EU-Data Grid project
  http://eu-datagrid.web.cern.ch/eu%2Ddatagrid/
- Earth System Grid https://www.earthsystemgrid.org/
- eScience http://www.nesc.ac.uk/
- KISTI http://testbed.gridcenter.or.kr/kor/
- China National Grid project
  http://www.cs.hku.hk/~clwang/grid/CNGrid.html

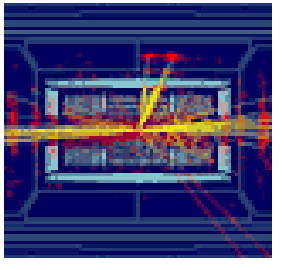
遊休資源利用：インターネットモデル

- SETI@Home http://setiathome.ssl.berkeley.edu/
- GIMPS http://www.mersenne.org/

GriPhyN (Grid Physics Network)

- Astronomy: The Sloan Digital Sky Survey
- LIGO: Detecting Einstein's Gravitational Waves
- High-Energy Particle Physics





The Sloan Digital Sky Survey

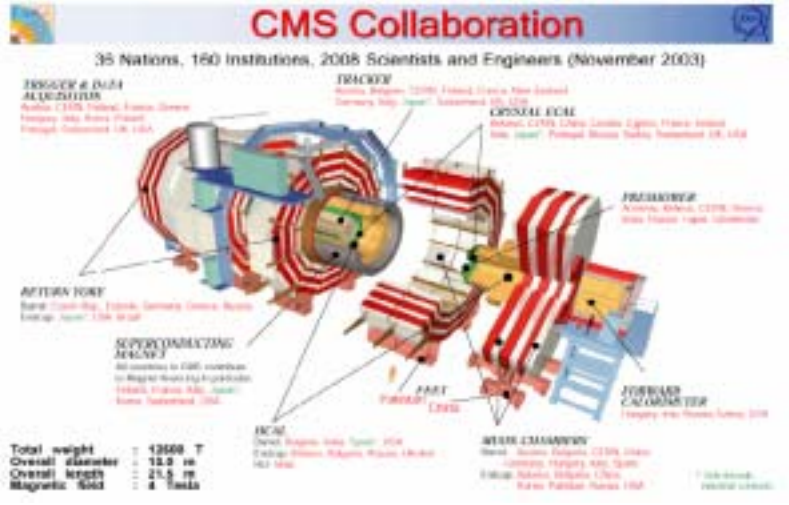
全天の4分の1の領域にある1億個を超える天体の位置と明るさ，赤方偏移を決定

データを全世界に公開

http://skyserver.nao.ac.jp/en/

参加機関：シカゴ大学，フェルミ加速器研究所，国立天文台の日本参加グループ，ジョンス・ホプキンス大学，マックス・プランク天文学研究所，マックス・プランク天体物理学研究所，ニューメキシコ州立大学，プリンストン大学，アメリカ海軍天文台，ワシントン大

大規模観測装置からのデータを広域分散利用

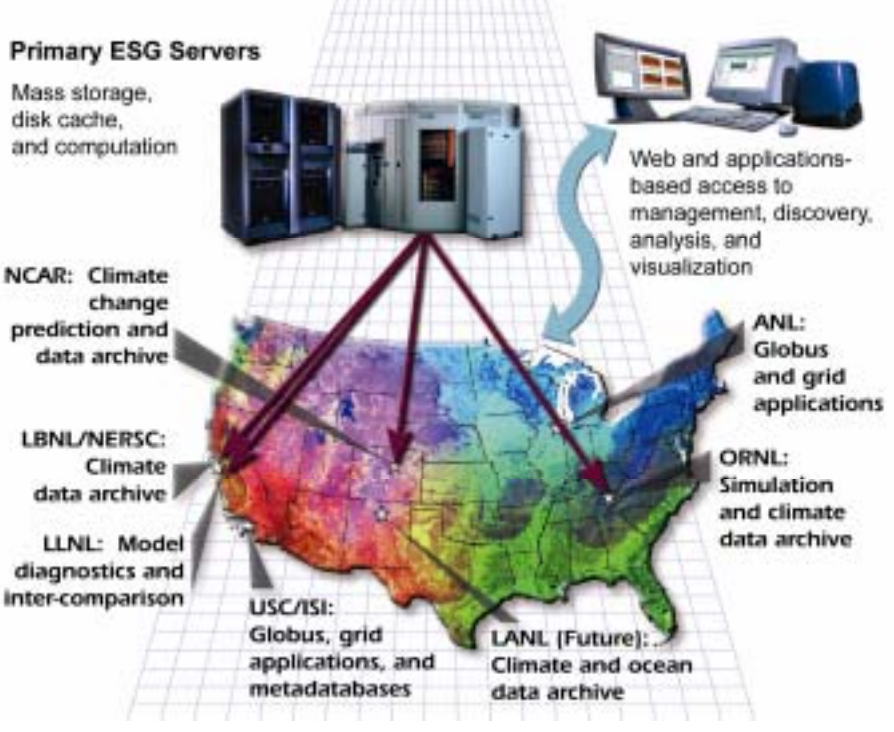


CERNのCMS

観測データおよび2次データが年間数10TB〜数100TBオーダーで増加，しかも分散.

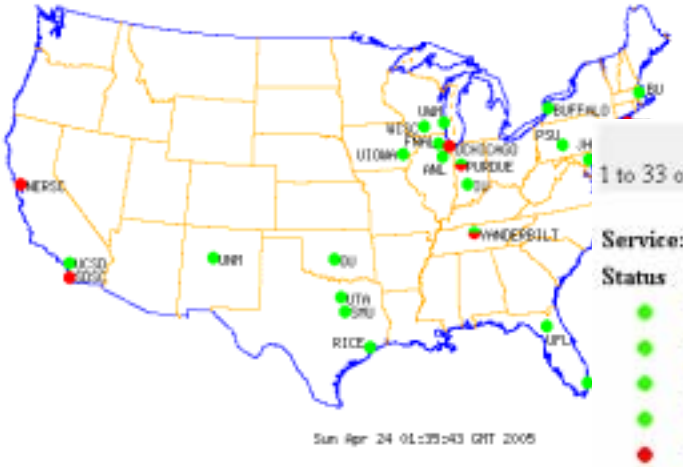
実データとその[カタログ情報]を別管理 (Virtual Data 技術)

Earth System Grid



The Open Science Grid Consortium

GridCat by Grid3 http://butternut.ucs.indiana.edu:8080/



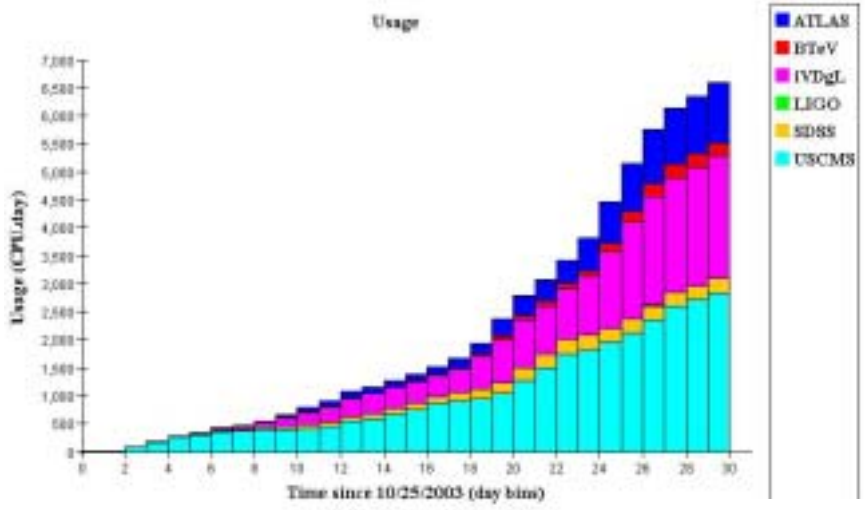
1 to 33 of 33 sort by: Service entries per page: 100 map: U.S. view: Summary

Service: CS(E) = Compute Service (Exemption), SS = Storage Service, LS = Login Service

Status	Site Name	Jobs	Disks	Service	Loc	Facility	CPUs
●	Vanderbilt			CS	TN	VANDERBILT	12
●	UWMilwaukee			CS	WI	UWM	296
●	UCSanDiego			CS	CA	UCSD	3
●	UWMadisonCMS			CS	WI	WISC	64
●	UCSanDiegoPG			CS	CA	SDSC	42
●	UBuffalo-CCR			CS	NY	BUFFALO	80
●	UNM_HPC			CS	NM	UNM	516
●	BNL_ATLAS_1			CS	NY	BNL	20
●	UFlorida-Grid3			CS	FL	UFL	42
●	UFlorida-PG			CS	FL	UFL	82
●	Rice-Grid3			CS	TX	RICE	10
●	PSU_GRID3			CS	PA	PSU	312
●	OU_OSCER			CS	OK	OU	272
●	FIU-CHEPREO			CS	FL	FIU	40
●	FNAL_CMS3			CS	IL	FNAL	216
●	HAMPTONU			CS	VA	HAMPTONU	6
●	UIowa			CS	IA	UIOWA	7
●	PDSF			CS	CA	NERSC	400

グリッド技術を利用して実運用されている広域分散システム

GriphyN VOごとの計算リソースの利用状況

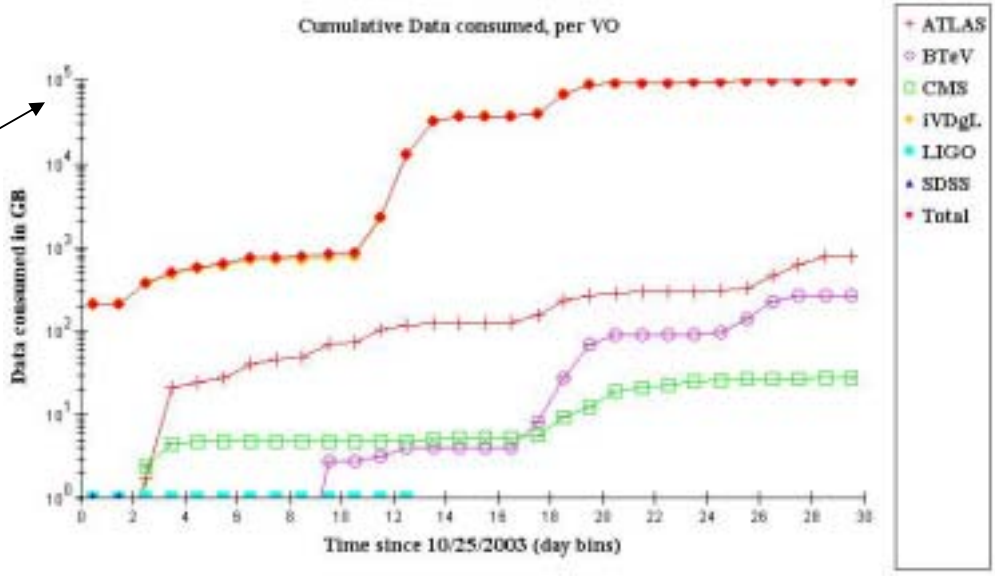


[VO(Virtual Organization)ごと]
という点に注意

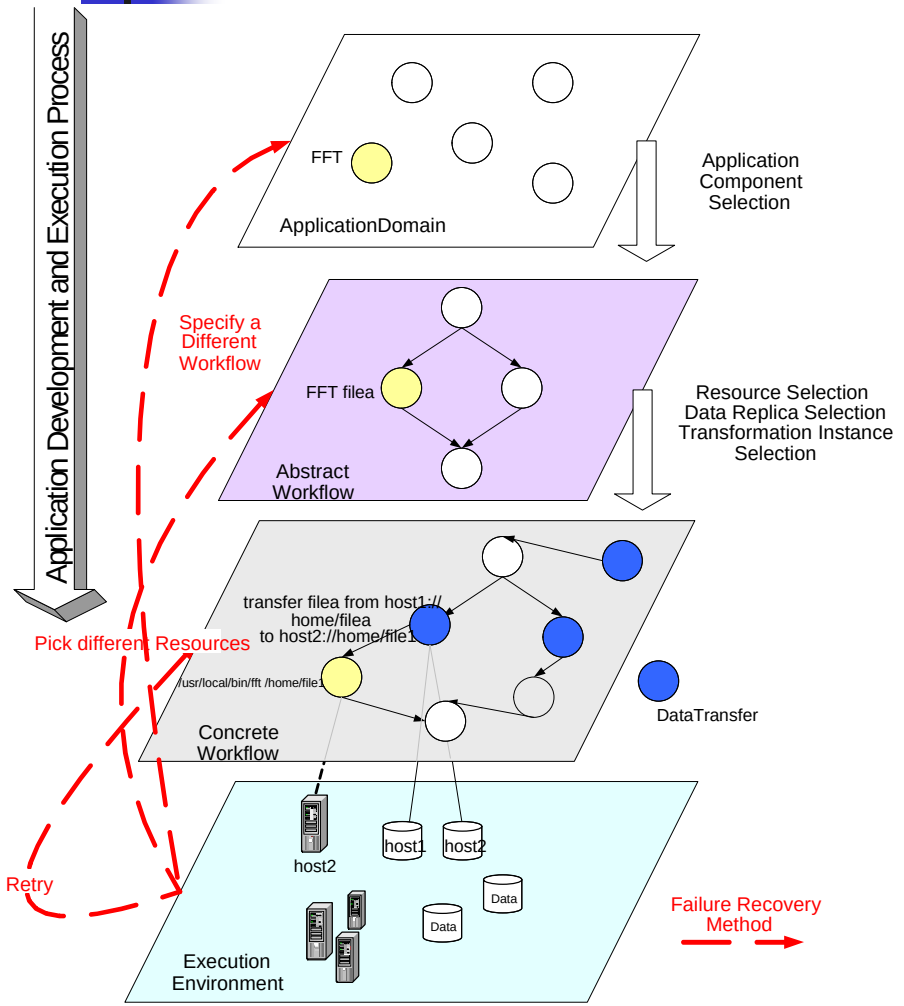
Grid3 – Cumulative Data Consumed
25 Oct ~ 25 Nov 2003

Grid3 – Cumulative CPU Usage
25 Oct ~ 25 Nov 2003

100TB



計算リソースとワークフローのMapping



アプリケーション Domainにおける必要コンポーネントの選択

コンポーネント間を結ぶフロー（抽象的なオブジェクト）

抽象オブジェクトと実リソースへのMapping（Run-Time）

データの所在，アクセス権，移動･複製可否，等を考慮してミドルウェアが判断する部分

障害対応もミドルの仕事

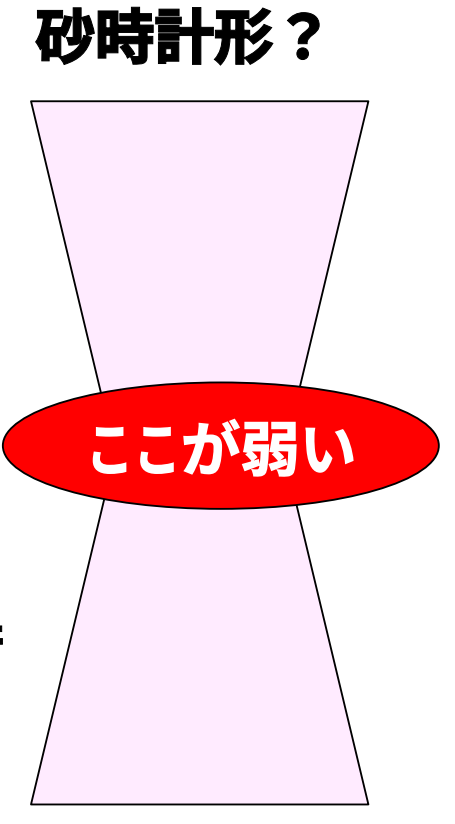
グリッドの階層構造　前回スライドから再掲

- 第四層　アプリケーション層
  分散スーパコンピューティング，データ集約型，
  オンデマンド型，協調研究型のアプリケーション

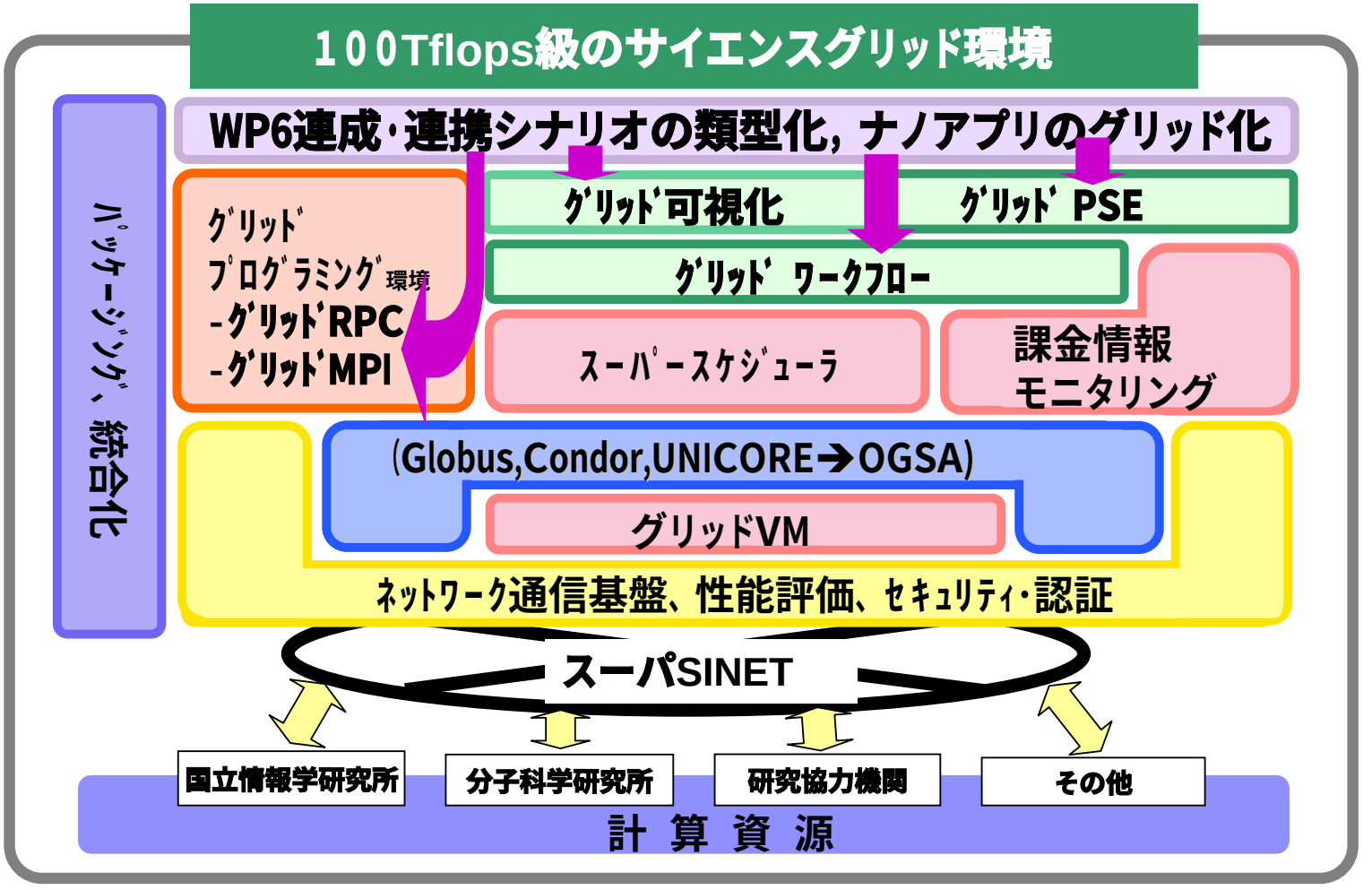
- 第三層　上位ツール
  分散プログラミングツール，可視化，PSE，
  ワークフローツール(二層へのI/F)

- 第二層　共通サービス層
  広域セキュリティー基盤，分散スケジューリング，
  分散リソース管理，広域ネットワークにおける高速通信技術

- 第一層　計算機資源等，インフラ
  ハードウエア，ストレージ，センサー，
  通信基盤・・・・



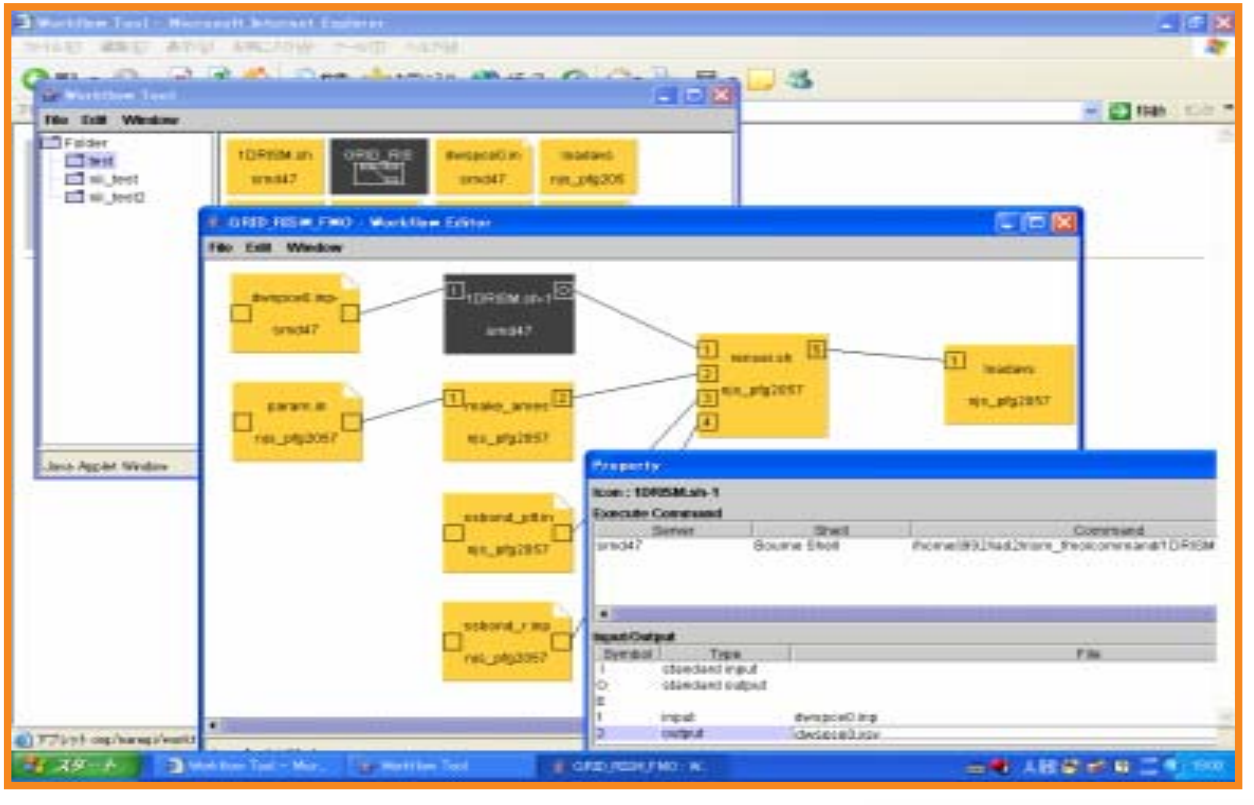
NAREGI の場合


１００Tflops級のサイエンスグリッド環境
WP6連成・連携シナリオの類型化，ナノアプリのグリッド化
パッケージング、統合化
ｸﾞﾘｯﾄﾞ
ﾌﾟﾛｸﾞﾗﾐﾝｸﾞ環境
-ｸﾞﾘｯﾄﾞRPC
-ｸﾞﾘｯﾄﾞMPI
ｸﾞﾘｯﾄﾞ可視化
ｸﾞﾘｯﾄﾞ PSE
ｸﾞﾘｯﾄﾞ ﾜｰｸﾌﾛｰ
ｽｰﾊﾟｰｽｹｼﾞｭｰﾗ
課金情報
モニタリング
(Globus,Condor,UNICORE➔OGSA)
グリッドVM
ネットワーク通信基盤、性能評価、セキュリティ・認証
スーパSINET
国立情報学研究所
分子科学研究所
研究協力機関
その他
計 算 資 源

NAREGI Grid Workflow: 概観

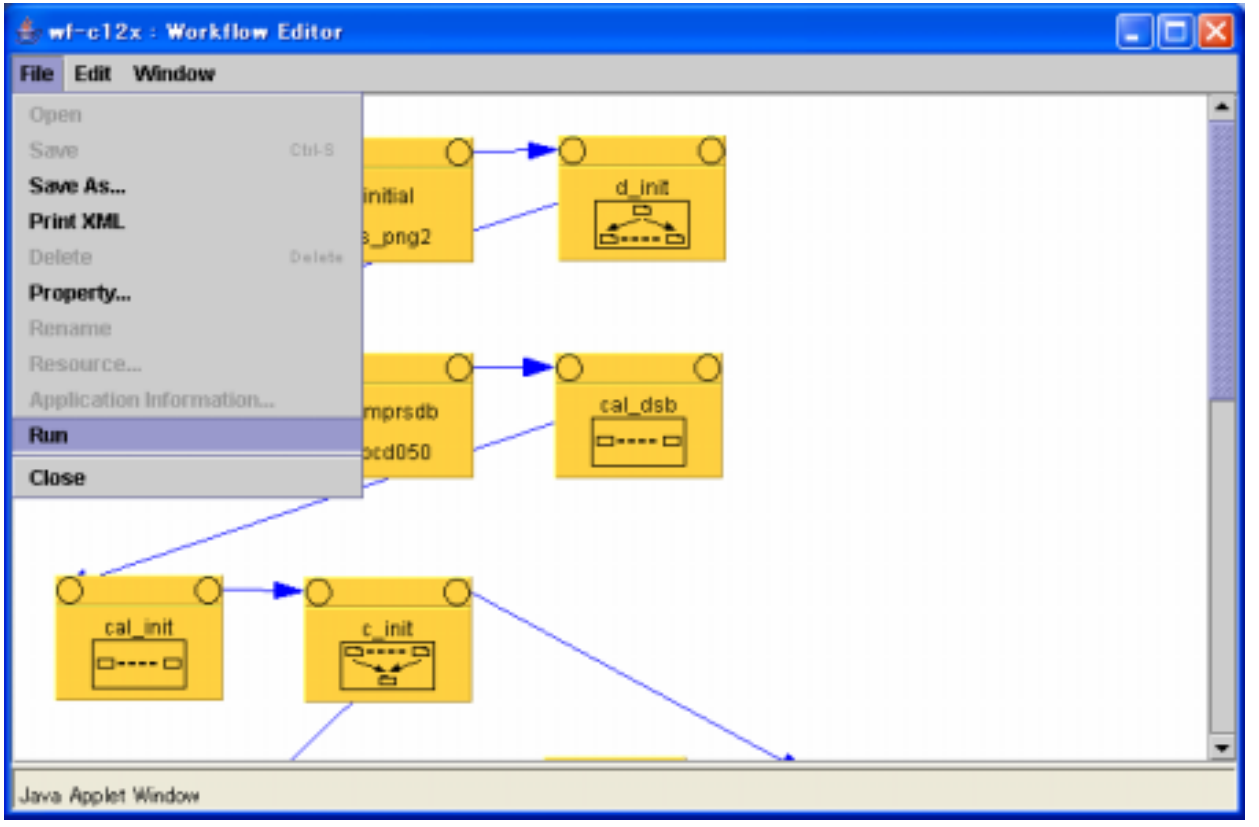
個々のアイコンは application program または input/output data.を表現

実行時の制御フローに従い，プログラムアイコンまたはデータを連結



NAREGI Grid Workflow: job submit

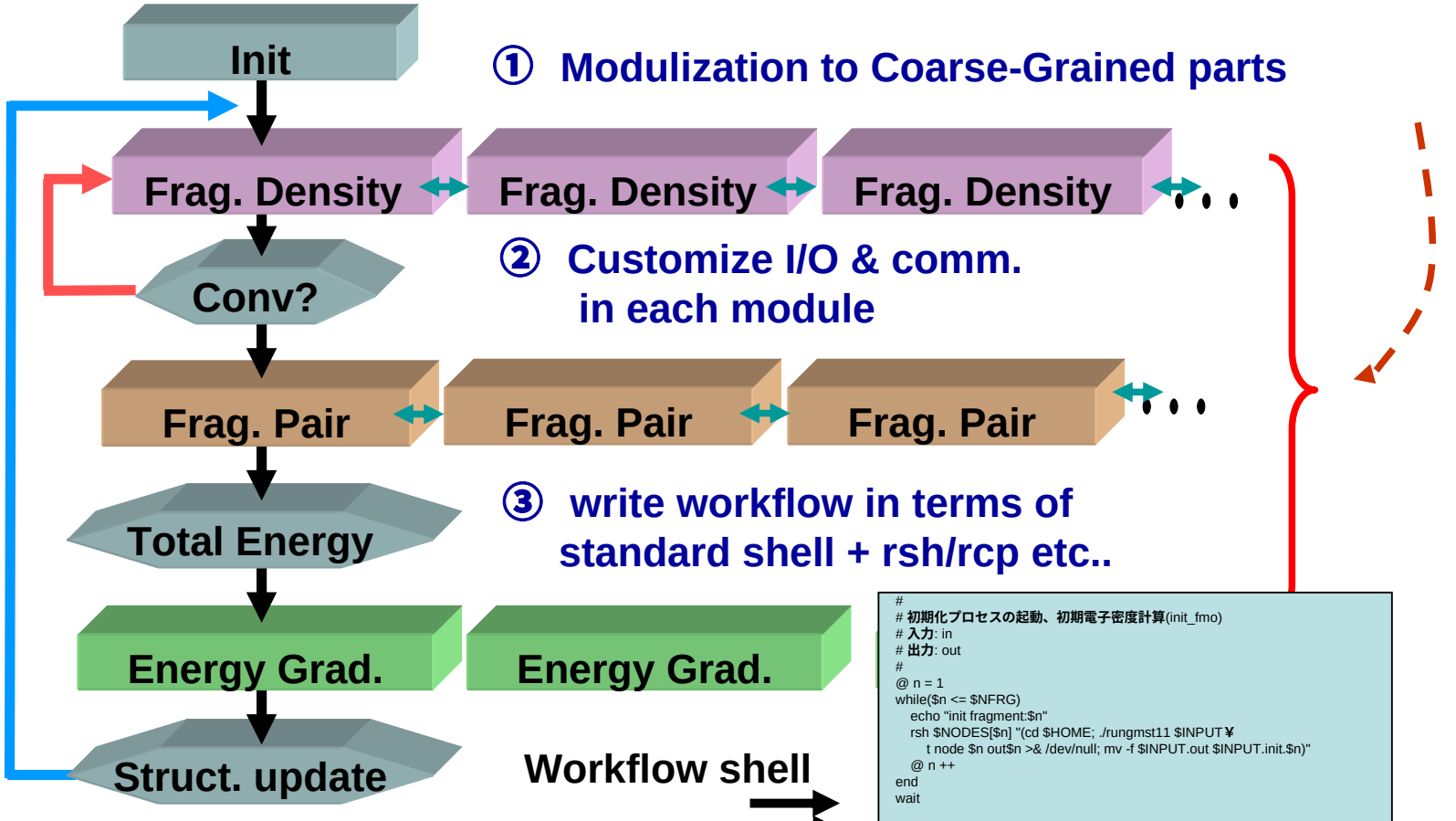
We can run application program according to the computational flow defined on the screen by selecting ‘Run’ menu.



グリッド化の視点　疎結合

- データの分散　vs　処理の分散
- 単機能モジュールの組合せにより複雑な処理フロー (シナリオ) を構成
  - ヘテロな分散計算環境を有効に活用
  - モジュールのグループ化
  - シナリオの柔軟な再構成化に対応
  - モジュール処理からグリッドサービスへ
- モジュールの演算粒度, 独立性, ライフタイム
  モジュール間の通信頻度, 遅延許容度

Grid enabling of the GAMESS FMO prog.



① Modulization to Coarse-Grained parts

② Customize I/O & comm. in each module

③ write workflow in terms of standard shell + rsh/rcp etc..

Workflow shell

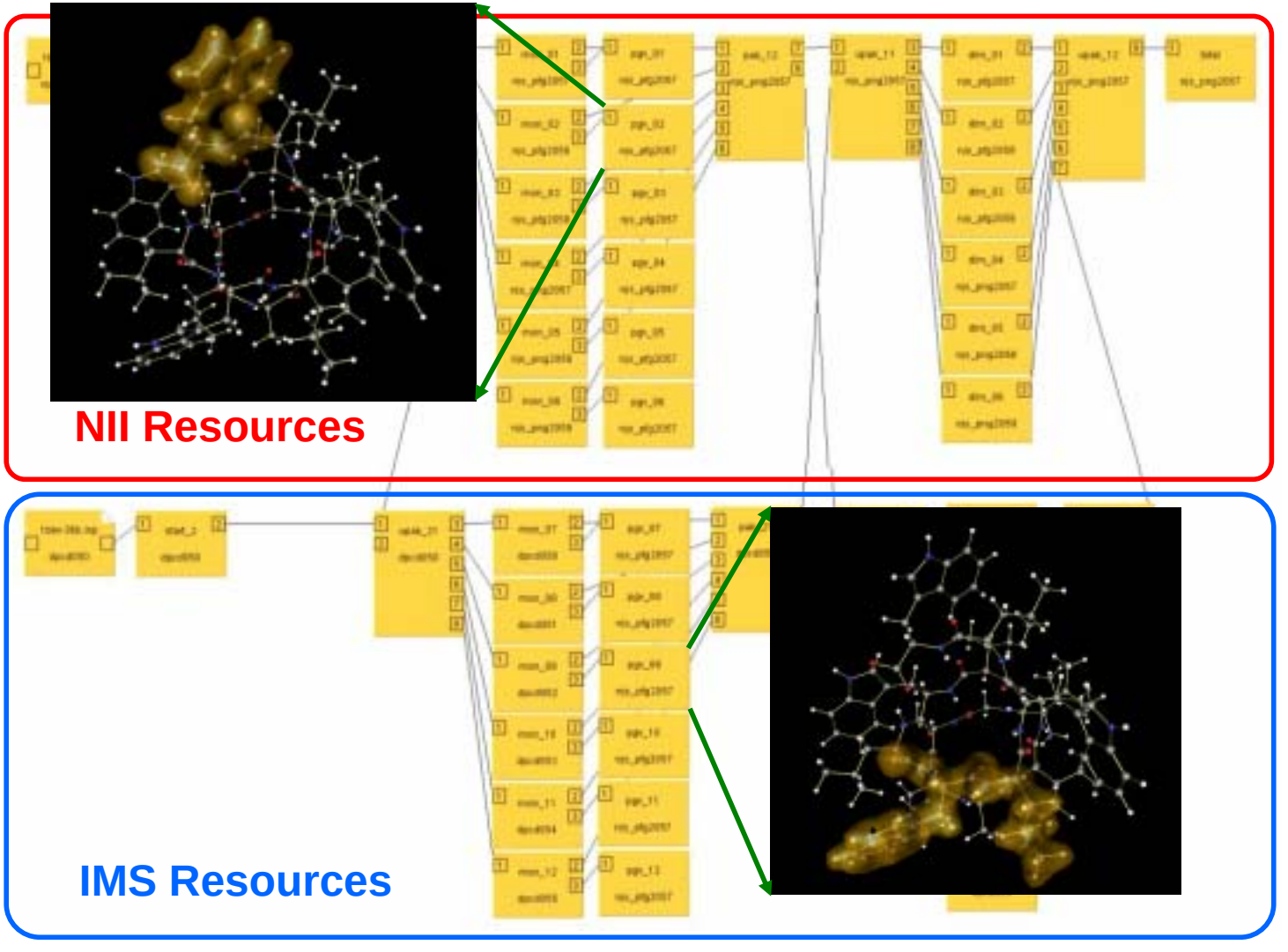
```
#
# 初期化プロセスの起動、初期電子密度計算(init_fmo)
# 入力: in
# 出力: out
#
@ n = 1
while($n <= $NFRG)
   echo "init fragment:$n"
   rsh $NODES[$n] "(cd $HOME; ./rungmst11 $INPUT¥
      t node $n out$n >& /dev/null; mv -f $INPUT.out $INPUT.init.$n)"
   @ n ++
end
wait

# データの収集
@ n = 1
while($n <= $NFRG)
   if($MY != $NODES[$n]) then
      rcp $NODES[$n]¥:$HOME/out$n out$n >& /dev/null
   endif
   @ n++
end
wait
```

Shell scripts

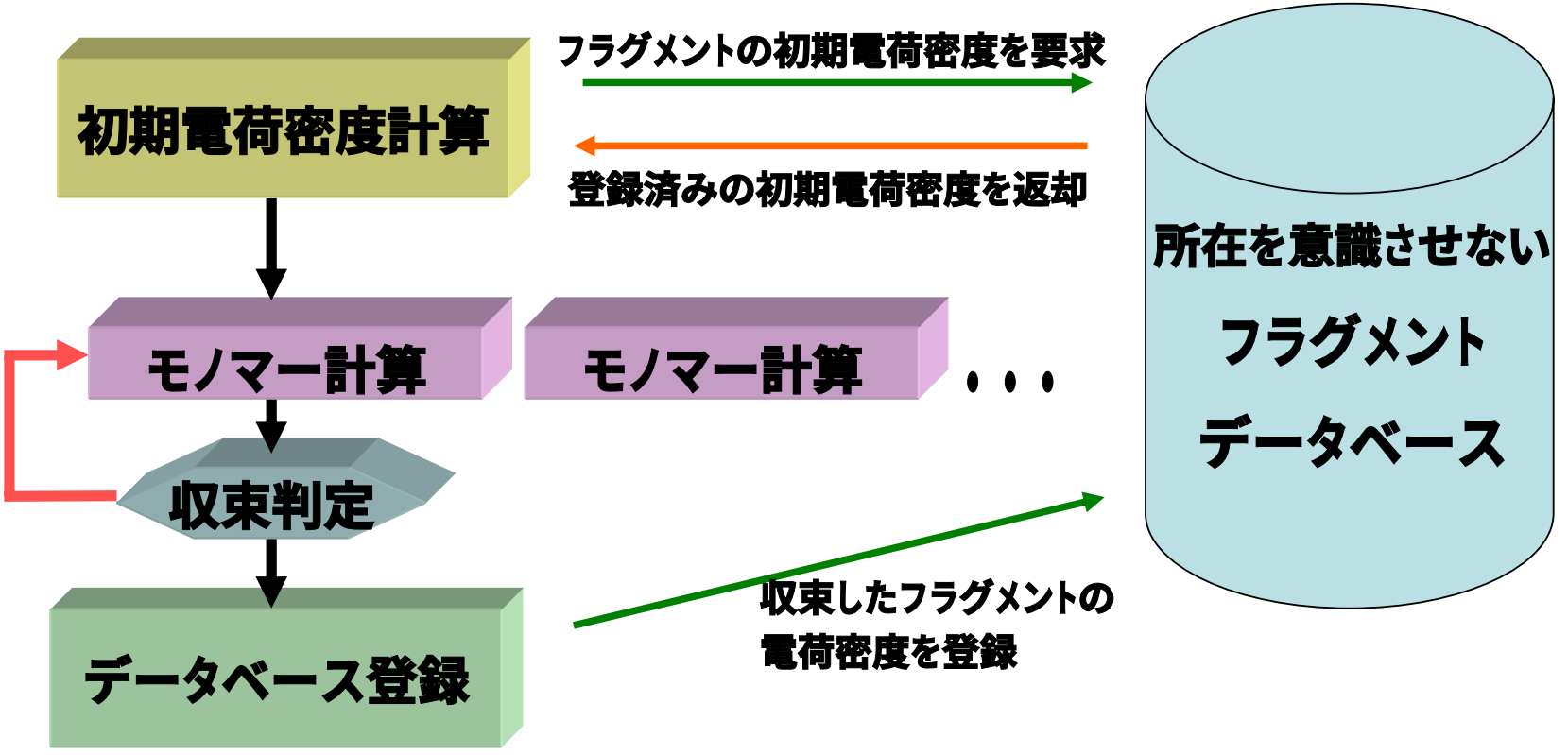
④ deploy each module and translate shell scripts into NAREGI workflow

NAREGIワークフローツールを利用した分散処理



フラグメントデータベースの活用

モノマー計算の結果をデースベースに蓄え、再利用可能とする機能を実装　(分散　データグリッド応用)



グリッド上における大規模計算

平成15年度

SSBOND分子（101原子、5フラグメント）を単一サイトにて計算

平成16年度

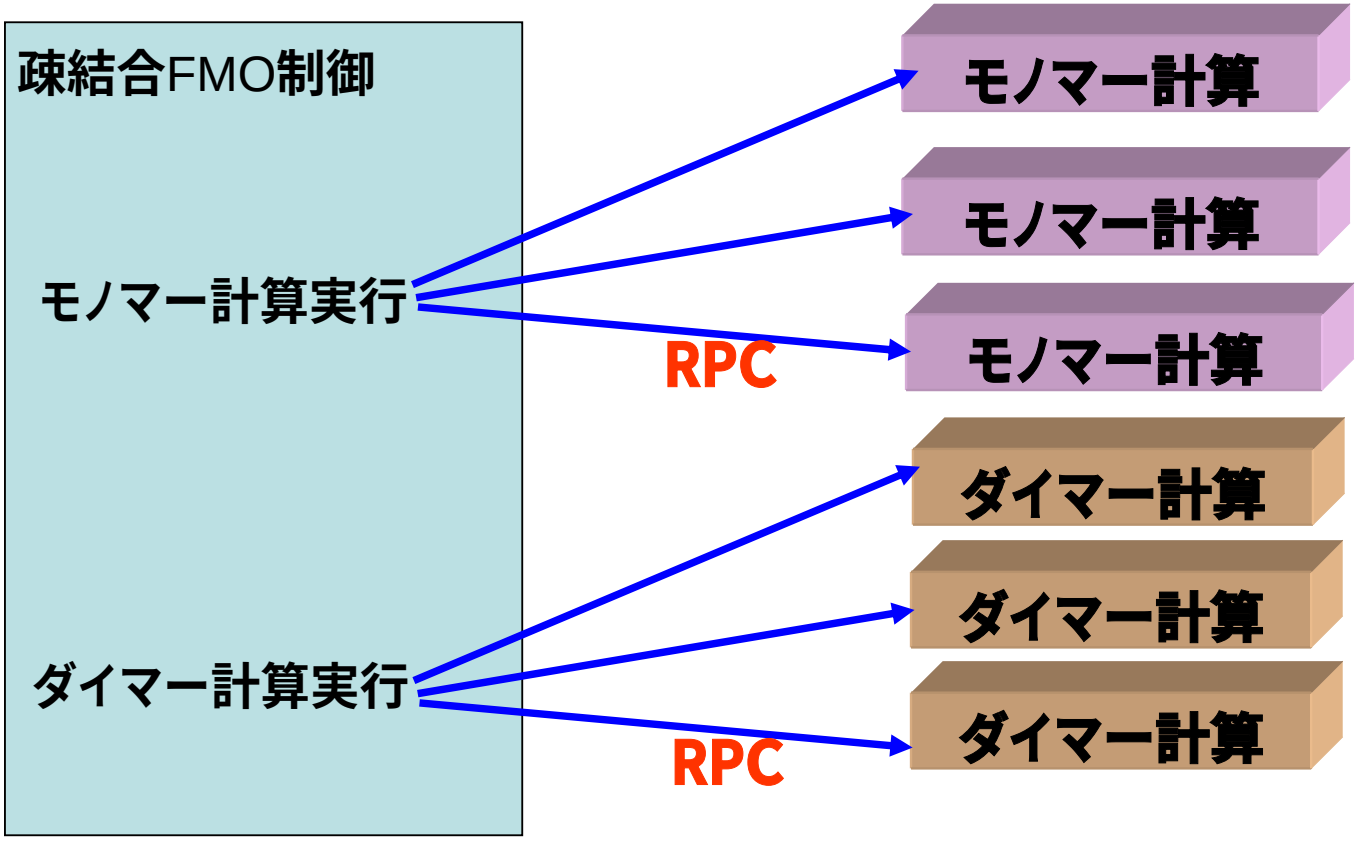
コラーゲン分子（327原子、12フラグメント）を複数サイトを対象に
NAREGI上位層とフラグメントDB機能を用いて計算

**血清アルブミン分子（18972原子、1122フラグメント）を
NII GridとIMS Gridを使用し計算を実施**

脂肪酸を7箇所の窪みに取り込み計算を行った結果、どの窪みに取り込んだ場合でもエネルギーが安定していることを確認

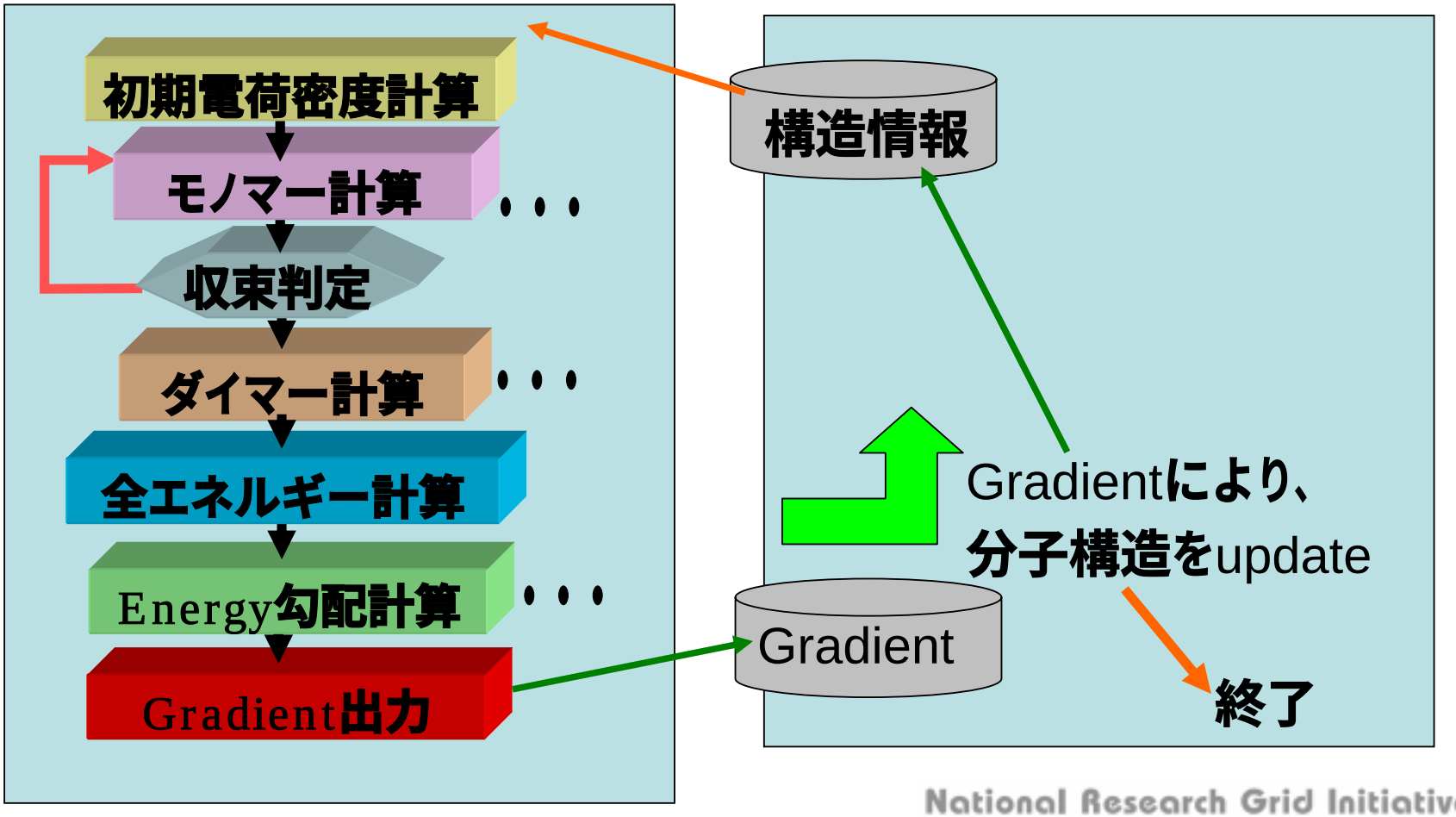
疎結合FMOアルゴリズムとＲＰＣ化

疎結合FMO制御プログラムより、疎結合FMOモジュールをRPCした（AISTグリッド研究センターとの共同研究）



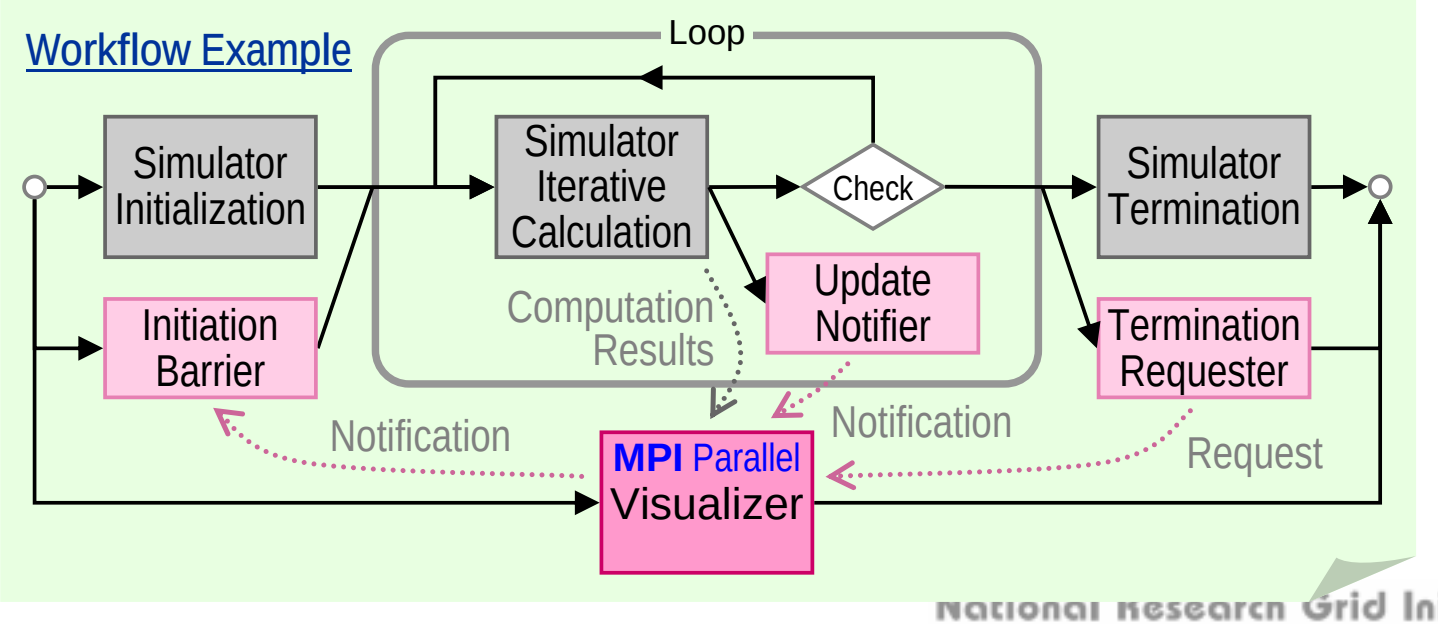
疎結合構造最適化

エネルギー勾配計算を疎結合FMOモジュールとして開発し、
Optimizer　TINKERと連携させた構造最適化計算を実現

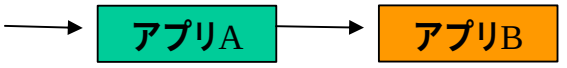
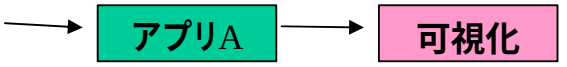
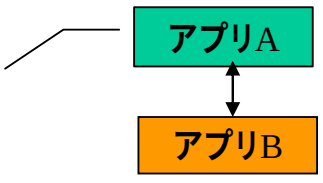
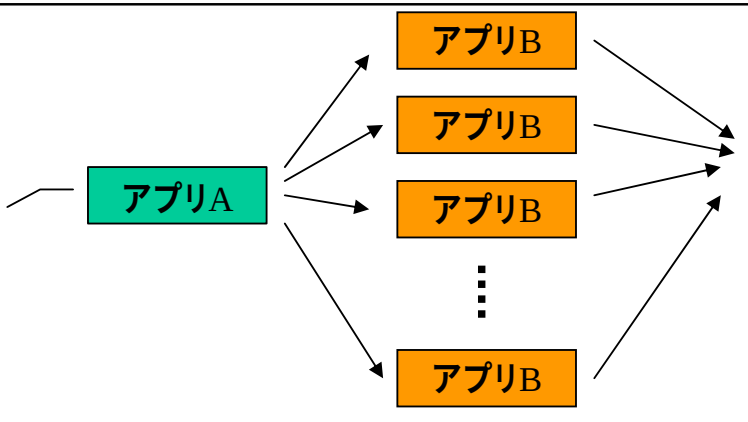
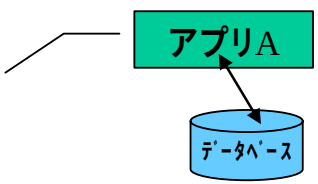


疎結合並列可視化

- Visualizer receives latest results via files from simulator.
- The following components work cooperatively.
  - Initiation Barrier: waits for starting of Visualizer.
  - Update Notifier: notifies Visualizer of new computation results.
  - Termination Requester: requests Visualizer to terminate.



フローパターン 要素(Primitives)

フロー単位	連携･連成フローパターン Primitives
単一･アプリ	
直列･連携型	
＋可視化	
平行･連成型	
分割統治型(1-N-1) Devide-and-Conquer Fork-and-join	
データベース連携型	

Macro フローの構造

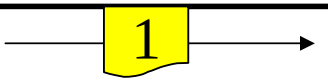
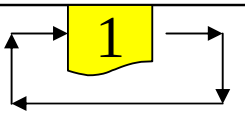
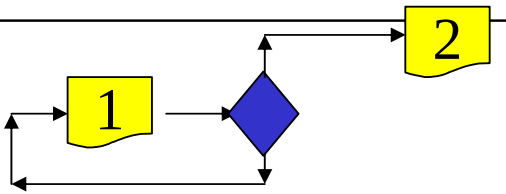
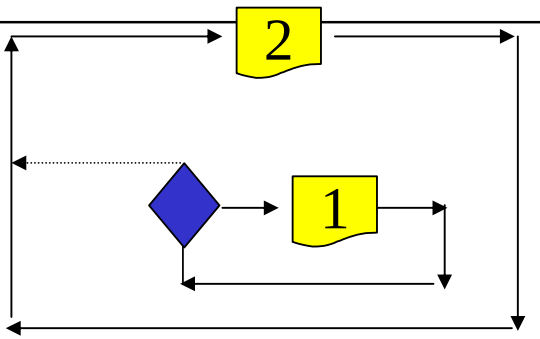
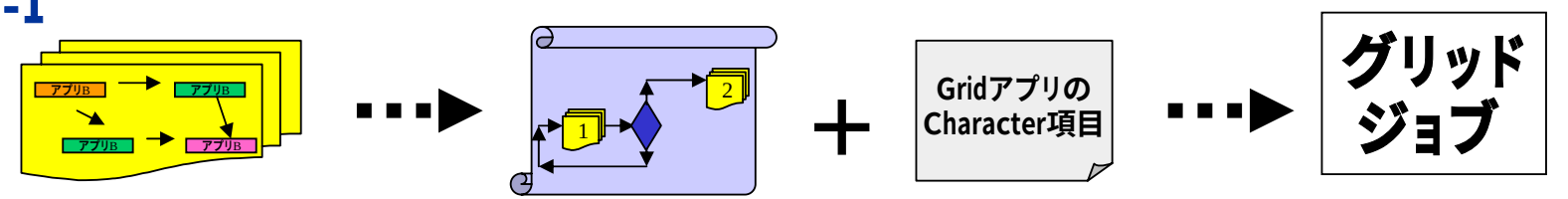
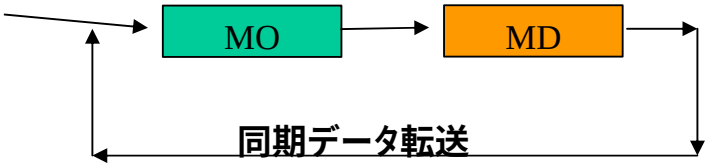
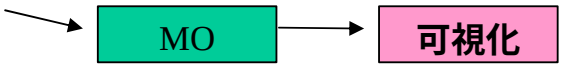
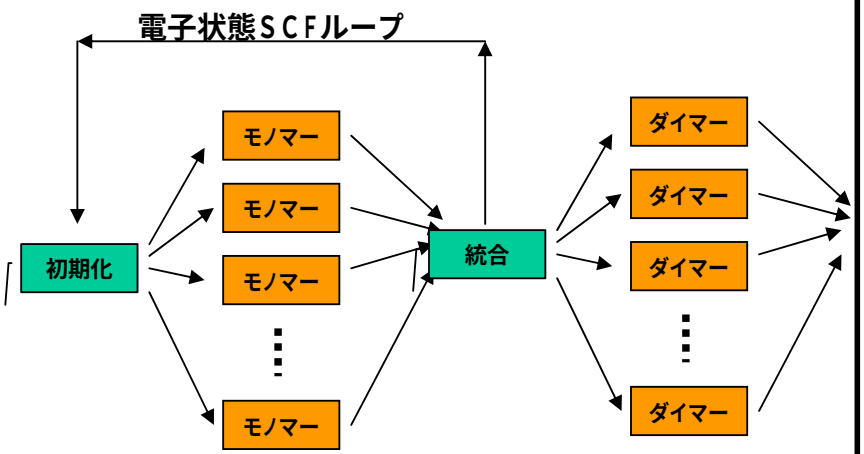
ループ構造/分岐	Macro フローの構造
なし / なし	
1重 / なし	
1重 / あり	
多重	

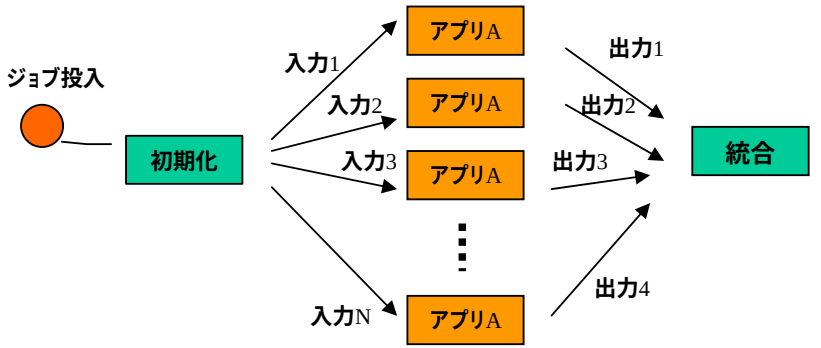
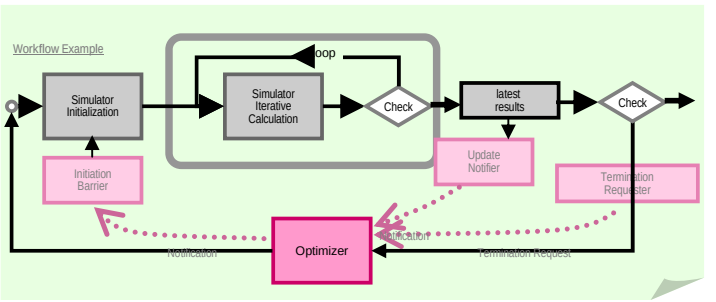
図-1



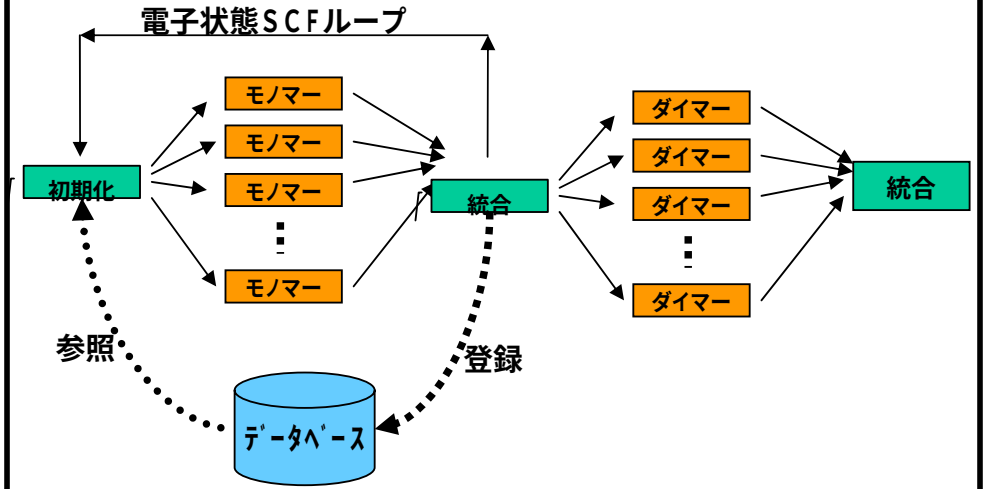
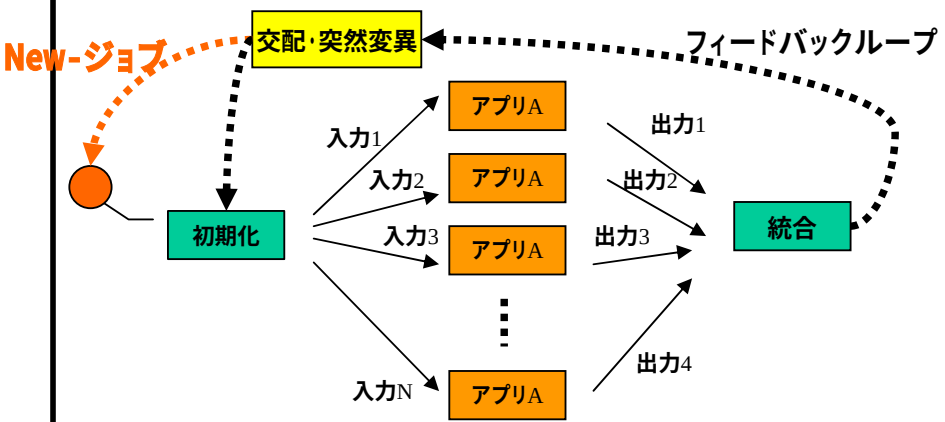
アプリケーションシナリオの例

アプリシナリオ例	グリッドジョブ ジョブ total フローパターン	内容
並列MDジョブ		MPI 並列MD 単一ジョブをIMSのPCクラスタ上で実行する
MO-MD 連携ジョブ		IMSでFMO計算した結果(Force)をNIIへ転送してNIIでMD計算し，更新した座標をIMSへ戻し再度MO計算を行う．(MD部の粒度が小さくサイト間連携ジョブは実は不向き．)
MO＋(同期)可視化		IMSでFMO計算した結果を九大で遠隔可視化する(c.f. 非同期可視化)
疎結合FMO計算		GAMESS-FMOのモノマー，ダイマー，エネルギー勾配計算部を単体で動作可能なモジュールに分割した「疎結合FMO」プログラムをIMSとNIIのＰＣクラスタ上で実行する．(備考) (1)構造最適化を行う際は，(同期型) 2重ループとする方法，および (非同期型) 連成最適化(別紙)とする方法がある．(2)MOの初期データを過去の計算から取得するケースはＤＢとの連携項目を参照．

アプリケーションシナリオの例（cont)

アプリシナリオ例	グリッドジョブ total フローパターン	内容
広義のパラメタサーベイ		Simulated Annealing, レプリカ交換分子動力学, 大域的最適化問題, Combinatorial 計算化学物質探索等. Master-Workerで行う場合はBarrier同期が必要. GridRPCならば非同期.
RISM-FMO連成	後述1	Mediator (Wp6)を介する溶媒-溶質連成計算.
疎結合可視化	後述2	最新のシミュレーションoutputを (非同期で) 可視化サービスに依頼し, 結果を遠隔可視化.
疎結合最適化	後述3 	パラメタを含むシミュレーションoutputから目的関数 (totalエネルギーなど) の解析的パラメタ微分を取得し, パラメタの更新を最適化サービスに依頼. 全体としては非同期のループ構造.

アプリケーションシナリオの例（cont)

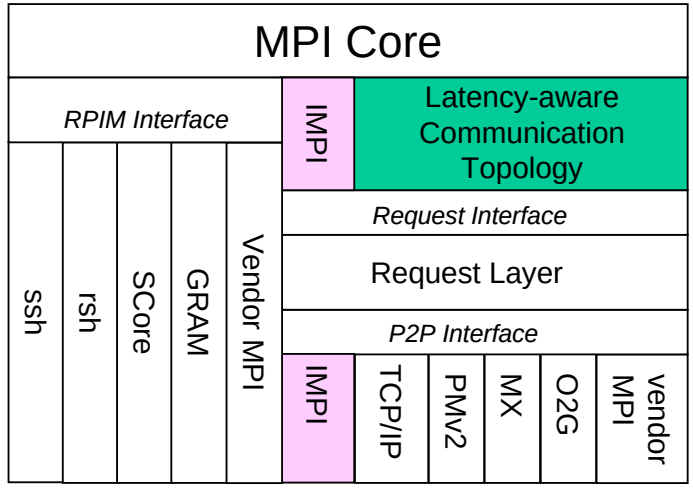
アプリシナリオ例	グリッドジョブ total フローパターン	内容
データベース連携型疎結合FMO		外部にFragment初期電子密度のデータベースがある場合の電子構造計算．疎結合FMO(一般にＭＯプログラム)は，まず当該Fragmentの情報が既にDBに登録されているか否かを判断し，登録済の場合はそれを入力として利用する．　新規のFragment計算ならば電子状態計算が収斂した後で，収束した電子密度をDBに登録する．
自律分散型?		遺伝的アルゴリズムを用いた大域的最適化問題や探索問題，マルチエージェント法に基づく自律的なcell automata , 生態系，ライフサイエンス分野等．　黒の点線のフィードバックならば単なるパラメタサーベイ型＋１重ループ．　赤の点線の場合は，統合データを加工した後，新規にグリッドジョブを投入する．

グリッドアプリケーションのフロー構造と Characterization

フロー単位	ループ構造/分岐	同期 (synchronization)	並列/分散処理手法	演算粒度	遅延許容度	分散効率 (負荷バランス)	要求する実行マシンの環境
単一・アプリ	なし / なし	該当なし	なし	大	許容大	良い	ANY ノード
直列・連携型	1重 / なし	同期	MPI	中	普通	普通	単一サイト・単一マシンノード
＋可視化	1重 / あり	非同期	**GridMPI**	小	許容小	悪い	複数サイト・単一マシンノード
平行・連成型	2重 / なし		**GridRPC**				複数サイト・異機種混在
分割統治型(1-N-1) Devide-and-Conquer Fork-and-join	2重 / あり		Mediator/ MPI or /GridMPI				
データベース連携型	多重	(備考) 他の特徴付・項目としてグリッドサービスの「ライフタイム」がありうる.	単一ファイル転送				(備考) その他, ローカルマシンのノード数, CPU時間, 主記憶容量, ファイル容量など, サイトの{物理,論理,運用} 上限値への適合性という環境要素がある.
			集団通信・集団ファイルcopy				

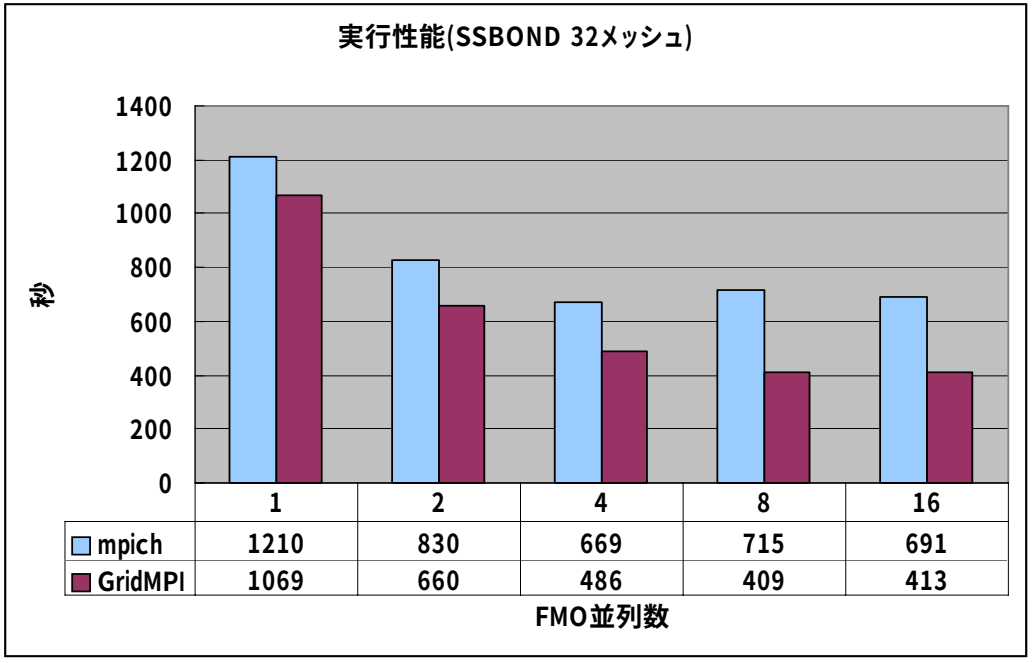
(講義の後半で紹介予定)

GridMPI + Mediator tools



Features

- MPI-1.2 compatible
- IMPI 0.0 full implementation
- YAMPII TCP/IP and Score
- Collective Communications are implemented based on the network latency and bandwidth characteristics



Mediatorによる連成計算方式

(講義の後半で紹介予定)

Mediatorは、アプリケーション間での高度な意味変換機能をサポートするパッケージ群である。
計算手法の異なるアプリケーション間の物理量変換は、グリッド上でMediatorを介して自動的に実行されるため、プログラムの修正を最小限に抑えた連成シミュレーションを構築できる。

1. 連成シミュレーションの構築手順

Mediatorの提供するAPIを用いてレガシーコードをコンポーネント化。グリッド標準の通信フレームワーク上で連成化。

①ソースコードにMediatorのAPIを追加

```
PROGRAM MAIN
CALL MPI_INIT
```

連成シミュレーションの管理
call dex_join (color, ローカルコミュニケータ)

離散点情報の設定
call dex_register_quantum
(位置関係, 離散化手法,離散点数,番号,
座標,要素情報,属性)

ナノアプリケーション
レガシーコード

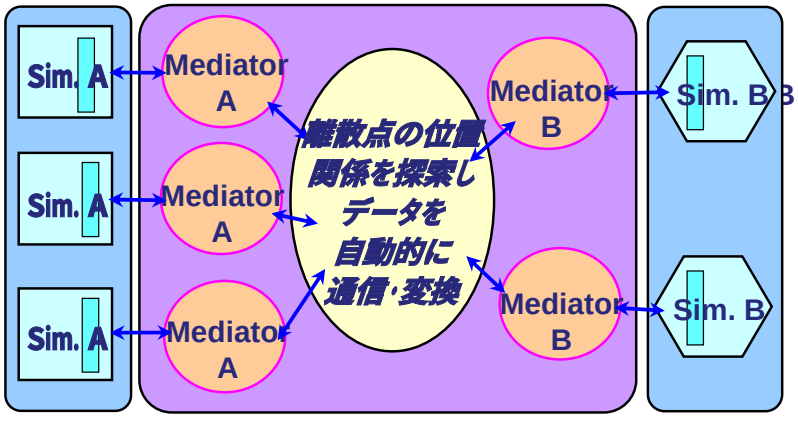
データ送受信
call dex_send(byte,物理量a) 又は
call dex_receive(byte,物理量b, 変換関数)

```
CALL MPI_FINALIZE
END
```

②コンパイル及びMediatorライブラリをリンク

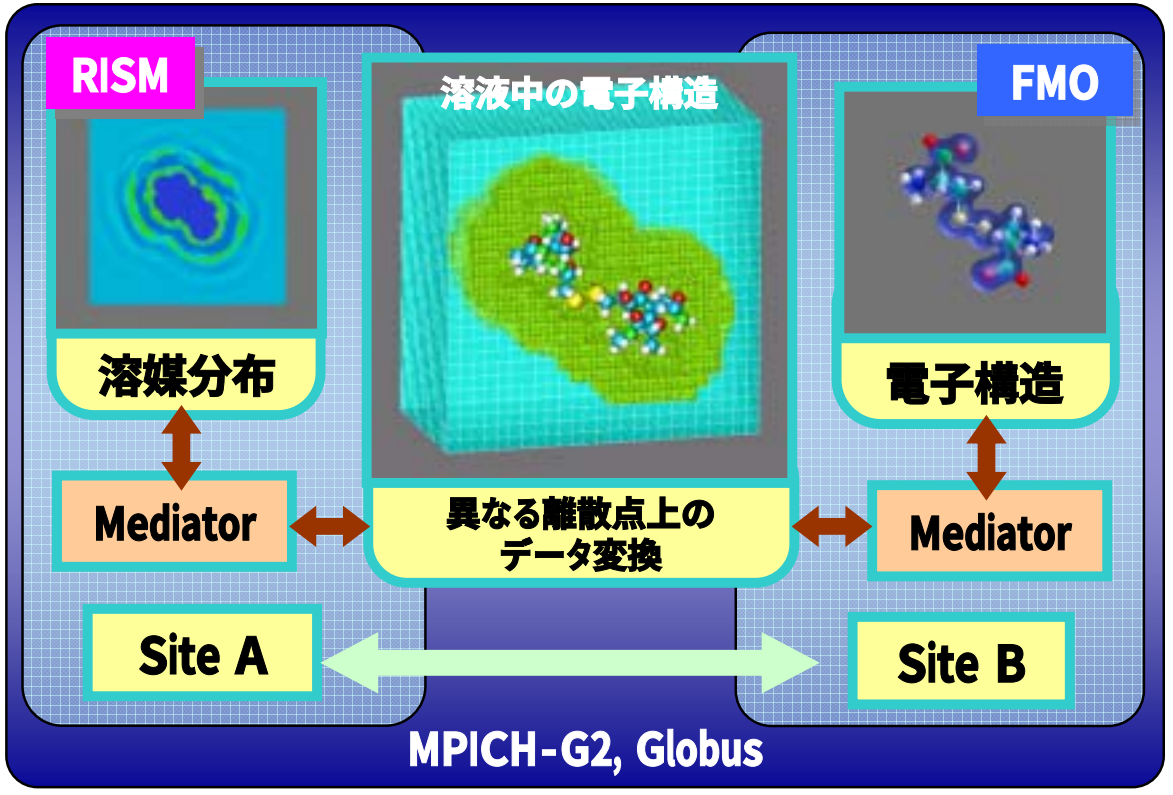
③アプリケーションとMediatorをGridMPIジョブとして実行

2. グリッド連成ミドルウェア Mediatorの特徴機能



- 高度な意味変換機能
  球内相関、矩形内相関、第一近接、最近接
- 離散点位置の高速・並列探索
  粒子法、差分法、有限要素法
- 標準通信ライブラリへの対応
  MPICH、Score、MPICH-G2、GridMPI

GridMPIによるグリッド連成ミドルウェアの開発と応用



RISM Reference Interaction Site Model

無限系に対する溶媒分布解析プログラム

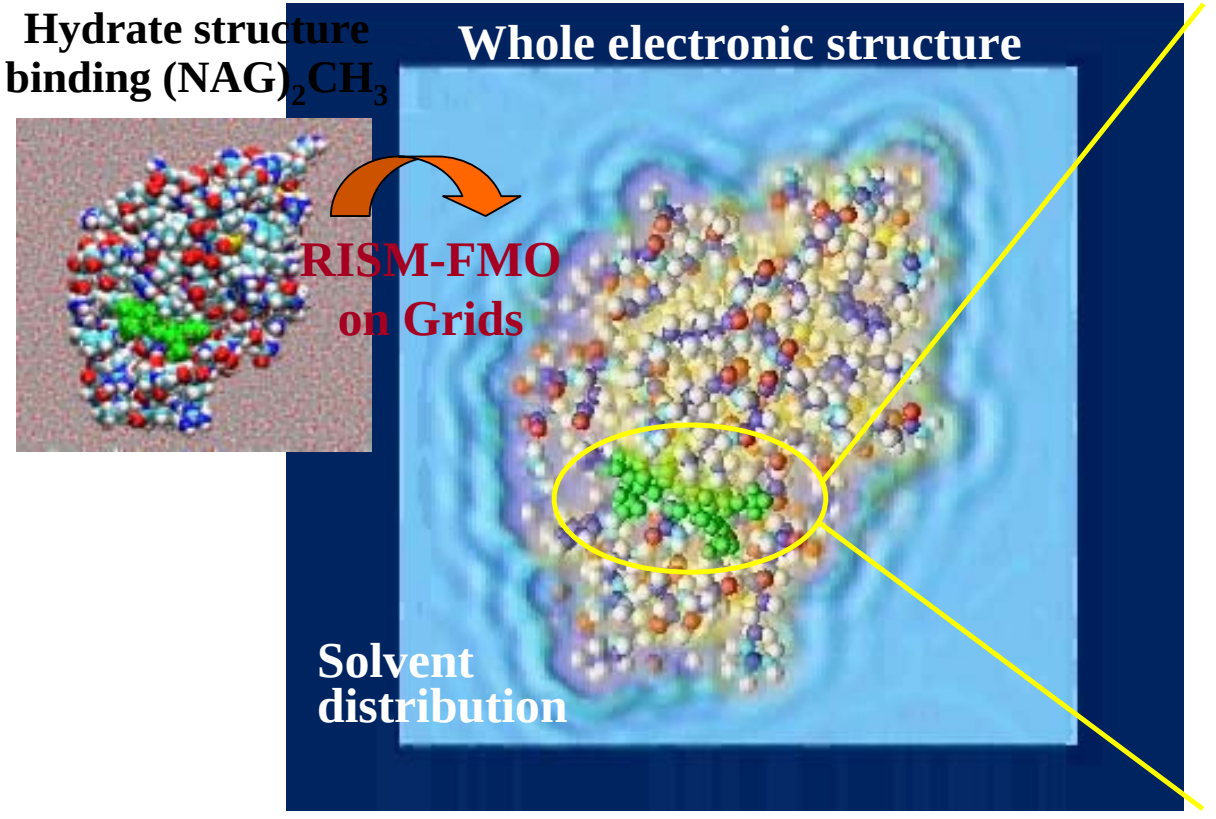
FMO Fragment Molecular Orbital method

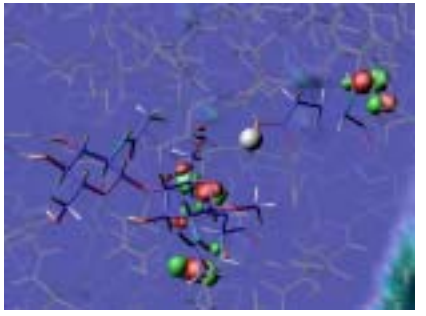
ナノ高分子の電子構造解析プログラム

グリッド連成ミドルウェア(Mediator)を開発し、複数のナノアプリケーション(RISM, FMO)をグリッド環境下で接続し、超巨大水分子を対象に溶液中の電子構造を計算

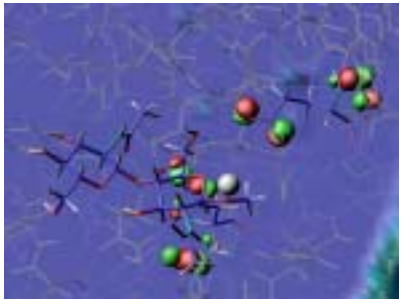
Lysozyme functionality in solvent

Whole electronic structure of Lysozyme in solvent is analyzed by using RISM-FMO and electronic density changes according to the position of proton transfer.





Proton at Glu35



Proton at Glycosidic O

今日のおわりに・・

- サイエンス分野のグリッド利用はまだ始まったばかり，対障害，ＶＯ管理など実運用ではいくつかの課題.
- アプリとミドルウェアそれぞれの役割を明確化
- 疎結合で連携できるモジュール，コンポーネントの集合，メタデータ，メタ通信等の抽象化
- サービスのネットワークとして・・

  インターフェースの標準化（サービスがサービスを呼び出す）

  再利用性，ライフタイム，

  ビジネス分野での利用とも関連しながら発展