第２回先進スーパーコンピューティング環境研究会（ASE 研究会）発表資料

ASE 研究会幹事　片桐孝洋

2008 年 8 月 20 日（水）13 時 30 分から 16 時 15 分まで、東京大学情報基盤センター大会議室にて、第 2 回先進スーパーコンピューティング環境研究会（ASE 研究会）が開催されました。

本号では、独立行政法人　理化学研究所　次世代生命体統合シミュレーション研究推進グループ　小野謙二　博士による招待講演「ペタスケールシミュレーションのソフトウェア基盤」の発表資料を掲載させていただきます。なお、概略につきましては、前号の記事(Vol.10, No.5, 79 頁, 2008.9)をご参照ください。

ASE 研究会では、先進スーパーコンピューティング環境を支援する研究話題の提供を目的に、年数回の活動を計画しております。第３回研究会につきましては、２００９年３月頃の開催を企画しております。

ASE 研究会の開催情報は、メーリングリストで発信をしております。研究会メーリングリストに参加ご希望の方は、ASE 研究会幹事の片桐（katagiri@cc.u-tokyo.ac.jp）までお知らせください。これからもご支援のほどを、よろしくお願い申し上げます。

以上

ペタスケールシミュレーションのソフトウェア基盤

小野謙二
理化学研究所
次世代計算科学研究開発プログラム
生命体基盤ソフトウェア開発・高度化チーム

ASE 研究会　2008

TOP500 JUNE 2008

Rank	Site	Processors	RMax	Processor	System Model
1	DOE/NNSA/LANL	122400	1026000	PowerXCell 8i	BladeCenter QS22 Cluster
2	DOE/NNSA/LLNL	212992	478200	PowerPC 440	BlueGene/L
3	Argonne National Laboratory	163840	450300	PowerPC 450	BlueGene/P
4	Texas Advanced Computing Center/Univ. of Texas	62976	326000	AMD x86_64 OpteronQuad	Sun Blade x6420
5	DOE/Oak Ridge National Laboratory	30976	205000	AMD x86_64 OpteronQuad	Cray XT4 QuadCore
6	Forschungszentrum Juelich (FZJ)	65536	180000	PowerPC 450	BlueGene/P
7	New Mexico Computing Applications Center (NMCAC)	14336	133200	Intel EM64T Xeon 53xx (Clovertown)	SGI Altix ICE 8200
8	Computational Research Laboratories, TATA SONS	14384	132800	Intel EM64T Xeon 53xx (Clovertown)	Cluster Platform 3000 BL460c
9	IDRIS	40960	112500	PowerPC 450	BlueGene/P
10	Total Exploration Production	10240	106100	Intel EM64T Xeon E54xx (Harpertown)	SGI Altix ICE 8200
16	The University of Tokyo	12288	82984	AMD x86_64 OpteronQuad	Hitachi Cluster
20	University of Tsukuba	10000	76460	AMD x86_64 OpteronQuad	Appro XtremeServers
24	Tokyo Institute of Technology	12344	67700	AMD x86_64 OpteronDual	Fire x4600 Cluster
34	Kyoto University	6656	50510	AMD x86_64 OpteronQuad	Fujitsu Cluster
49	The Earth Simulator Center	5120	35860	NEC	SX6

Peta-scale/2008, Exa-scale/2018（SciDAC2008）

HPCとアプリの動向

- MPP
- マルチコア
- アクセラレータ
- ヘテロ環境
  - 開発環境
- Petaプロジェクトの分野
  - ライフサイエンス，宇宙・航空，天文，ナノマテリアル
  - 防災，地球物理，ものつくり，・・・
- データ構造
  - 直交系
  - 非構造系
  - 粒子系
  - AMR
- 規模
  - TB~PB
- シミュレータ
  - 大規模並列計算
  - 連成解析　マルチフィジックス，マルチスケール

ペタスケールシミュレーションの課題

- プリプロセス
  - モデル作成
  - 境界条件設定
- シミュレーション
  - MPP，マルチコア，Accelerator
- ポストプロセス
  - 分散並列可視化環境

大規模
分散並列

集約的な開発環境の必要性
世界規模の競争
有限のリソース有効利用
コラボレーションの促進

Libraries, Frameworks and Middlewares

- PETSc- The Portable Extensible Toolkit for Scientific Computation
- KeLP- Kernel Lattice Parallelism
- POOMA- Parallel Object-Oriented Methods and Applications
- SAMRAI- Structured Adaptive Mesh Refinement Application Infrastructure
- Overture - CFD Application Framework
- CACTUS - Framework on GRID Environment
- HPC(HEC)-MW - Middleware for coupling...
- MpCCI - Generalized coupler

Capability

- For developer
  - Rapid prototyping by reuse of library collections
  - Easy extension from serial to parallel code
  - Source portability and management
  - Various coding style; OO or procedural
  - Management of coupling problem
- For end-users
  - Unified UI, concept
    - Easy access for various solvers on SPHERE
  - Controllable coupling simulation by scripting

Point of View

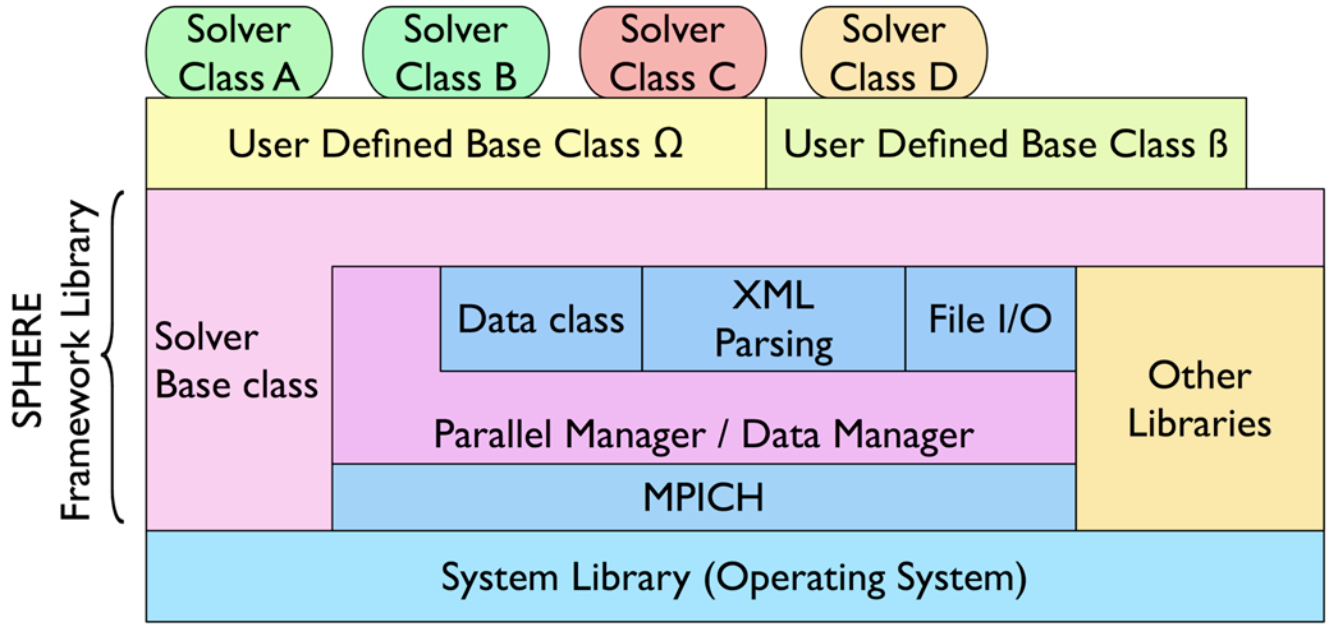
- Productivity
  - Accelerate efficiency of program development
  - Reduce cost of management and maintenance for simulators
- Usability
  - Provide unified User Interface for end-users
  - Higher level description to describe coupling and parallel code
- High performance
  - Retain balance among productivity, usability and performance

Framework SPHERE

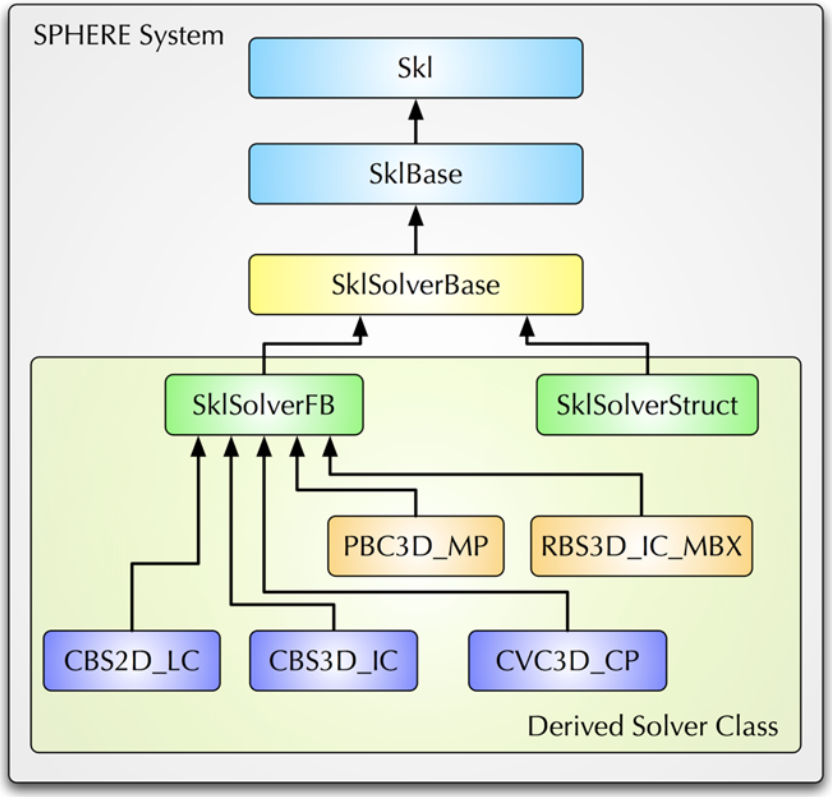
- Functionality
  - Provide mechanism of efficient code development
    - Management of data array
    - Skeleton of program structure
    - Reuse of both methods in libraries and program structure
    - Support for parallelization
  - Provide usable run-time environment
    - Select solvers that are registered on SPHERE
    - Parallel execution
    - Coupling control between solvers
- Multi...
  - Multi-platform (Linux, Mac, Win, NEC SX, SGI Altix)
  - Multi-language - Interoperability (C, C++, f77/f90...)

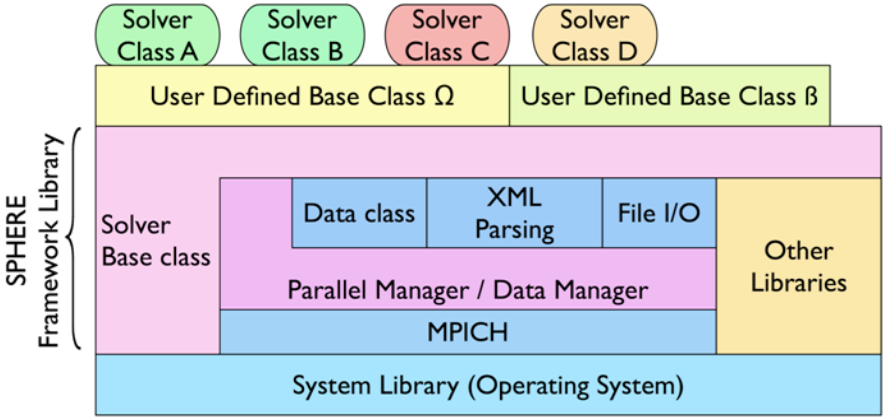
Functionality for solver development

- Libraries (MPI, File I/O, XML parser, ...)
- Support of parallelism based on domain decomposition
- Flow control for coupled solvers
- Skeleton of unsteady physical simulation code
- Difference programming by inheritance



Inheritance and Base class





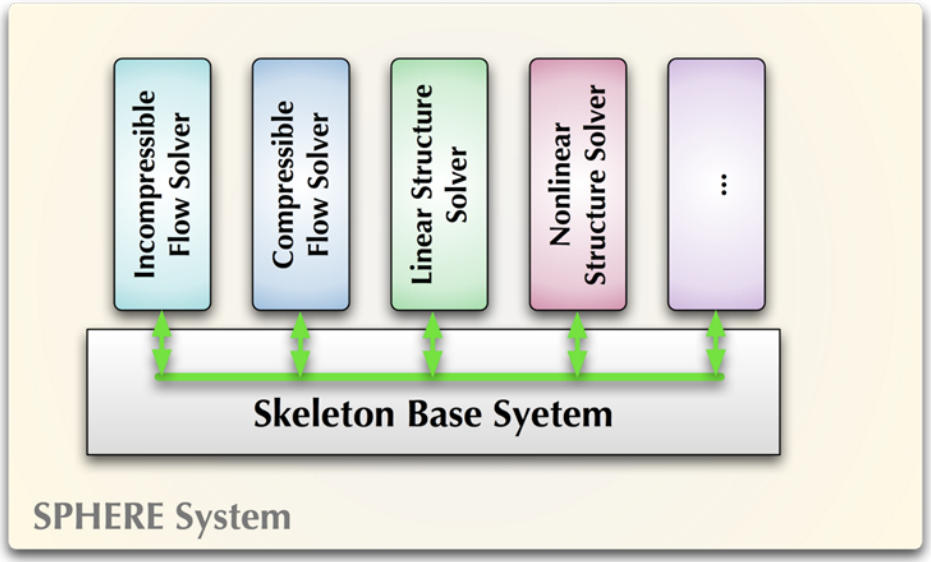
- Skeleton of unsteady physical simulation code
- Difference programming by inheritance

Run-time environment

- Select solvers that are registered on SPHERE
- Parallel execution
- Coupling control between solvers

```
XML file :
<SphereConfigSolverType="Incompressible_flow_solver" >
```

```
Command line :
    $ sphere input.xml
```



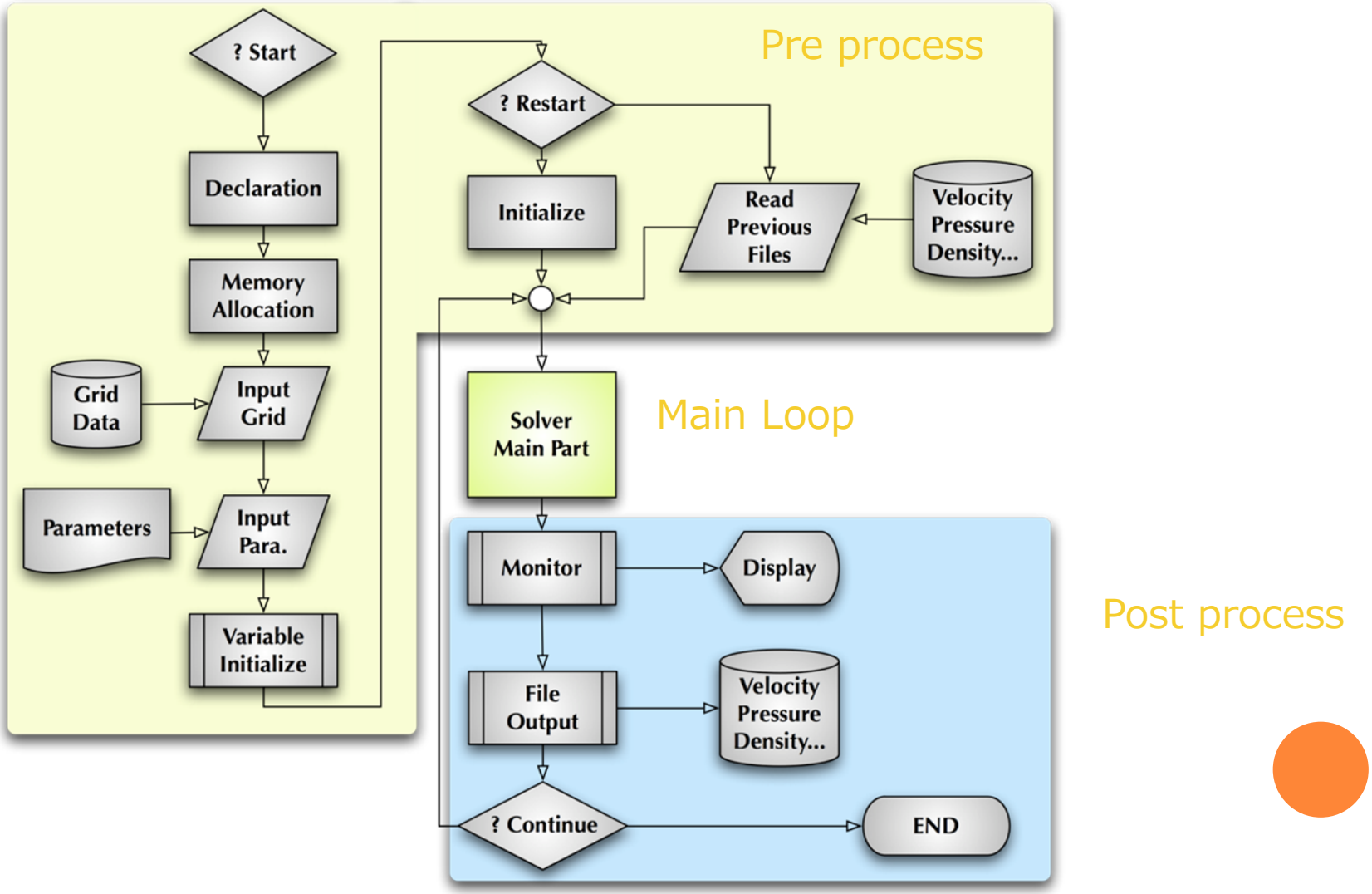
Parameter description by XML

- Management of parameter and run-time information libxml2 library (open source)

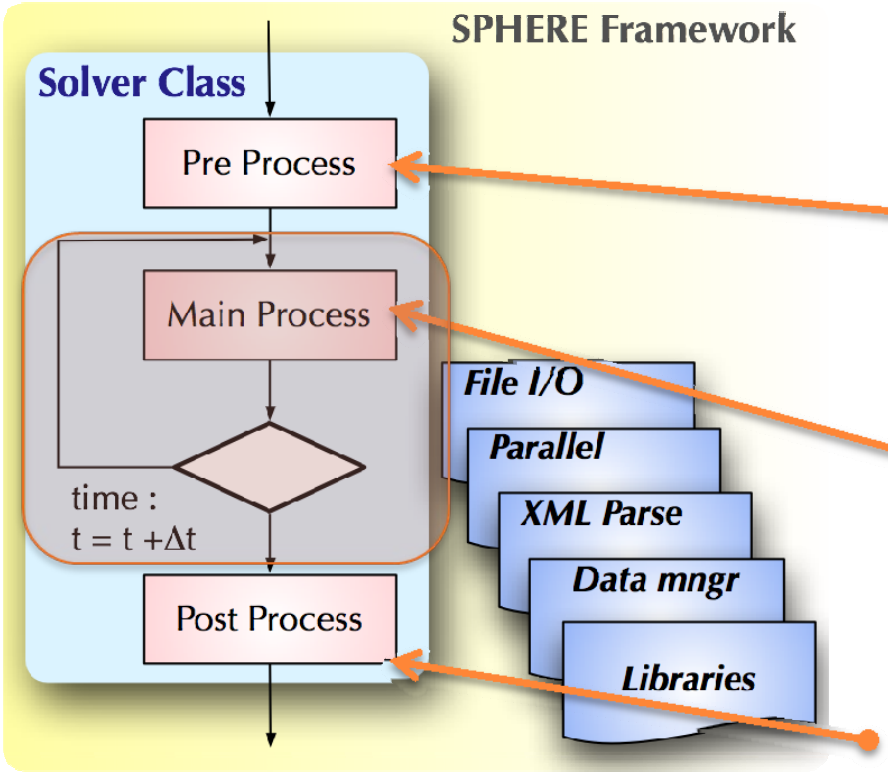
```
<?xml version="1.0"?>
<DomainInfo>
<VoxelOrigin ox="7.72e-01" oy="8.75e-01" oz="8.67e-02" />
<VoxelSize   ix="292" jx="262" kx="53" />
<VoxelPitchdx="1.225e-4" dy="1.225e-4" dz="1.225e-4" />
</DomainInfo>
```

- Enable to working with
  - database
  - other collaborative applications

Class abstraction of unsteady physical solvers

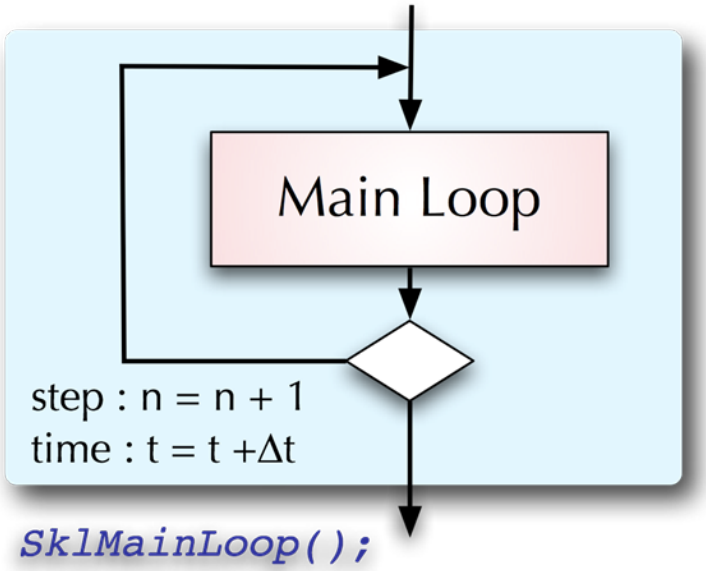


Abstracted Methods




```
boolSklSolverBase::SklSolverEx
ec()
{
intinit_ret =
SklSolverInitialize();

   ...

intloop_ret =
SklMainLoop();

   ...

SklSolverPost();

   return true;
}
```

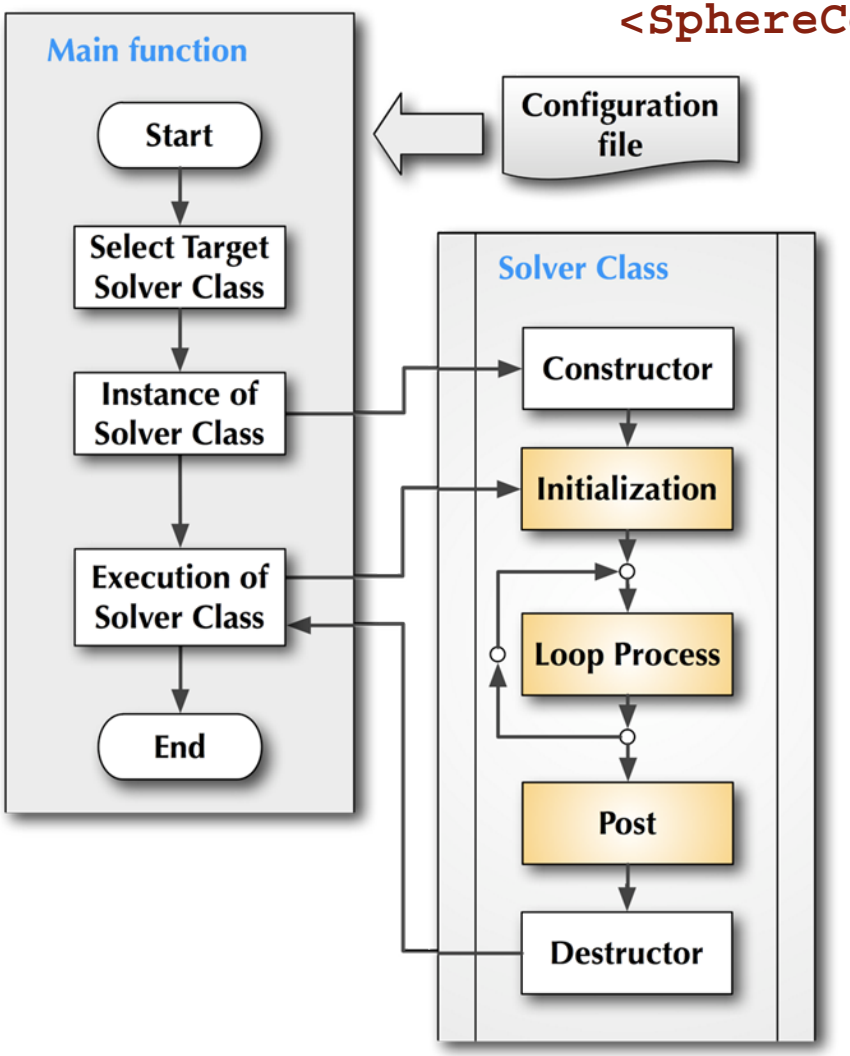
Methods that developer should describe




```
intSklSolverBase::SklMainLoop()
{
int ret = 1;
  register unsigned inti;

for(i=1; i<=m_maxStep; i++){
m_currentStep = i;

intloop_ret = SklSolverLoop(i);

    if( loop_ret == 0 ) break;
  }
  return ret;
}
```

Invocation of SPHERE and solver class object

```
<?xml version="1.0"?>
<SphereConfigSolverType="CBS3D_IC" >
```

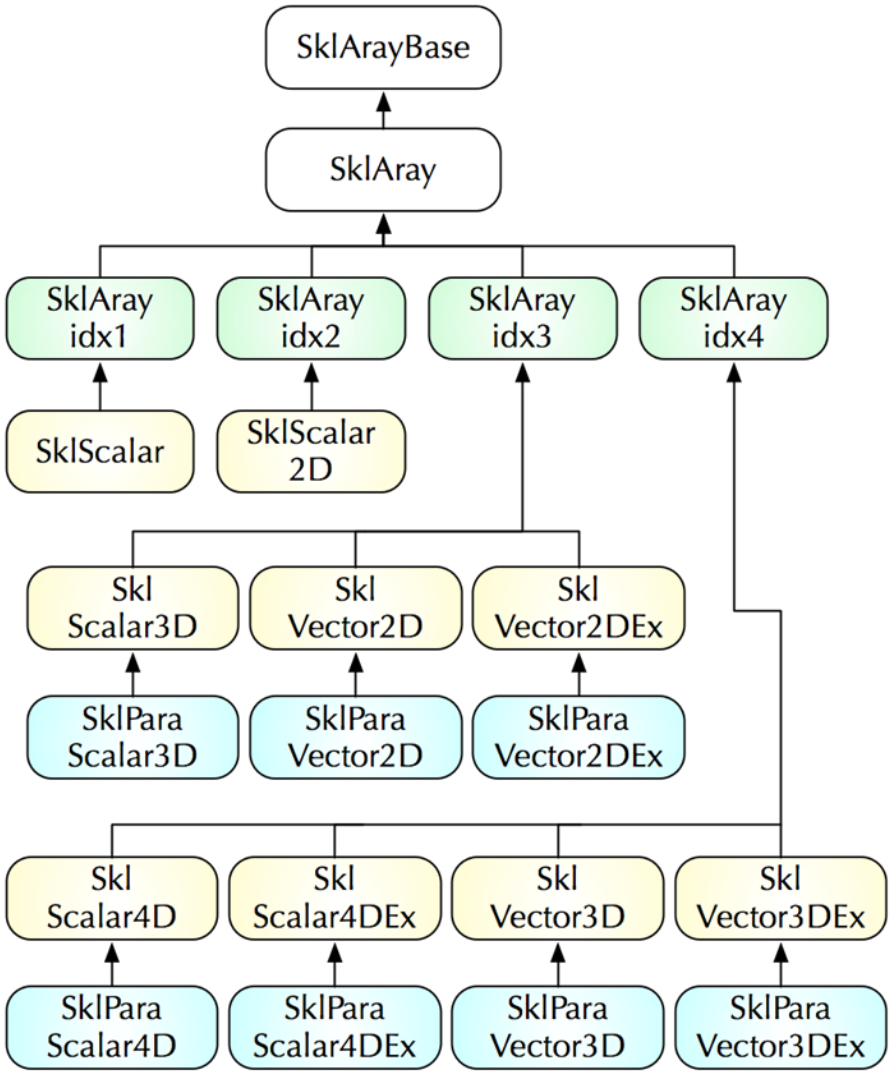


- Select Solver class to invoke
- Instantiation of solver class
- Migration of execution from SPHERE to solver class, and return SPHERE again

Data class

- Management of Array - create, allocation, delete...
- Corresponds to domain decomposition of parallelization
- Communication in between each domain
- Send/receive of data in case of solver coupling
- Based on concept of data class (Ohta, 1999)
- Extension for flexibility
  - Data class + Manager class

Hierarchy and functions of data class



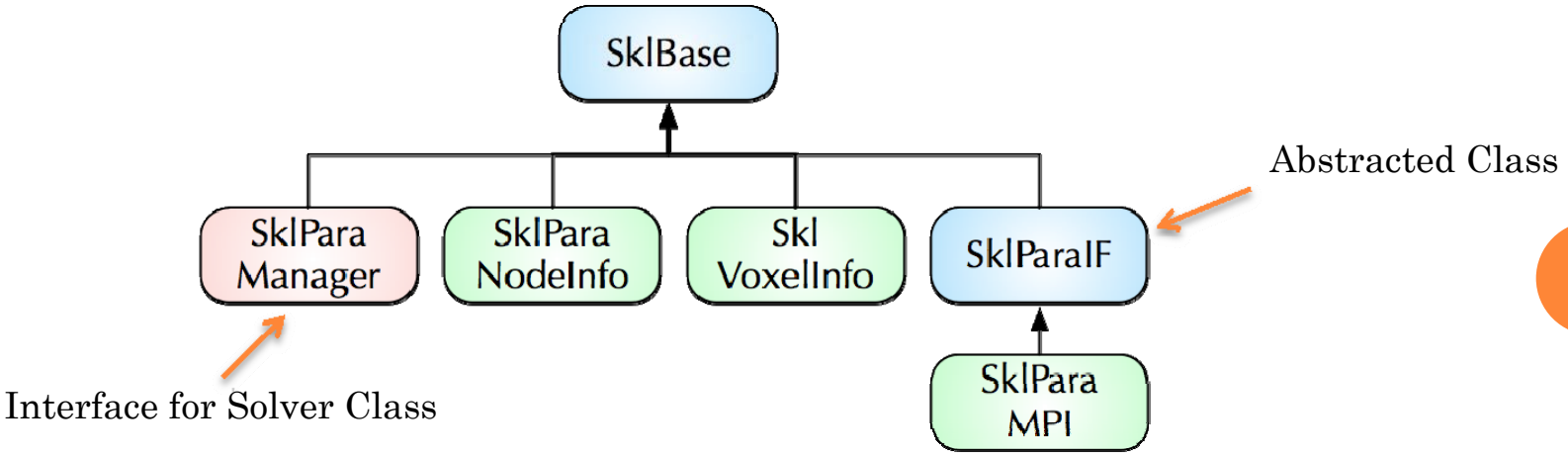
- Management and Operation for Multi-dimensional Array
  - N-dimensional Indexing
  - Vector and Scalar Arrays
- Parallelization
  - Treatment of Guide Cell
  - Communication in between Nodes
- Current status
  - Cartesian, Multibox
- Developing...
  - Octree, BCM, UNS

Large-scale parallel computation

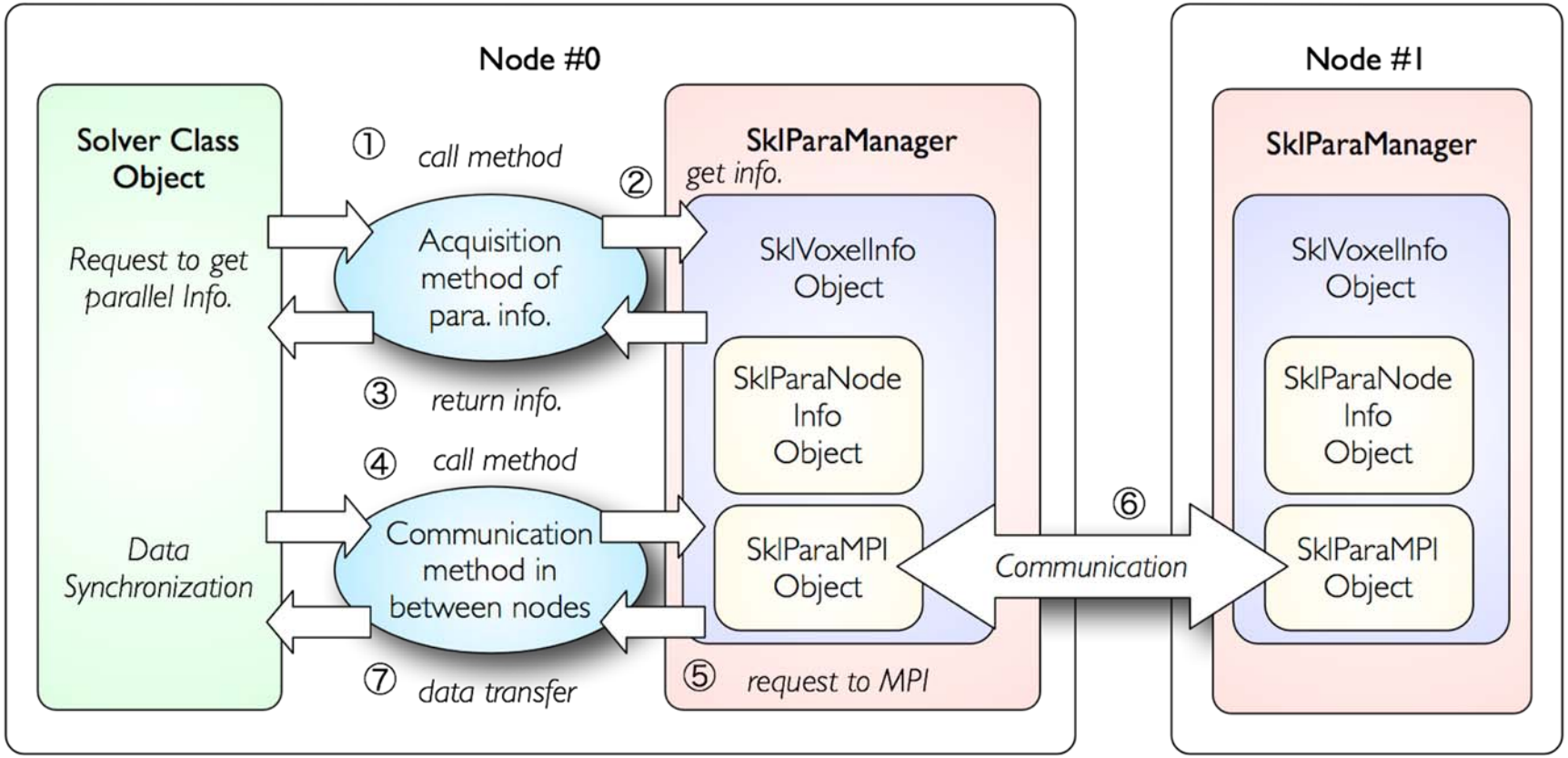
- Parallel procedure build in SPHERE
  - Based on domain decomposition
  - Takes charge of part of synchronization among domains
  - Only specify instruction to synchronize
- Data structure
  - Currently, uniformly spaced Cartesian
  - Non-uniformly Cartesian, Octree, UNS, Multibox, Multilevel, BCM,…

Functions of parallel manager class

- Manage data class and parallel environment
  - Specify serial/parallel library
  - Initialize and finalize methods for parallel environment
  - Obtain parallel information such as node ID, # of ID, neighbor node ID...
  - Communication between nodes
  - Instantiation of parallel data class and registration

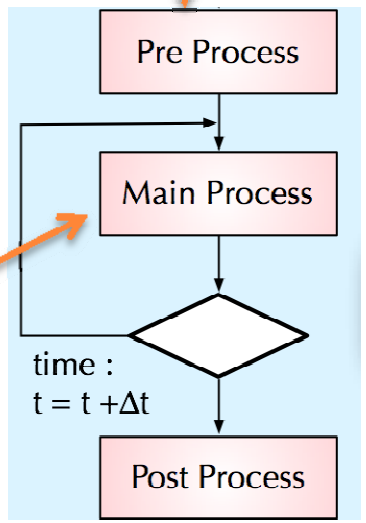


Behavior of solver and parallel manager class



Pseudo code 1

```
intSklSolverInitialize(){
SklCfgGetVoxelSize(wi, wj, wk);
GetCfgVoxelDivisionMethod(nx, ny, nz);
ParaMngr.SklVoxelInit(wi,wj,wk,nx,ny,nz);
VoxelInitilize();
AllocateMatrix();
InitilizeMatrix();
  return 1;
}
intSklSolverLoop(const unsigned step){
hbmtf_jacobi_(nn, p, ...);
  return 1;
}
```



Pseudo code 2

```
subroutine hbmtf_jacobi(nn, p, ...)
  integer, dimension(3) :: sta, end
  real, dimension()     :: p
  real(4)               :: wgosa, gosa
  integer               :: i,j,k,loop, nn, ierr

  do loop=1,nn
gosa= 0.0
    do i,j,k=sta(),end()
p(i,j,k)=....
enddo
call SklCommBndCell(p, 1, ierr)
    call SklAllreduce(gosa, wgosa, ...)
    call BoundaryCondition()
enddo
  return
end subroutine hbmtf_jacobi
```

Benchmark code

- Poisson Solver
- Jacobi Relaxation Method
- Load/Store : Arithmetic = 1 : 1
- { C | f77 | f90 }, { Static | Dynamic }
- MPI Parallelism based on Domain Decomposition

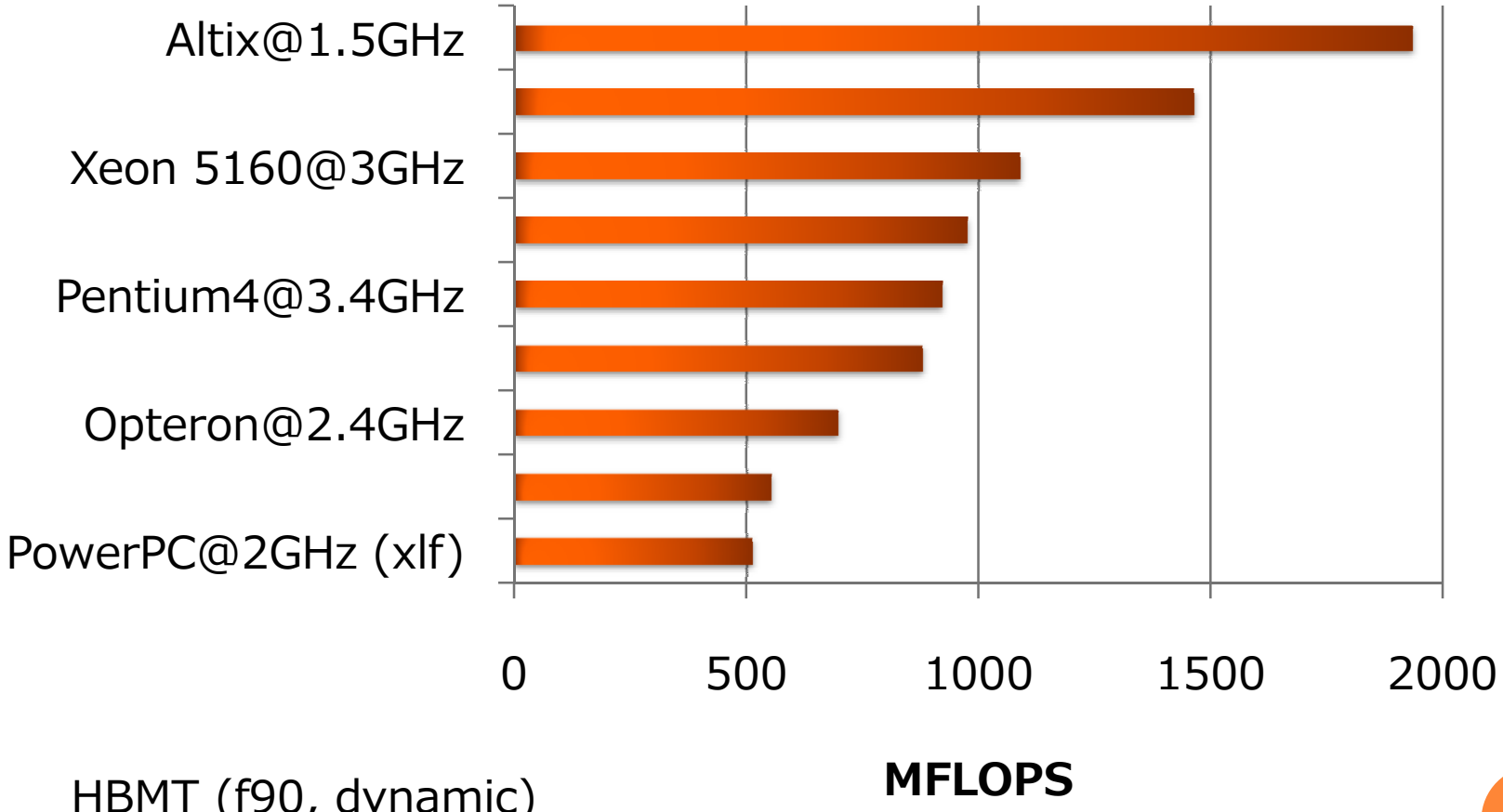
Language and memory allocation

Code	Allocation	MFLOPS	Array size
org_f77	static	1,021	256x128x128
org_f90	dynamic	907	256x128x128
org_C	static	1,144	128x128x256
org_C	dynamic	399	128x128x256
sph_f90	dynamic	921	256x128x128
sph_C++	dynamic	525	256x128x128
sph_C++	dynamic	537	128x128x256

Intel Compiler 9.0

On Pentium4 3.4 GHz

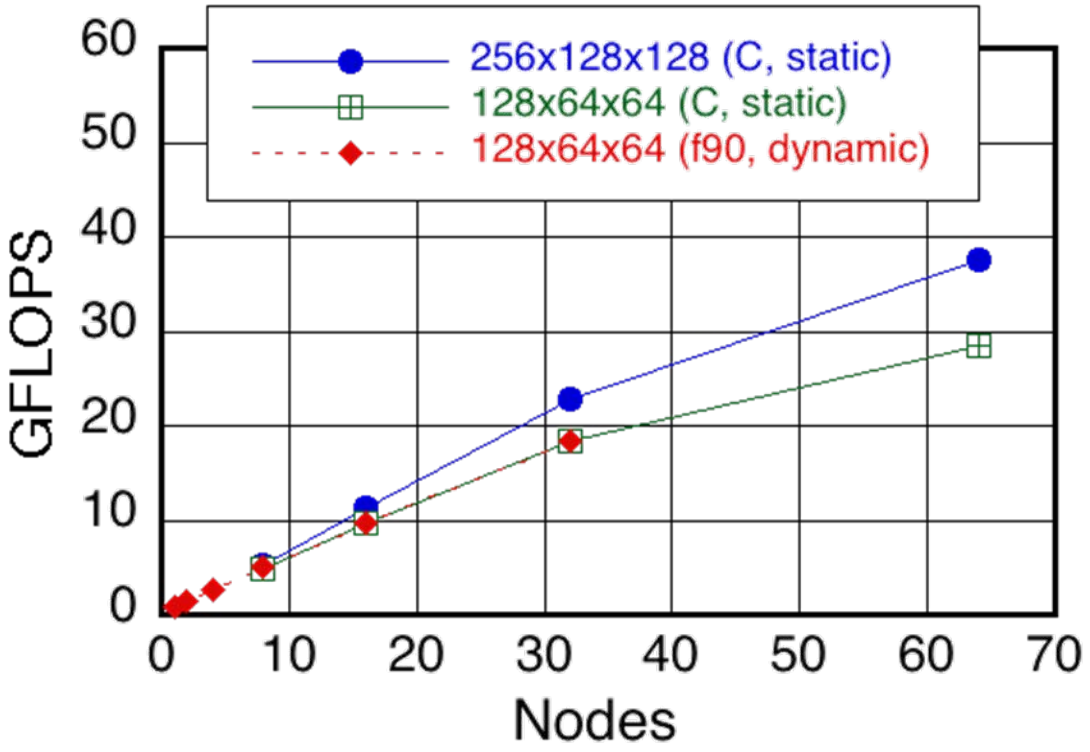
Serial performance on several platforms



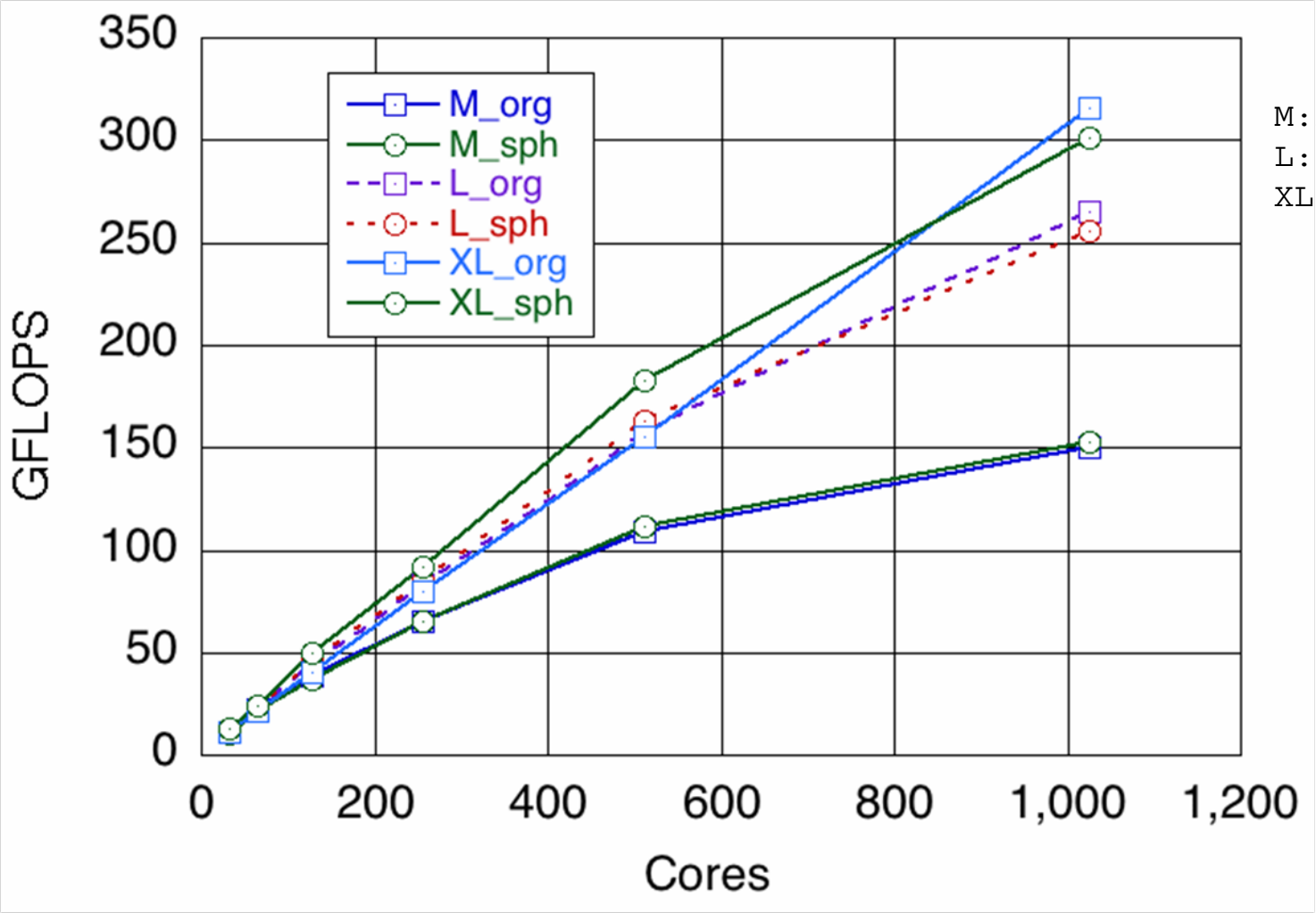
HBMT (f90, dynamic)
ifort 9.1
icc 9.1

Parallel performance of original code on RSCC

	RSCC
CPU	Xeon 3.06GHz 512kB Dual
Chipset	Prestonia
Memory	1GB/CPU
OS	RH8 +Score5.6.1
Compiler	Fujitsu LPL -Kfast
Interconnect	Infini-Band (8Gbps)



Comparison of parallel performance on RSCC

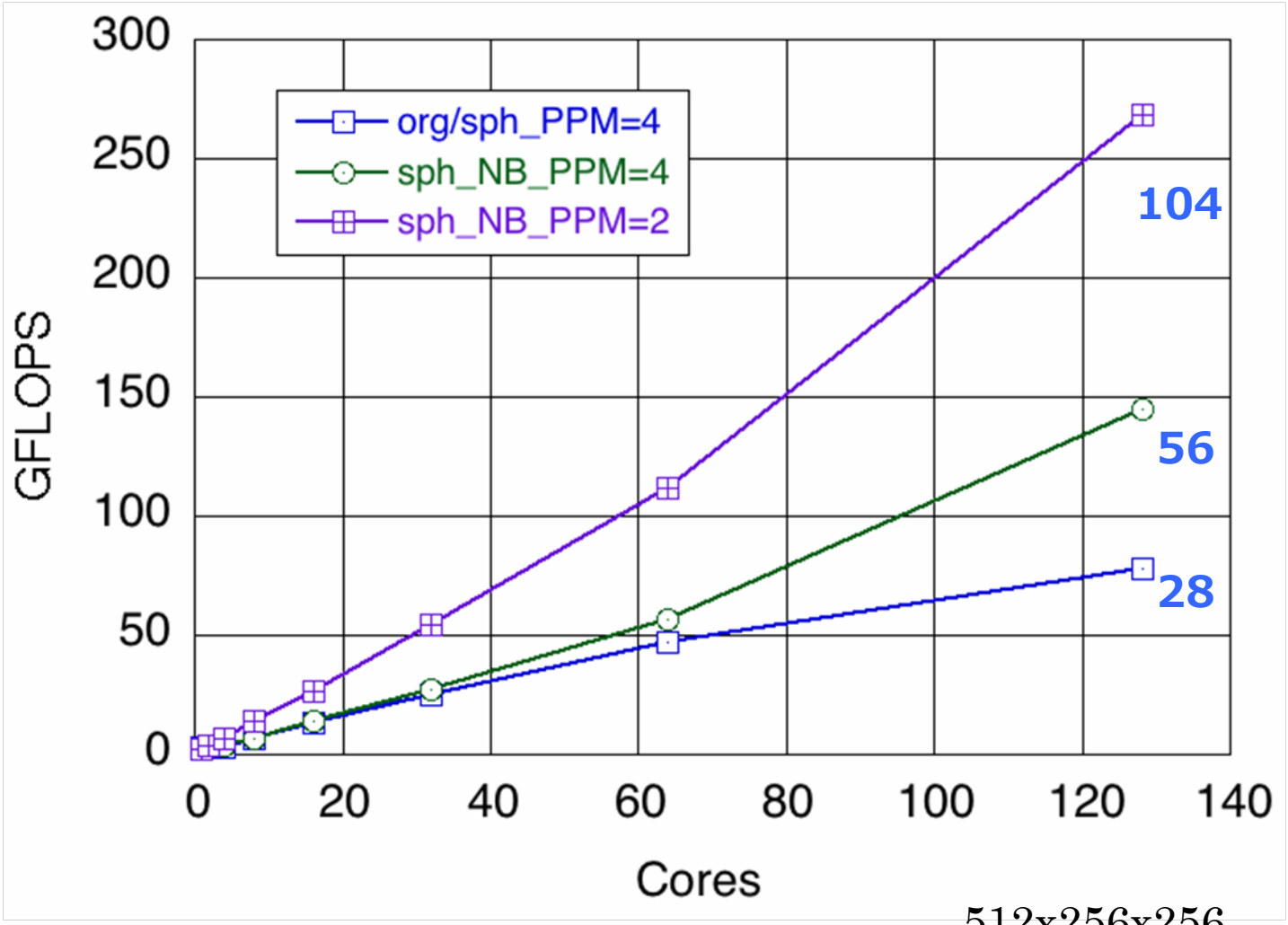


M:256x128x128
L:512x256x256
XL:1024x512x512

Specification of machines

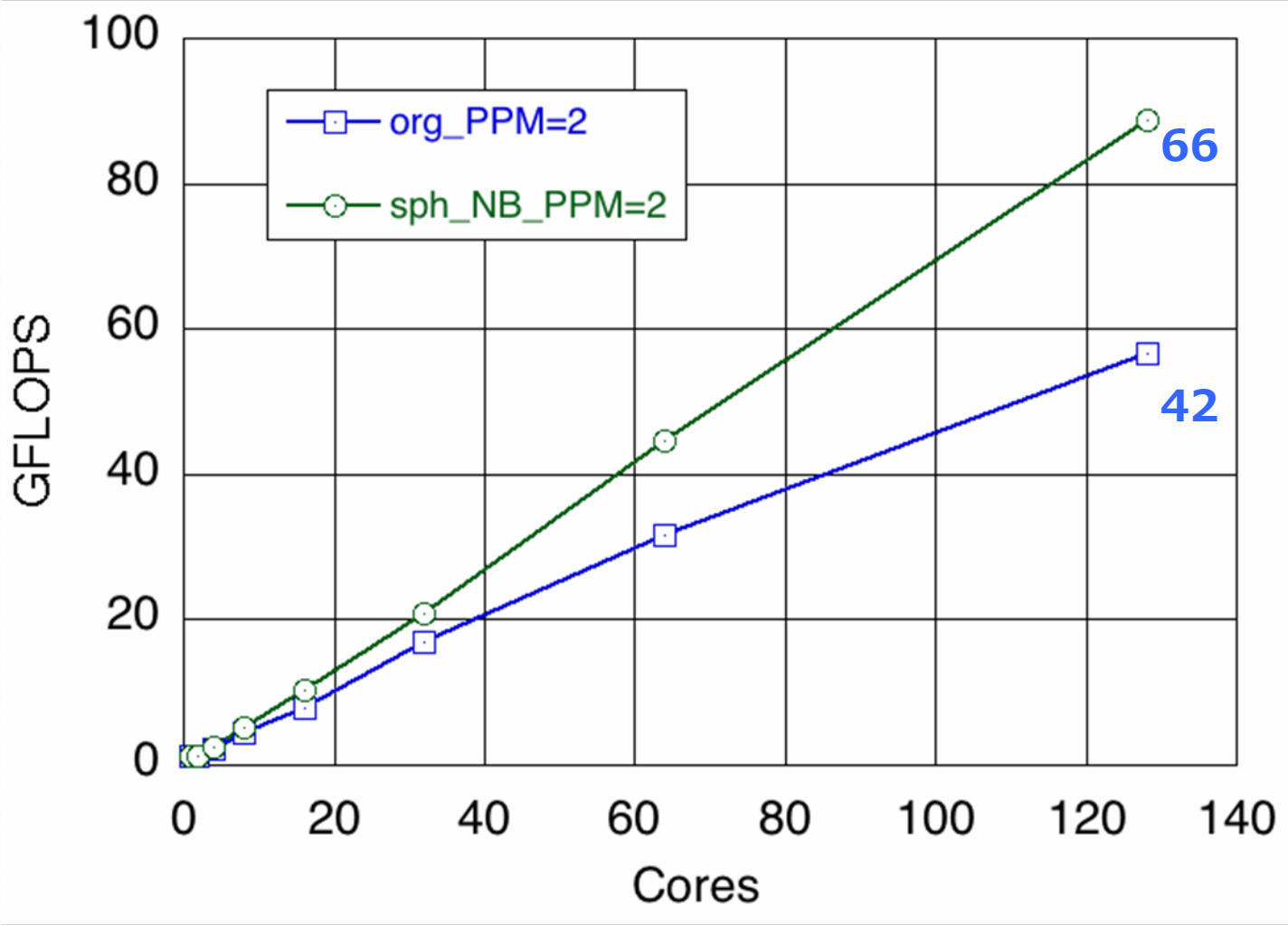
	Altix 4700	Xeon Cluster
CPU	Itanium2 Montecito	Xeon 5160
Frequency	1.6 GHz	3.0 GHz
Core / CPU per node	2 / 1	2 / 2
Node	64	8
Cache / core	9MB (3^{rd})	2GB (2^{nd})
OS	SGI Pro pack5	SGI Pro pack5
Compiler	Intel 9.1 –O3	Intel 9.1 –O3 -axT
Interconnect	CC NUMA	Infiniband
MPI library	MPT1.14	Voltaire MPI

Non-blocking performance on Altix 4700



512x256x256

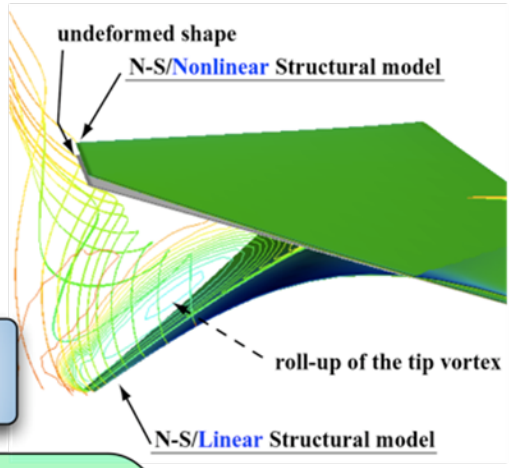
Non-blocking performance on Xeon 5160 cluster



反復解法クラス

- 大型疎行列の連立一次方程式
  - Poisson
  - 陰解法
- アーキテクチャー
  - MPP
  - Multicore, Hybrid
  - Vector
- Library
  - Multicolor SOR-CMA, MG, FMM
  - Trillinos

COUPLING HANDLING



Data flow description
&
Control flow description

Acting Force

Fluid Simulation
$\partial_t Q + \partial_x E + \partial_y F + \partial_z G = 0$

Structural Simulation
$M \cdot \ddot{u} + C \cdot \dot{u} + Q = Faero$

Displacement

Compressible Solver

Shell Solver

interface Information

interface Information

Base Class for Fluid code

Base Class for Structural code

XML Parsing

Data class

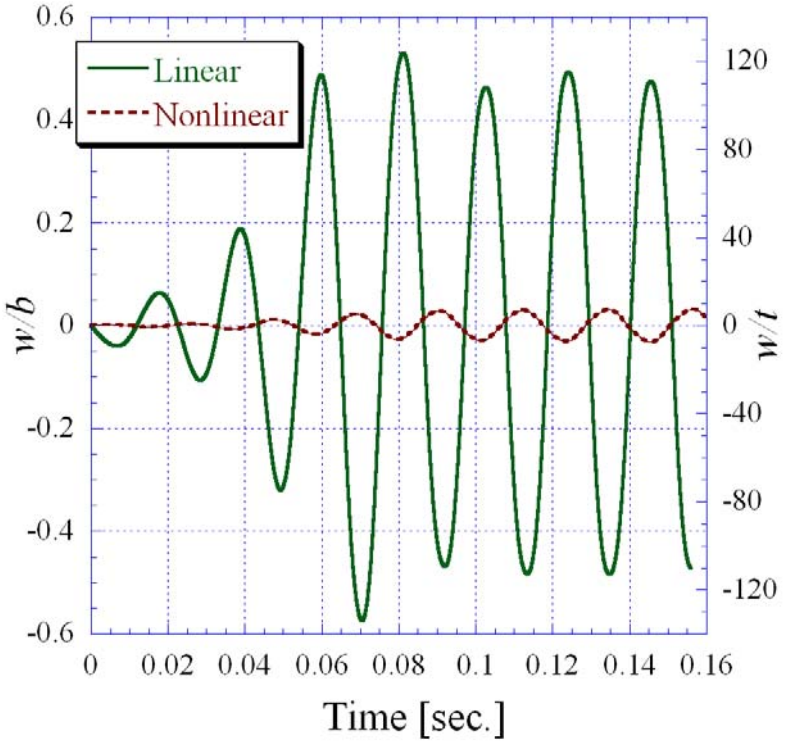
File I/O

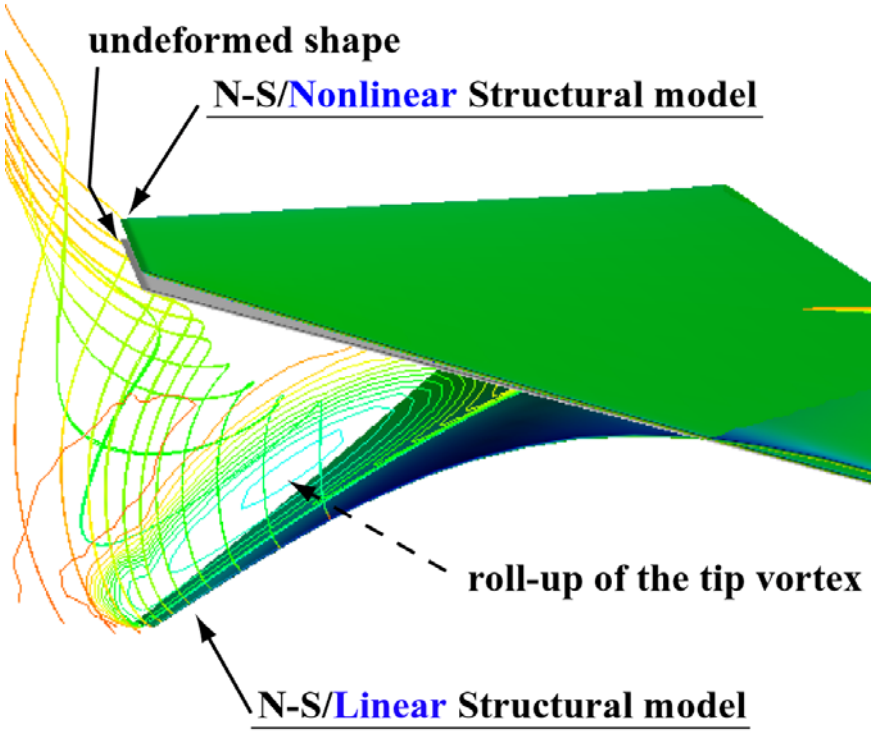
Solver Base class

Parallel Manager / Data Manager

MPICH

翼フラッター時の非線形挙動





ソルバーの準備

- コンセプトの理解
  - オブジェクト指向
- 記述する部分の把握
  - 提供クラスのヘッダを読む
- コーディングの方法
  - Common文の削除
  - 動的メモリ確保
- パラメータ読み込み
  - XMLパーサーの利用
- ソルバークラスのSPHEREへの登録
  - 簡単な手順
- Fortranのcommon文は利用しない
- フレームワークの機能を用い，動的なメモリ確保
- 配列領域はメモリの先頭アドレスをFortranサブルーチンへ引数渡し
- メインルーチンはC++

アプリ・ミドルウェアの機能

- 実行制御
  - マルチスケール/フィジックス連成
  - N次元モデルの統合（循環器系解析の0次元～3次元のハイブリッド化）
  - 分散高並列
  - スクリプトorビジュアルフローによる実行制御
- 開発支援
  - ライブラリ提供（I/O，数値ライブラリ），コード共有
  - ひな形によるラピッドプロト
  - アプリレベルのMPI並列/マルチコア開発支援
  - 外部アプリとの連携（最適化ソフトなど）
- コード記述方法の検討
  - 高レベル記述　と　実行性能　のバランス
  - 抽象化レベル（支配方程式，スキーム，格子系，解法など）

次世代HPCにおける大規模データの可視化環境

小野謙二
理化学研究所
次世代計算科学研究開発プログラム
生命体基盤ソフトウェア開発・高度化チーム

ASE研究会　2008

大規模並列計算の後処理における問題点

- データの大規模性
  - 空間規模，時系列データ，多変数
- データの分散性
  - 分散並列，GRID
- システムの複雑性
  - ヘテロ環境，ファイルシステム，ネットワーク

- データコピー，移動が高コスト
  - 処理時間，MMU/HDD容量
- データは動かせない
  - 適切なツールなしにはデータアクセスさえ不可
  - 動かせたとしても，処理手続きが煩雑，面倒
  - ユーザの心理的ハードルは高い

- 本当にやりたい解析以外の処理が少なくない
  - ファイル操作，処理準備
  - 「考えること」に集中できる環境整備が必要

大規模計算の後処理システムの方向性

- データを動かさない
  - その先には．．．
    - データ共有・協調・協同作業
    - リモート，仮想組織
- 集約的なデータサービスが中心となる
  - アクセス，分析，処理
  - データ，結果，知識，リソースの共有
    - グループ単位のデータリポジトリ
    - データのブラウズ，検索，分析

  - 現象理解，知識の共有

巨大リソースを利用したシミュレーション

- チームによるプロジェクト推進
  - 複数研究者のコラボレーション
- 結果の共有化の仕組みが必要
  - コミュニティ群（アプリケーション毎ほか）
  - 共有する情報
    - データ，メソッド，プログラム，結果，知識，．．．
  - きめ細かなprivilege制御がシステムに必要

ポストフェーズの特徴

- 計算と異なり主観的作業
  - ユーザの数だけ処理のシナリオがある
- インタラクティブ性が効率に大きく影響
  - 計算処理システムと同じ運用では，ユーザの利便性・生産性が高まらない
- ポストフェーズのリソース運用が効率化の鍵

- 専用リソース（CPU/GPU cluster）の有用性
  - 柔軟な処理への対応
  - 多様なソフトウェアシステムの利用が想定される
  - 標準的なLinux clusterシステムが既存資産を利用可能で利便性が高い

提案するポスト処理

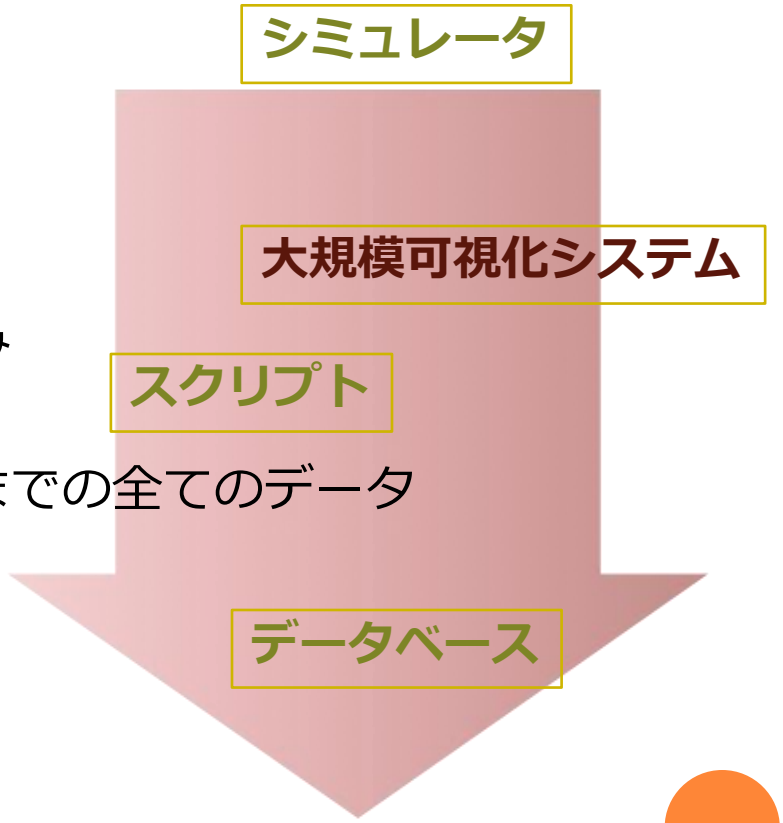
- 後処理の各プロセスを緩やかに有機的に結合したシステム
- スクリプトベースのツール・アプリ群
- 共同・協同作業のサポートの織り込み

- 研究者のリクエストのボトムアップ

- 研究者の「考えること」を支援
  - システムの大規模化によって生じる複雑性が問題

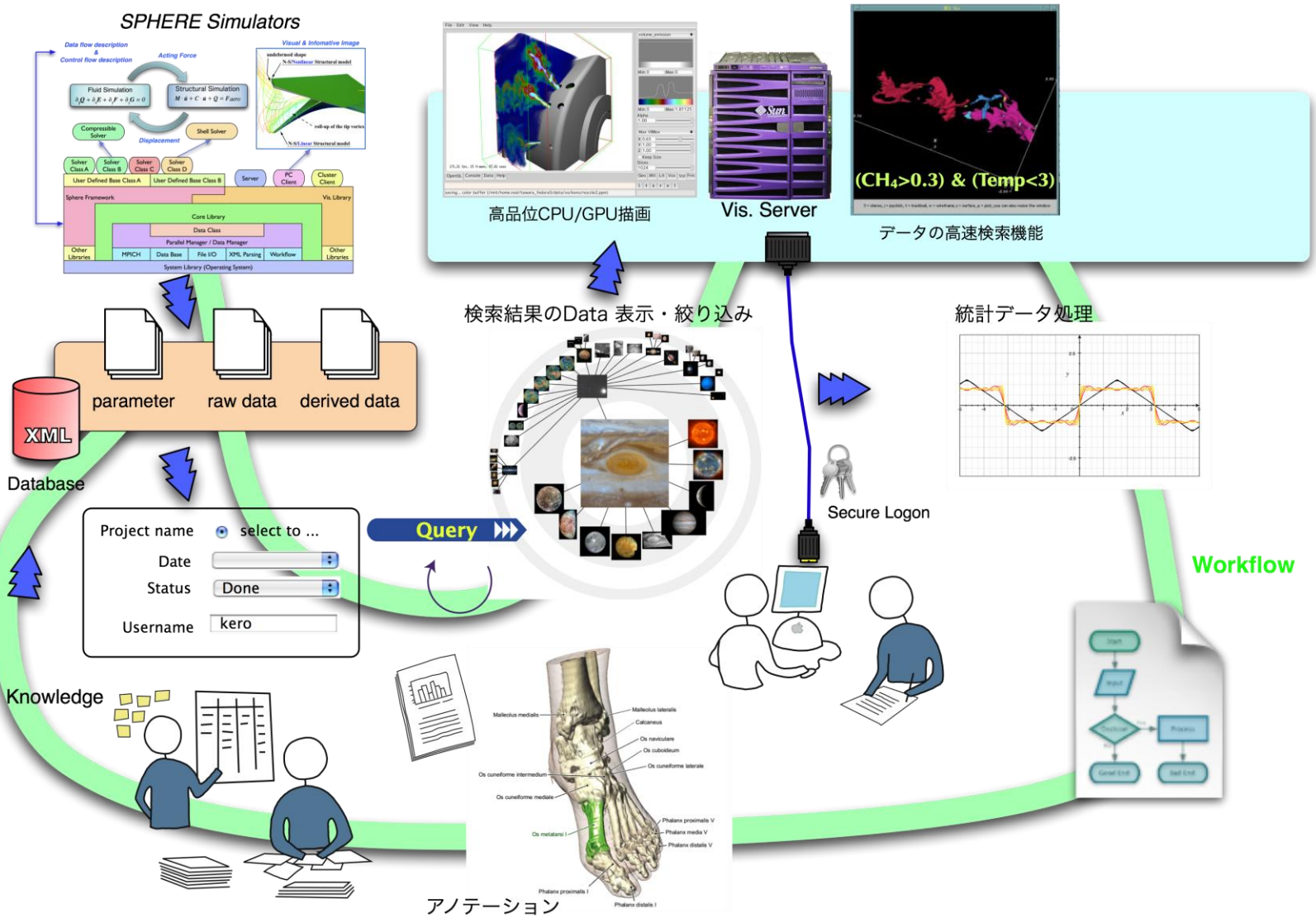
ポスト処理のコンセプト

- シミュレーション結果から，有用な知識・情報を効率的に引き出す
  - 様々な可視化手法による場の把握
  - 生成された2次データの利用
  - イメージ化
  - データ分析
  - 多様・大量なデータの管理のしくみ
- データ管理
  - シミュレーション開始から分析後までの全てのデータ
  - 個々のデータの関連づけ
- 処理の効率化
  - 定型作業の自動処理
- ユーザ作業環境
  - 単独で動作するコマンド，アプリの連携・緩やかな結合



シミュレーションの価値を増幅させる，知識を得るためのツール群

ポスト処理のイメージ


SPHERE Simulators
Data flow description & Control flow description
Acting Force
Visual & Infomative Image
Fluid Simulation
Structural Simulation
Displacement
Compressible Solver
Shell Solver
Sphere Framework
Vis. Library
Core Library
Data Class
Parallel Manager / Data Manager
MPICH
Data Base
File I/O
XML Parsing
Workflow
System Library (Operating System)
高品位CPU/GPU描画
Vis. Server
(CH4>0.3) & (Temp<3)
データの高速検索機能
検索結果のData 表示・絞り込み
統計データ処理
parameter
raw data
derived data
XML
Database
Project name
select to ...
Date
Status
Done
Username
kero
Query
Secure Logon
Workflow
Knowledge
アノテーション

開発中の可視化システム

- リモート可視化とローカル可視化
  - 共通クライアントによる統一環境の提供
- リアルタイム可視化とポスト可視化
  - ファイル経由の可視化を原則
- インタラクティブ可視化とバッチ可視化
- SWレンダリングとHWレンダリング
- 大規模データハンドリング
- 並列可視化
- 可視化基本ライブラリと機能モジュール群の構造
  - 将来の機能拡張と実装の容易性を担保
- 移植性
  - Linux, Windows, Max OSX, PC cluster, Supercomputer
- 他の可視化アプリとの連携
  - リモート可視化プロトコルの共通化

まとめ

- ポスト処理環境の提案
  - ユーザの可視化シナリオベース
- 大規模データ可視化システム
  - インタラクティブ性能の分析
  - 要素技術の検討
  - システム設計