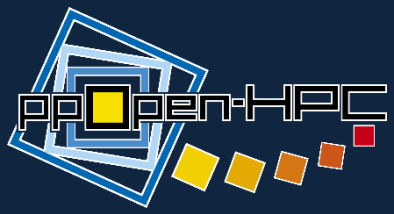
全国共同利用施設
東京大学情報基盤センター
Information Technology Center, The University of Tokyo

東京大学
THE UNIVERSITY OF TOKYO



ppOpen-HPCの概要
自動チューニング機構を有するアプリケーション開発・実行環境

松本正晴，片桐孝洋，中島 研吾
東京大学情報基盤センター

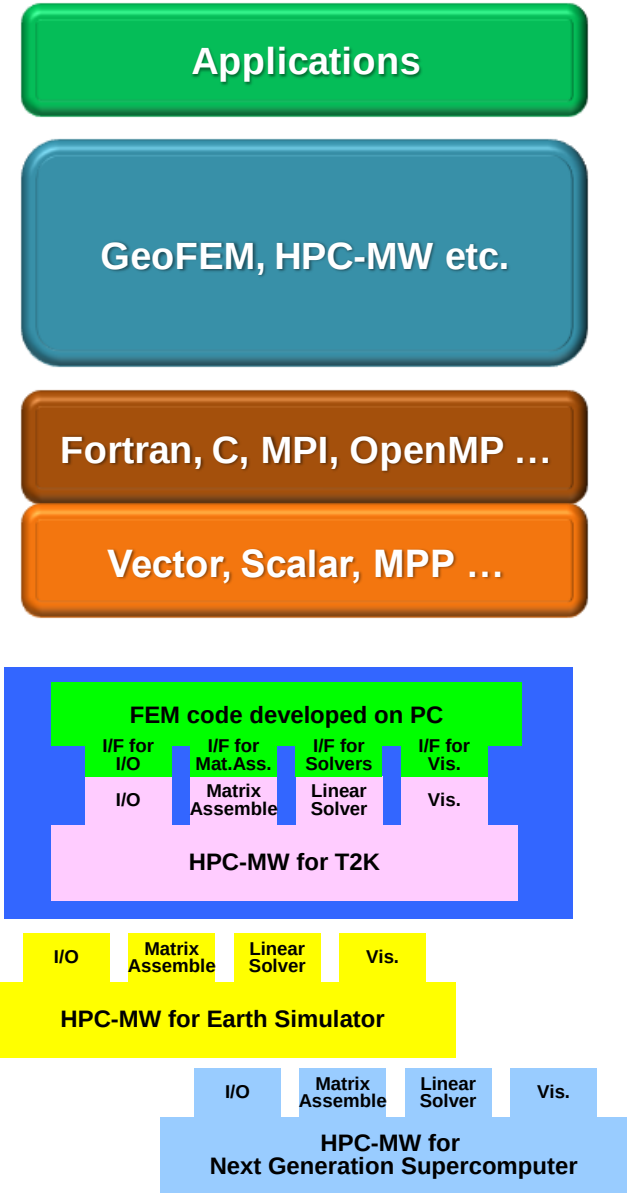
第44回お試しアカウント付き並列プログラミング講習会
「ライブラリ利用：高性能プログラミング初級入門」
2015年3月26日～27日

背　景(1/2)

- 大規模化，複雑化，多様化するハイエンド計算機環境の能力を充分に引き出し，効率的なアプリケーションプログラムを開発することは困難
- 有限要素法等の科学技術計算手法：
  - プリ・ポスト処理，行列生成，線形方程式求解等の一連の共通プロセスから構成される。
  - これら共通プロセスを抽出し，ハードウェアに応じた最適化を施したライブラリとして整備することで，アプリケーション開発者から共通プロセスに関わるプログラミング作業，並列化も含むチューニング作業を隠蔽できる。
  - アプリケーションMW，HPC-MW，フレームワーク

背　景(2/2)

- A.D.2000年前後
  - GeoFEM，HPC-MW
  - 地球シミュレータ，Flat MPI，FEM
- 現在：より多様，複雑な環境
  - マルチコア，GPU
  - ハイブリッド並列
    - MPIまでは何とかたどり着いたが・・・
    - 「京」でも重要
  - CUDA，OpenCL，OpenACC
  - ポストペタスケールからエクサスケールへ
    - より一層の複雑化



http://www.top500.org/

R_{max}：実効性能（TFLOPS），R_{peak}：ピーク性能（TFLOPS），Power: kW

	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	National Supercomputer Center in Guangzhou, China	**Tianhe-2A** Intel Xeon E5-2692, TH Express-2, IXeon Phi2013 NUDT	3120000	33863 (= 33.9 PF)	54902	17808
2	Oak Ridge National Laboratory, USA	**Titan** Cray XK7/NVIDIA K20x, 2012 Cray	560640	17590	27113	8209
3	Lawrence Livermore National Laboratory, USA	**Sequoia** BlueGene/Q, 2011 IBM	1572864	17173	20133	7890
4	RIKEN AICS, Japan	**K computer**, SPARC64 VIIIfx , 2011 Fujitsu	705024	10510	11280	12660
5	Argonne National Laboratory, USA	**Mira** BlueGene/Q, 2012 IBM	786432	8586	10066	3945
6	Swiss National Supercomputing Ctr. (CSCS), Switzerland	**Piz Daint** Cray XC30, Xeon E5-2670 8C, NVIDIA K20x, 2013 Cray	115984	6271	7789	2325
7	TACC, USA	**Stampede** Xeon E5-2680/Xeon Phi, 2012 Dell	462462	5168	8520	4510
8	Forschungszentrum Juelich (FZJ), Germany	**JuQUEEN** BlueGene/Q, 2012 IBM	458752	5009	5872	2301
9	DOE/NNSA/LLNL, USA	**Vulcan** BlueGene/Q, 2012 IBM	393216	4293	5033	1972
10	U.S. government, USA	- Cray CS-Storm, Xeon E5-2660v2 10C, NVIDIA K40, 2014 Cray	72800	3577	6131	1499
48	**ITC/U. Tokyo Japan**	**Oakleaf-FX SPARC64 IXfx, 2012 Fujitsu**	**76800**	**1043**	**1135**	**1177**

TOP500 List Highlights, Nov 2014

- Total combined performance of all 500 systems has grown to 309 Pflop/s, compared to 274 Pflop/s in June and 250 Pflop/s one year ago. This increase in installed performance also exhibits a noticeable slowdown in growth compared to the previous long-term trend.
- There are 50 systems with performance greater than 1 petaflop/s on the list, up from 37 six months ago.
- The No. 1 system, Tianhe-2, and the No. 7 system, Stampede, use Intel Xeon Phi processors to speed up their computational rate. The No. 2 system, Titan, and the No. 6 system, Piz Daint, use NVIDIA GPUs to accelerate computation.
- A total of 75 systems on the list are using accelerator/co-processor technology, up from 62 from November 2013. Fifty of these use NVIDIA chips, three use ATI Radeon, and there are now 25 systems with Intel MIC technology (Xeon Phi). Intel continues to provide the processors for the largest share (85.8 percent) of TOP500 systems.
- Ninety-six percent of the systems use processors with six or more cores and 85 percent use eight or more cores.
- HP has the lead in systems with 179 (36 percent) compared to IBM with 153 systems (30 percent). HP had 182 systems (36.4 percent) six months ago, and IBM had 176 systems (35.2 percent) six months ago. In the system category, Cray remains third with 62 systems (12.4 percent).

http://www.top500.org/

東大情報基盤センターのスパコン

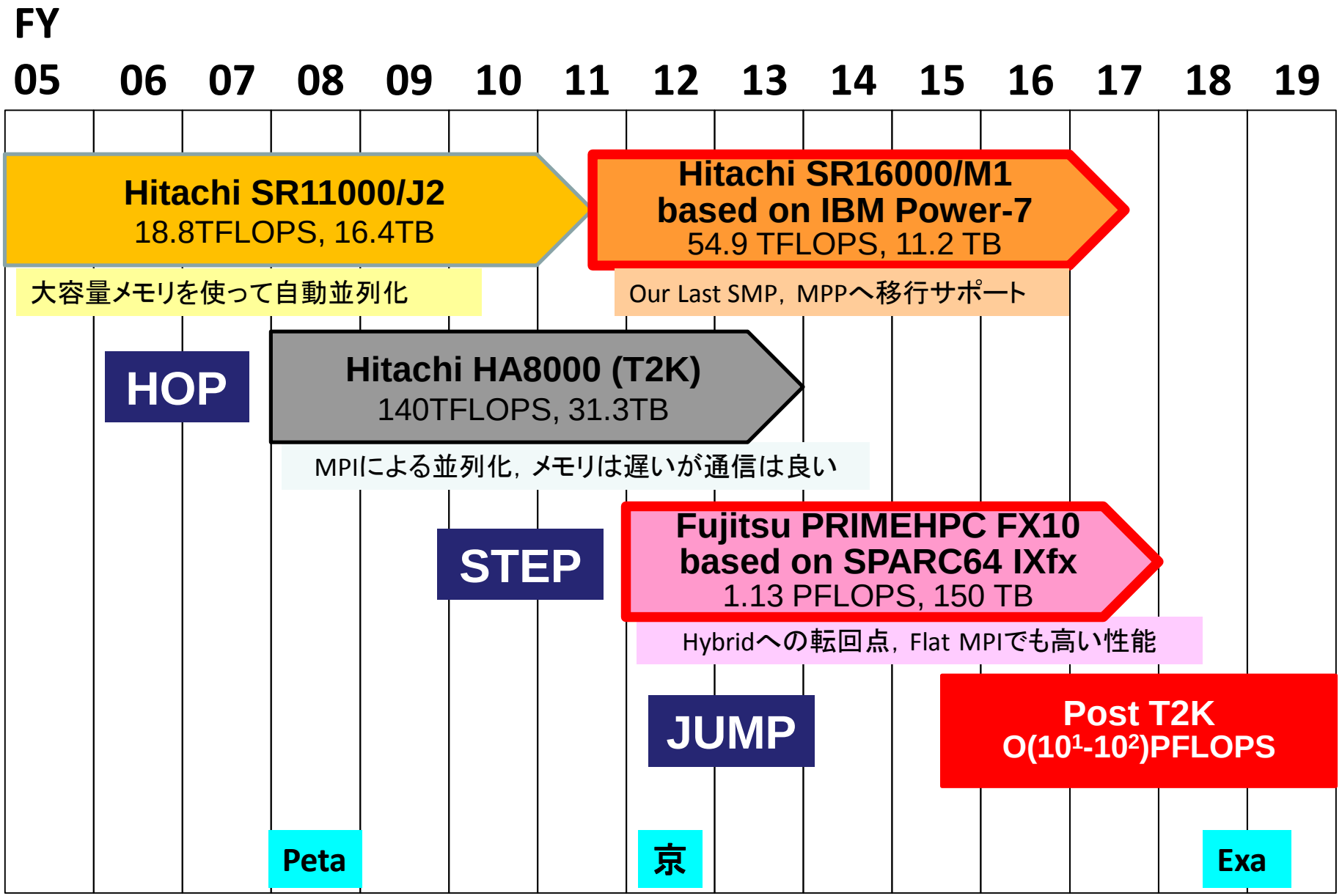
1システム～6年，3年周期でリプレース

Oakleaf-FX (Fujitsu PRIMEHPC FX10)		T2K-Todai（2014年3月退役）（Hitachi HA8000-tc/RS425 ）		Yayoi (Hitachi SR16000/M1)	
Total Peak performance	: 1.13 PFLOPS	Total Peak performance	: 140 TFLOPS	Total Peak performance	: 54.9 TFLOPS
Total number of nodes	: 4800	Total number of nodes	: 952	Total number of nodes	: 56
Total memory	: 150 TB	Total memory	: 32000 GB	Total memory	: 11200 GB
Peak performance / node	: 236.5 GFLOPS	Peak performance / node	: 147.2 GFLOPS	Peak performance / node	: 980.48 GFLOPS
Main memory per node	: 32 GB	Main memory per node	: 32 GB, 128 GB	Main memory per node	: 200 GB
Disk capacity	: 1.1 PB + 2.1 PB	Disk capacity	: 1 PB	Disk capacity	: 556 TB
SPARC64 Ixfx 1.84GHz		**AMD Quad Core Opteron 2.3GHz**		**IBM POWER 7 3.83GHz**	

“Oakbridge-fx” with 576 nodes installed in April 2014 (separated) (136TF)

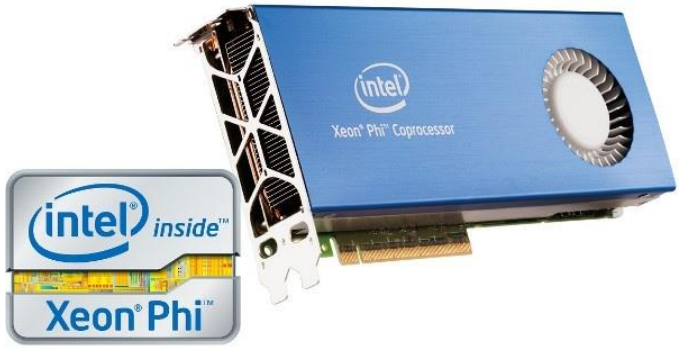
Total Users > 2,000

東大情報基盤センターのスパコン



Post T2K System

- 20-30 PFLOPS, FY.2015
- Many-core based (e.g. (only) Intel MIC/Xeon Phi)
- Joint Center for Advanced High Performance Computing (最先端共同HPC基盤施設，JCAHPC, http://jcahpc.jp/)
  - 筑波大学計算科学研究センター，東京大学情報基盤センター
- Programming is still difficult, although Intel compiler works.
  - (MPI + OpenMP)
  - Tuning for performance (e.g. prefetching) is essential
  - Some framework for helping users needed



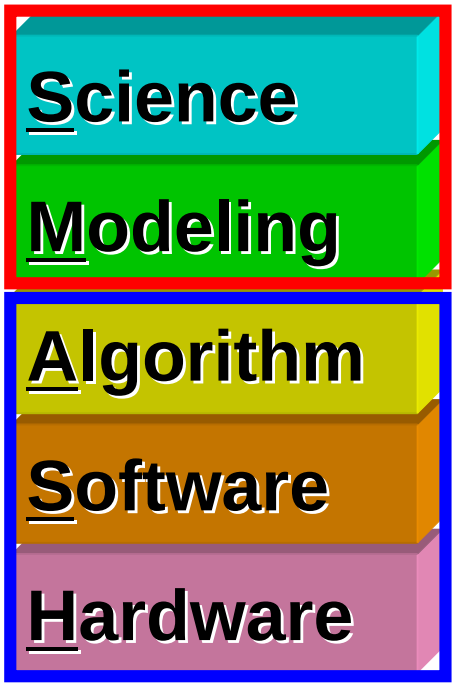
Key-Issues for Appl's/Algorithms towards Post-Peta & Exa Computing

Jack Dongarra (ORNL/U. Tennessee) at ISC 2013

- Heterogeneous/Hybrid Architecture
- Communication/Synchronization Reducing Algorithms
- Mixed Precision Computation
- Auto-Tuning/Self-Adapting
- Fault Resilient Algorithms
- Reproducibility of Results

HPCミドルウェア：何がうれしいか

- アプリケーション開発者のチューニング（並列，単体）からの解放
  - SMASHの探求に専念
    - 一生SMASHと付き合うのはきつい
  - SMASHをカバー
- コーディングの量が減る
- 教育にも適している
- 問題点
  - ハードウェア，環境が変わるたびに最適化が必要となる



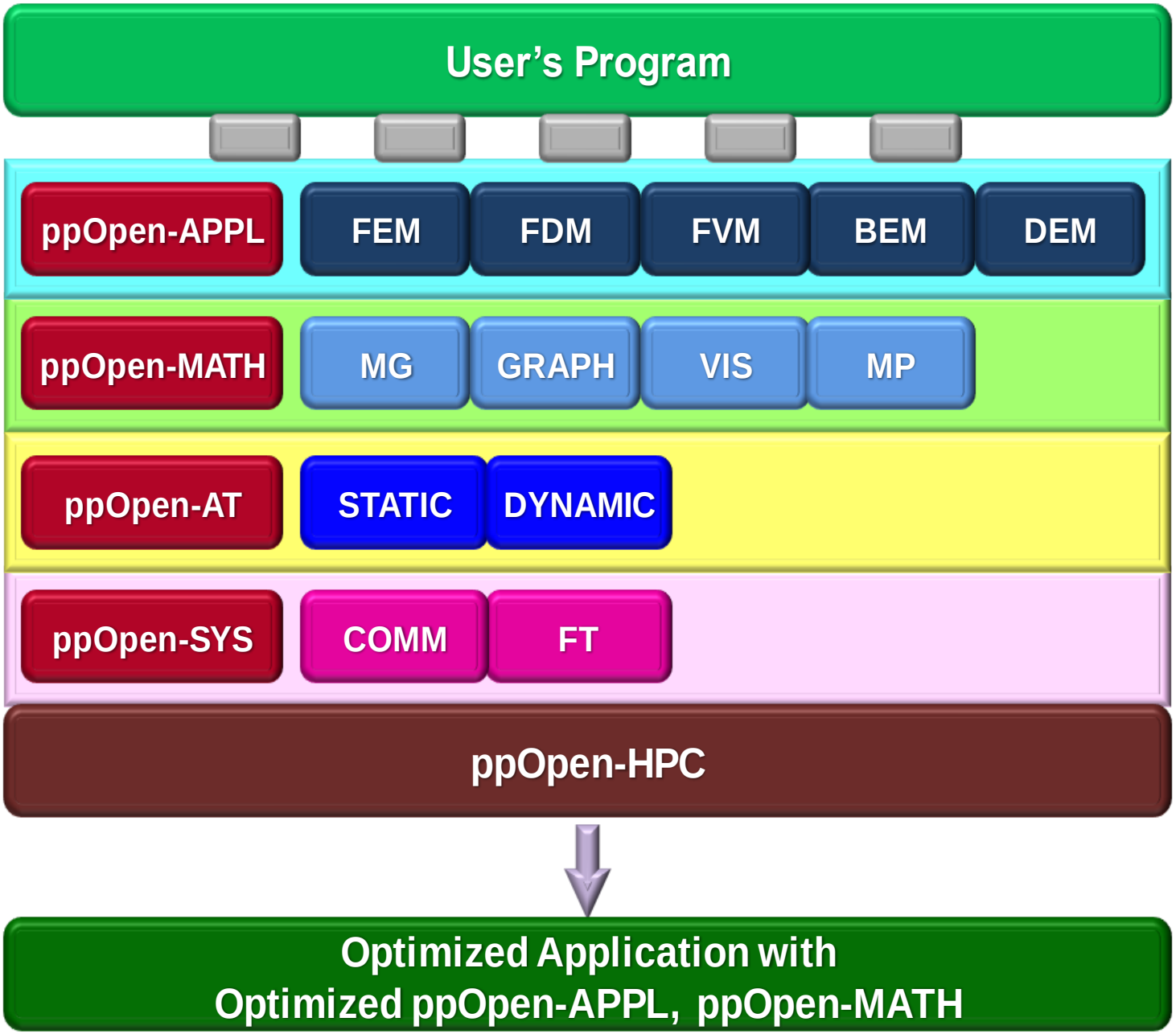
HPCのCo-Design

ppOpen-HPC

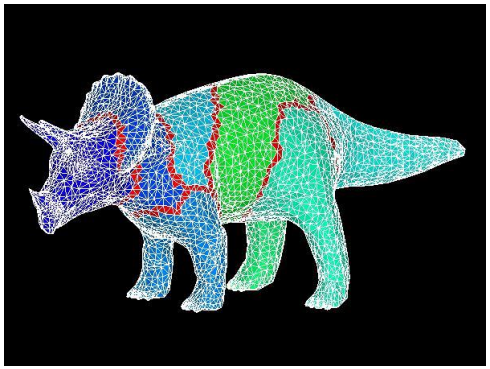
- 東京大学情報基盤センターでは，メニィコアに基づく計算ノードを有するポストペタスケールシステムの処理能力を充分に引き出す科学技術アプリケーションの効率的な開発，安定な実行に資する「自動チューニング機構を有するアプリケーション開発・実行環境：ppOpen-HPC」を開発中。
  - 科学技術振興機構戦略的創造研究推進事業（CREST）研究領域「ポストペタスケール高性能計算に資するシステムソフトウェア技術の創出（Post-Peta CREST）」（2011～2015年度）（領域統括：米澤明憲教授（理化学研究所計算科学研究機構））
  - PI：中島研吾（東京大学情報基盤センター）
  - 東大（情報基盤センター，大気海洋研究所，地震研究所，大学院新領域創成科学研究科），京都大学術情報メディアセンター，北海道大学情報基盤センター，海洋研究開発機構
  - 様々な分野の専門家によるCo-Design

概 要(1/3)

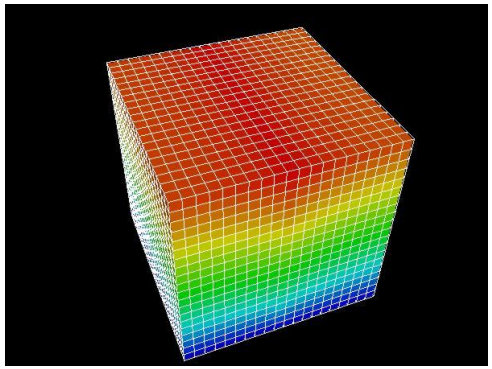
- メニーコアクラスタによるポストペタスケールシステム上での科学技術アプリケーションの効率的開発，安定な実行に資するppOpen-HPCの研究開発を計算科学，計算機科学，数理科学各分野の緊密な協力のもとに実施している。
  - 6 Issues in Post-Peta/Exascale Computingを考慮
  - “pp”: Post Peta
- 東大情報基盤センターに平成27年度導入予定のO(10)PFLOPS級システム（ポストT2K，Intel MIC/Xeon-Phiベース）をターゲット：
  - スパコンユーザーの円滑な移行支援
- 大規模シミュレーションに適した5種の離散化手法に限定し，各手法の特性に基づいたアプリケーション開発用ライブラリ群，耐故障機能を含む実行環境を実現する。
  - ppOpen-APPL：各手法に対応した並列プログラム開発のためのライブラリ群
  - ppOpen-MATH：各離散化手法に共通の数値演算ライブラリ群
  - ppOpen-AT：科学技術計算のための自動チューニング（AT）機構
  - ppOpen-SYS：ノード間通信，耐故障機能に関連するライブラリ群


User's Program
ppOpen-APPL
FEM
FDM
FVM
BEM
DEM
ppOpen-MATH
MG
GRAPH
VIS
MP
ppOpen-AT
STATIC
DYNAMIC
ppOpen-SYS
COMM
FT
ppOpen-HPC
Optimized Application with
Optimized ppOpen-APPL, ppOpen-MATH

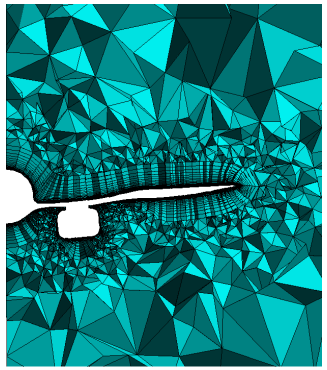
対象とする離散化手法
局所的，隣接通信中心，疎行列



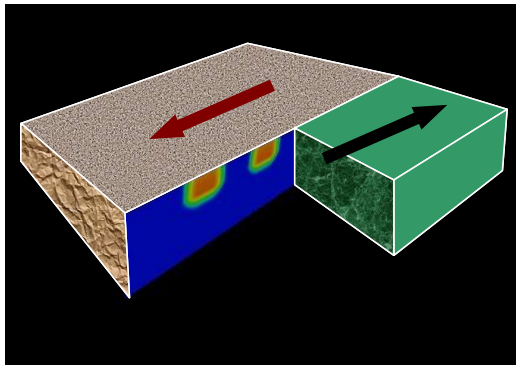
有限要素法
Finite Element Method
FEM



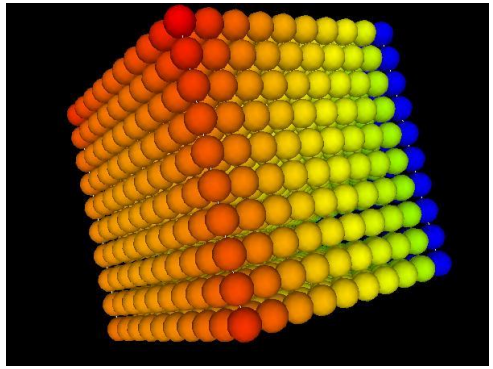
差分法
Finite Difference Method
FDM



有限体積法
Finite Volume Method
FVM



境界要素法
Boundary Element Method
BEM



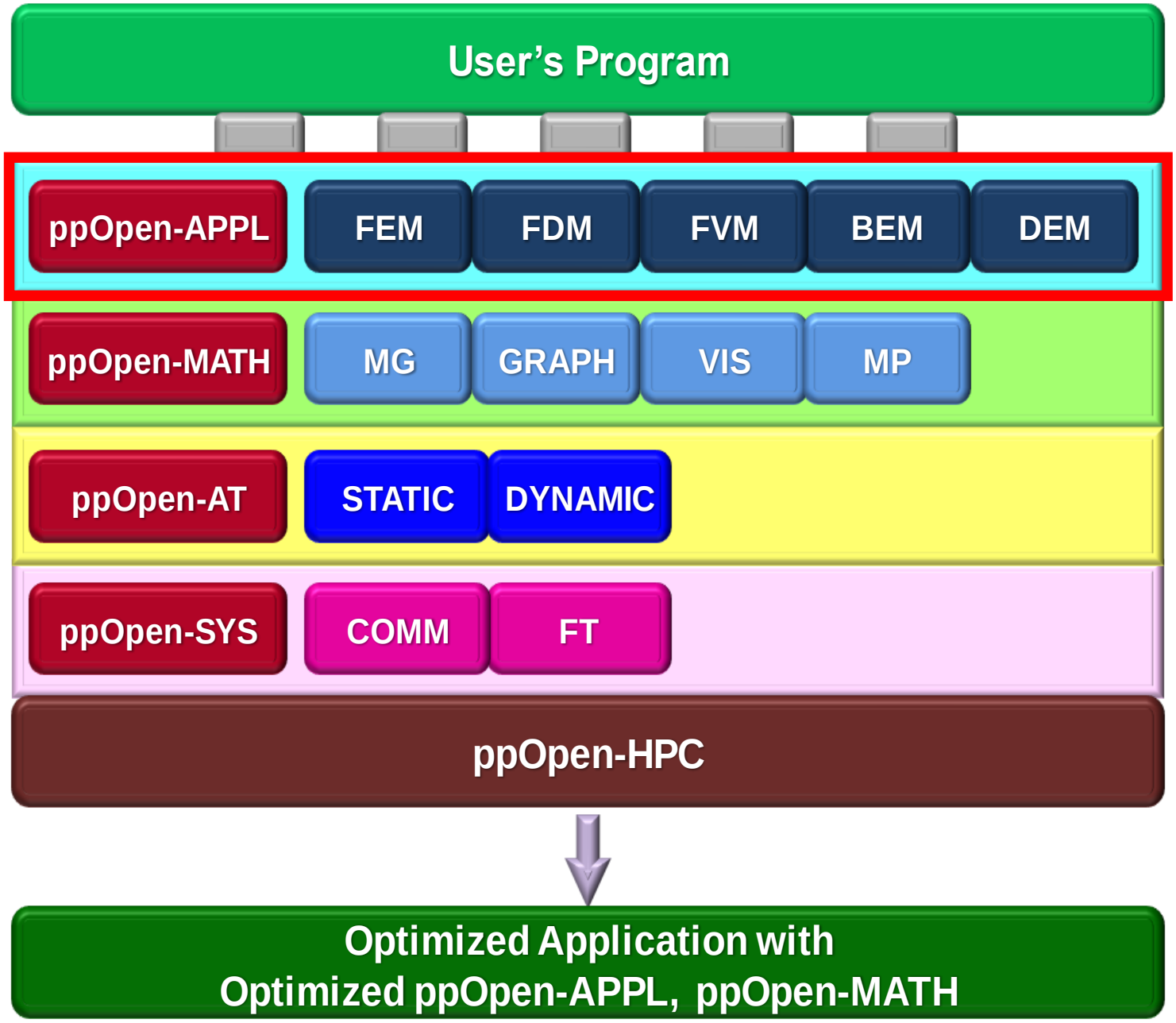
個別要素法
Discrete Element Method
DEM

概 要(2/3)

- 先行研究において各メンバーが開発した大規模アプリケーションに基づきppOpen-APPLの各機能を開発，実装
  - 各離散化手法の特性に基づき開発・最適化
    - 共通データ入出力インタフェース，領域間通信，係数マトリクス生成
    - 離散化手法の特性を考慮した前処理付き反復法
    - 適応格子，動的負荷分散
  - 実際に動いているアプリケーションから機能を切り出す
  - 各メンバー開発による既存ソフトウェア資産の効率的利用
    - GeoFEM，HEC-MW，HPC-MW，DEMIGLACE，ABCLibScript
- ppOpen-ATはppOpen-APPLの原型コードを対象として研究開発を実施し，その知見を各ppOpen-APPLの開発，最適化に適用
  - 自動チューニング技術により，様々な環境下における最適化ライブラリ・アプリケーション自動生成を目指す

概 要(3/3)

- 平成24年11月にマルチコアクラスタ向けに各グループの開発したppOpen-APPL，ppOpen-AT，ppOpen-MATHの各機能を公開（Ver.0.1.0）
  - http://ppopenhpc.cc.u-tokyo.ac.jp/
  - 平成25年11月にVer.0.2.0公開
  - 平成26年11月にVer.0.3.0公開
- 現在は各機能の最適化，機能追加，ppOpen-APPLによるアプリケーション開発とともに，Intel Xeon/Phi等メニーコア向けバージョンを開発中


User's Program
ppOpen-APPL
FEM
FDM
FVM
BEM
DEM
ppOpen-MATH
MG
GRAPH
VIS
MP
ppOpen-AT
STATIC
DYNAMIC
ppOpen-SYS
COMM
FT
ppOpen-HPC
Optimized Application with
Optimized ppOpen-APPL, ppOpen-MATH

ppOpen-APPL

- A set of libraries corresponding to each of the five methods noted above (FEM, FDM, FVM, BEM, DEM), providing:
  - I/O
    - netCDF-based Interface
  - Domain-to-Domain Communications
  - Optimized Linear Solvers (Preconditioned Iterative Solvers)
    - Optimized for each discretization method
  - H-Matrix Solvers in ppOpen-APPL/BEM
  - Matrix Assembling
  - AMR and Dynamic Load Balancing
- Most of components are extracted from existing codes developed by members

FEM Code on ppOpen-HPC

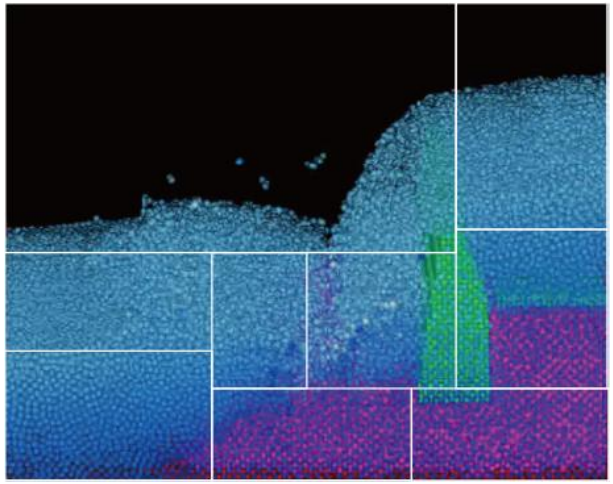
Optimization/parallelization could be hidden from application developers

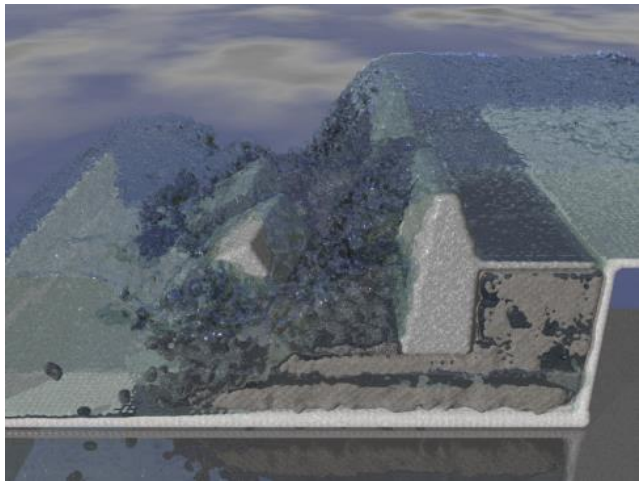
```
Program My_pFEM
use ppOpenFEM_util
use ppOpenFEM_solver

call ppOpenFEM_init
call ppOpenFEM_cntl
call ppOpenFEM_mesh
call ppOpenFEM_mat_init

do
  call Users_FEM_mat_ass
  call Users_FEM_mat_bc
  call ppOpenFEM_solve
  call ppOPenFEM_vis
  Time= Time + DT
enddo

call ppOpenFEM_finalize
stop
end
```

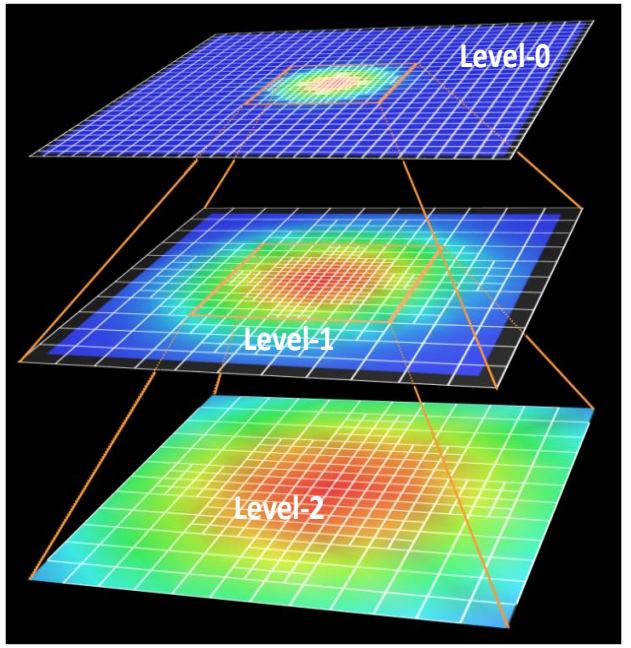
Target Applications

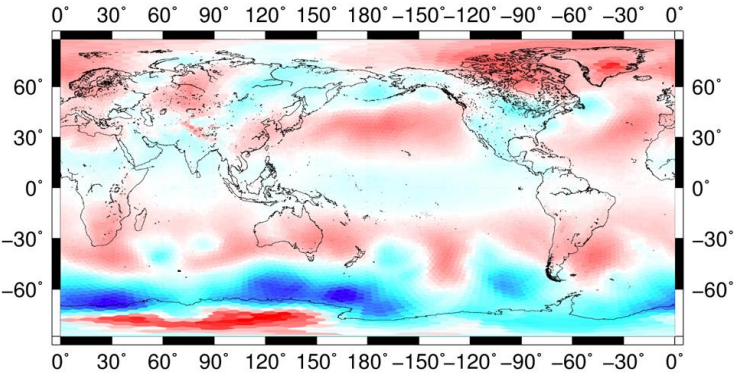
- Our goal is not development of applications, but we need some target appl. for evaluation of ppOpen-HPC.

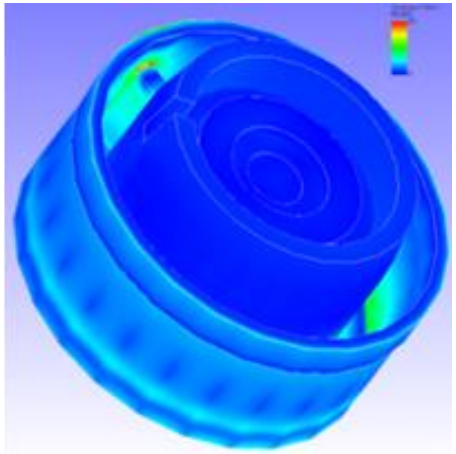
- ppOpen-APPL/FEM
  - Incompressible Navier-Stokes
  - Heat Transfer, Solid Mechanics (Static, Dynamic)
- ppOpen-APPL/FDM
  - Elastic wave propagation
  - Incompressible Navier-Stokes
  - Transient Heat Transfer, Solid Mechanics (Dynamic)
- ppOpen-APPL/FVM
  - Compressible Navier-Stokes, Heat Transfer
- ppOpen-APPL/BEM
  - Electromagnetics, Solid Mechanics (Quasi Static) (Earthquake Generation Cycle)
- ppOpen-APPL/DEM
  - Incompressible Navier-Stokes, Solid Mechanics (Dynamic)

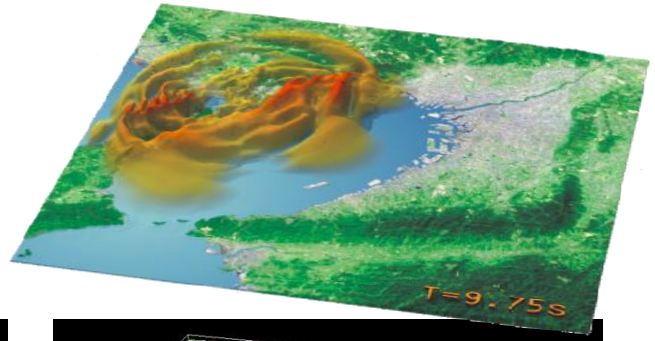


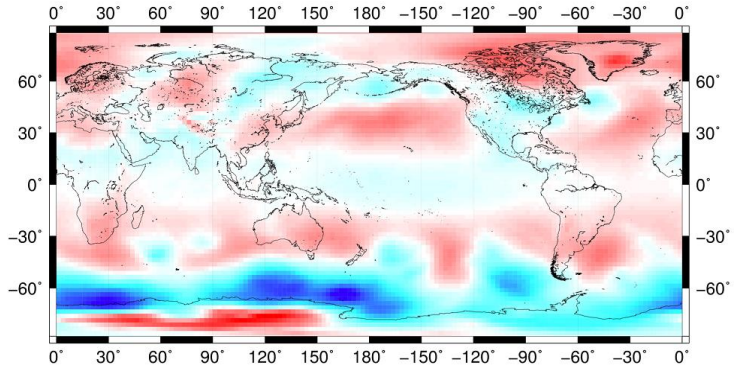


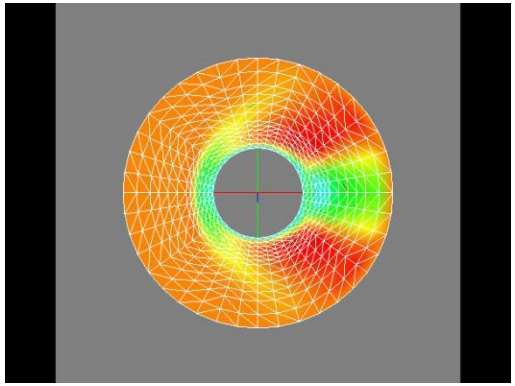












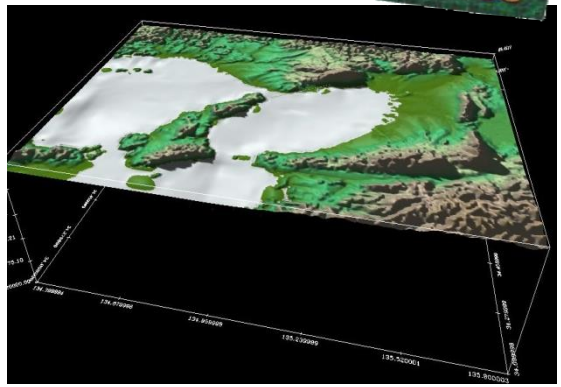
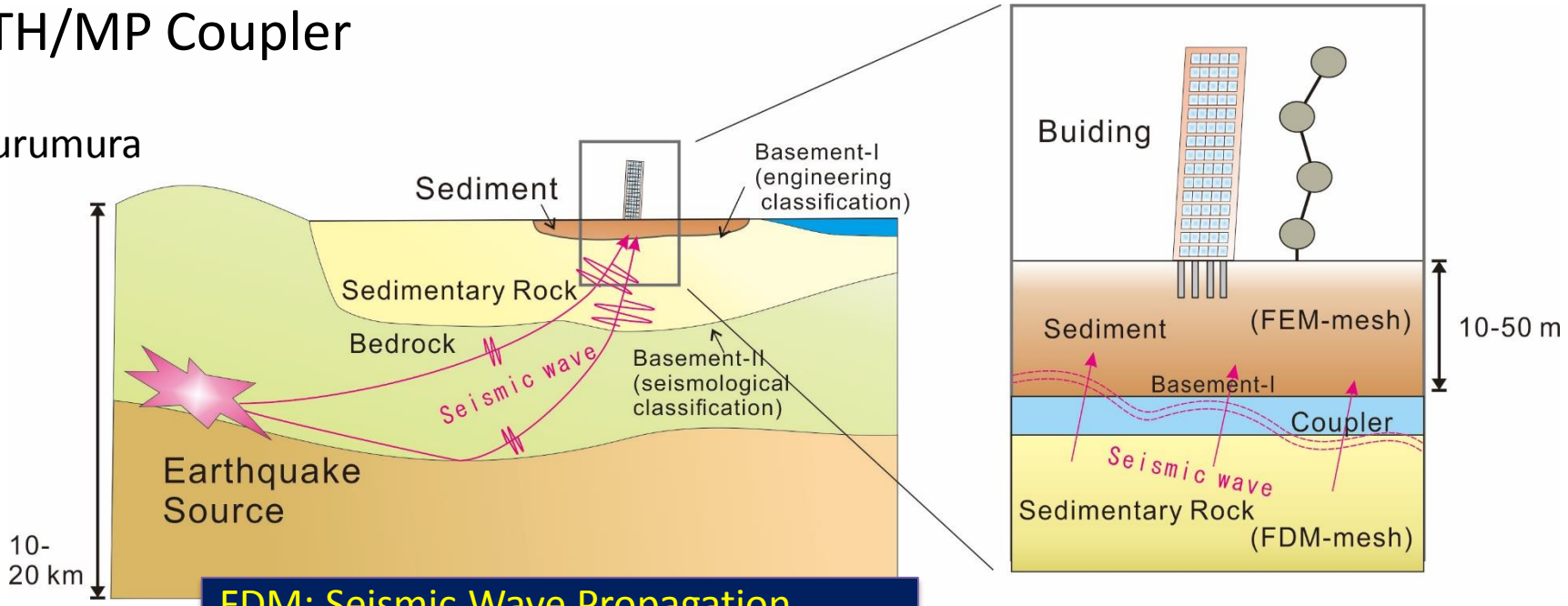


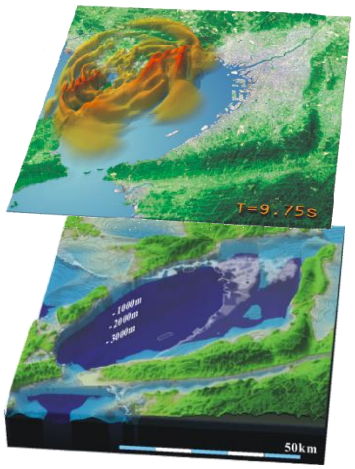
図 1　NICAM と IO モジュールの海面気圧

Large-Scale Coupled Simulations in FY.2014

Challenge (FY2014) : A test of a coupling simulation of FDM (regular grid) and FEM (unconstructed grid) using newly developed ppOpen-MATH/MP Coupler

c/o T.Furumura

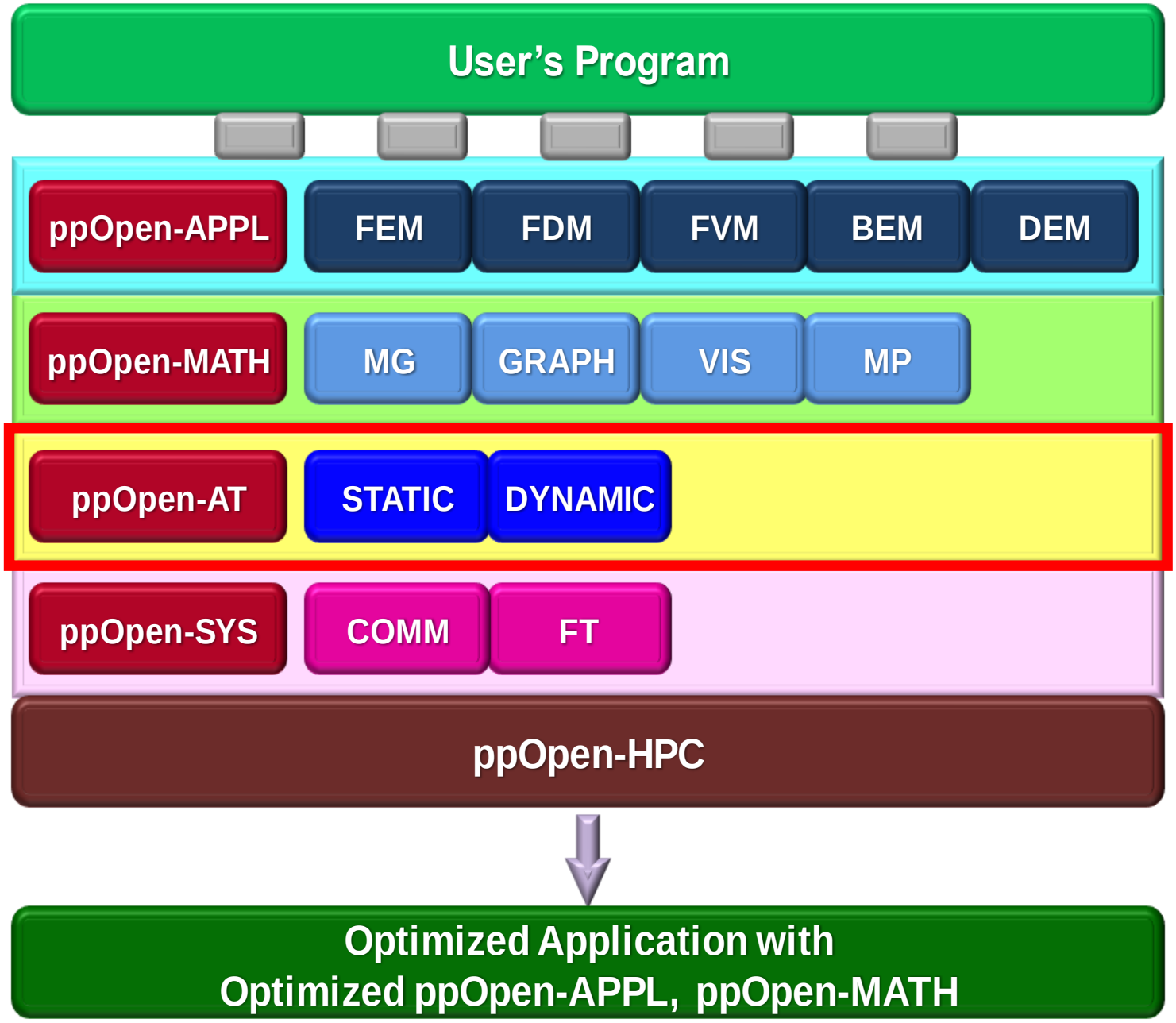




FDM: Seismic Wave Propagation
Model size: 80x80x400 km
Time : 240 s
Resolution (space) : 0.1 km (regular)
Resolution (time) : 5 ms
(effective freq.<1 Hz)

FEM: Building Response
Model size : 400x400x200 m
Time : 60 s
Resolution (space) : 1 m
Resolution (time) : 1 ms

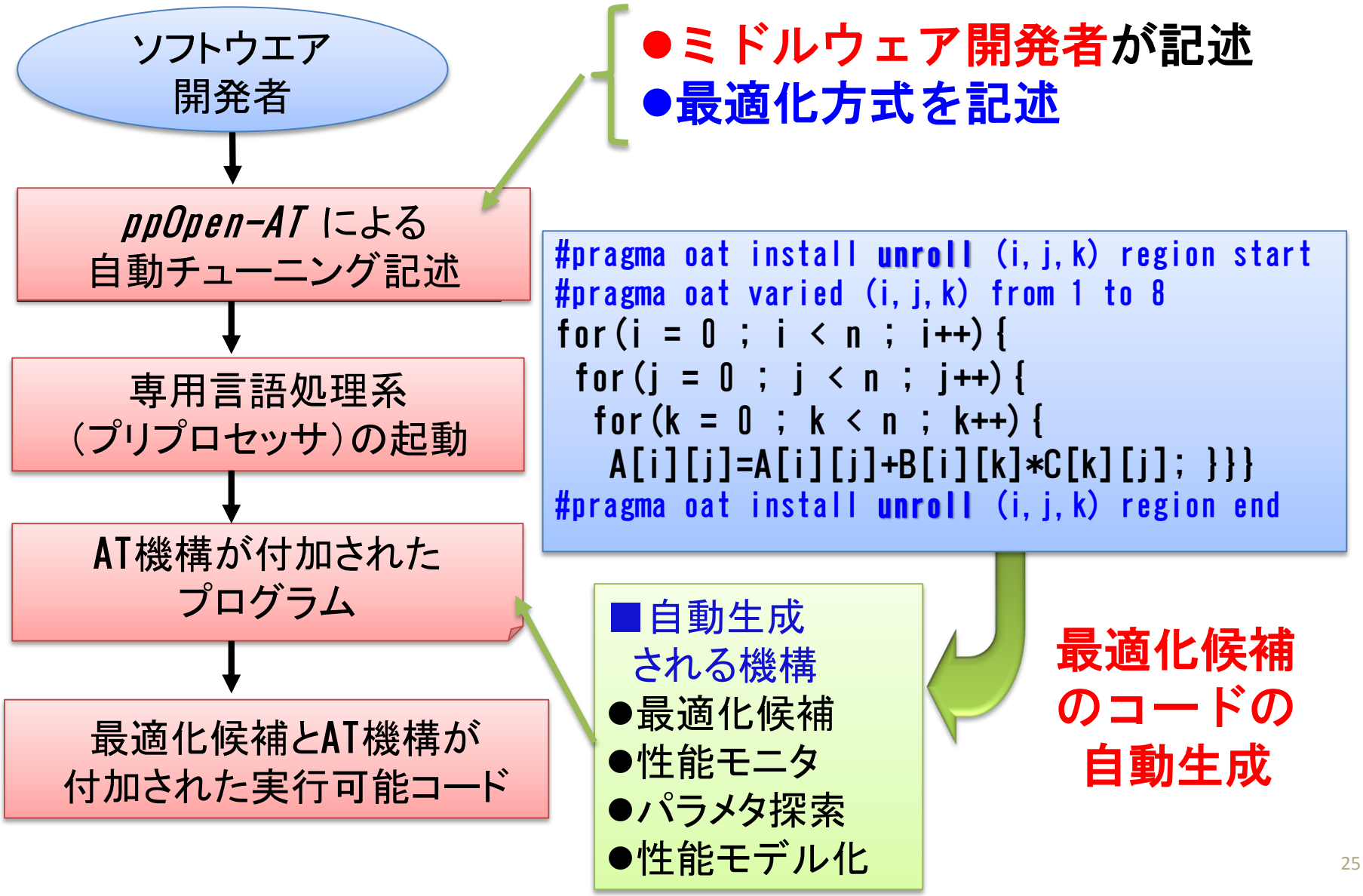
ppOpen-MATH/MP: Space-temporal interpolation, Mapping between FDM and FEM mesh, etc.


User's Program
ppOpen-APPL
FEM
FDM
FVM
BEM
DEM
ppOpen-MATH
MG
GRAPH
VIS
MP
ppOpen-AT
STATIC
DYNAMIC
ppOpen-SYS
COMM
FT
ppOpen-HPC
Optimized Application with
Optimized ppOpen-APPL, ppOpen-MATH

ppOpen-AT

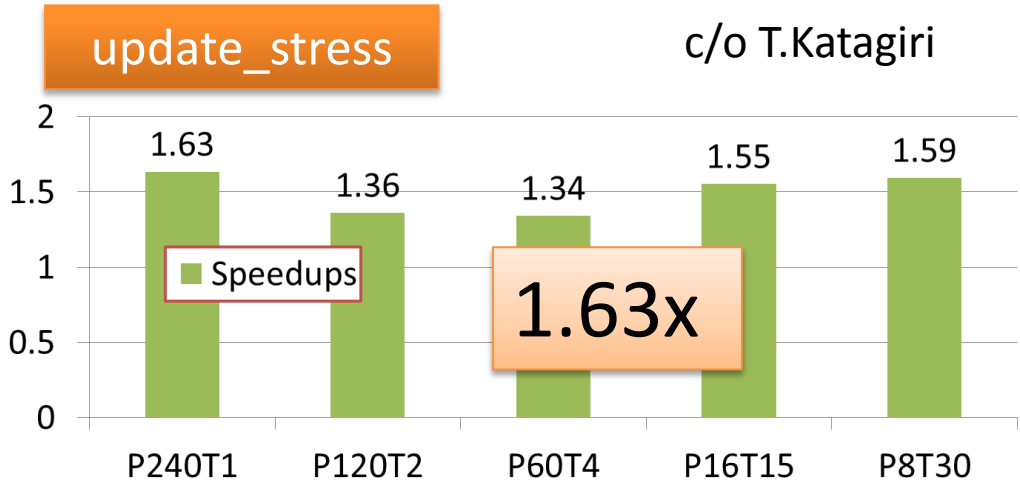
- Automatic tuning (AT) enables smooth and easy shifts to further development on new and future architectures, through use of ppOpen-AT/STATIC and ppOpen-AT/DYNAMIC.
- A special directive-based AT language based on ABCLibscript is developed for specific procedures in scientific computing, focused on optimum memory access.
  - Geometries
  - Problem size
  - H/W Parameters

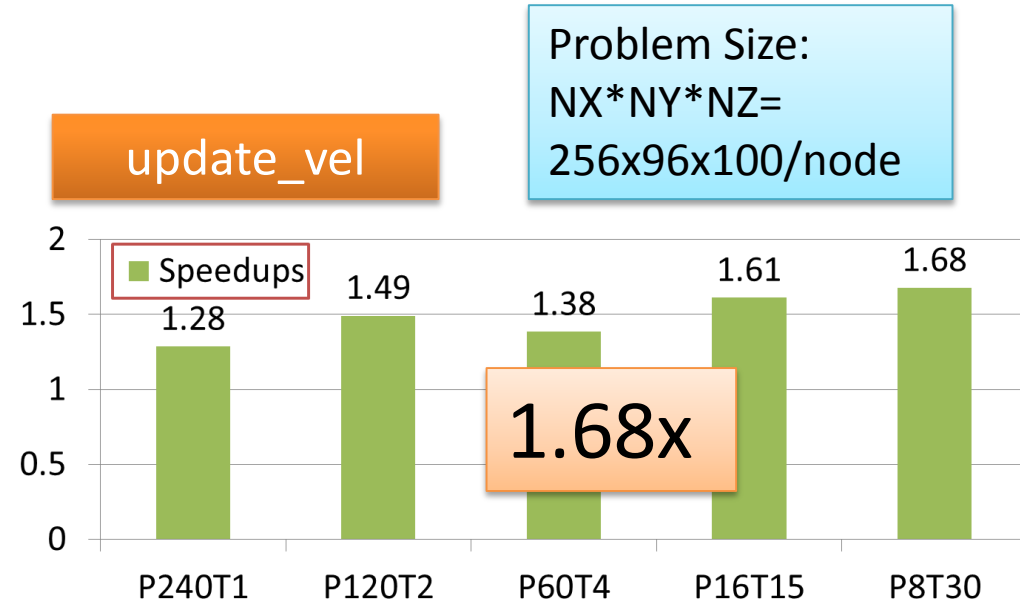
AT専用言語*ppOpen-AT*によるソフトウェア開発手順



Optimization of ppOpen-APPL/FDM (Seism3D) by ppOpen-AT

- **A single node of Intel Xeon Phi (60 cores, 240 threads)**
  - P240T1: Flat MPI (240 process with 1 thread)
  - P8T30: 8 proc's with 30 threads
  - Speed-ups based on execution without auto-tuning
- **`update_stress`**
  - 3-nested FDM loops, with a lot of operations
  - "Loop Splitting" is effective
- **`update_vel`**
  - 3-nested FDM loops, medium amount of operations
  - "Loop Fusion" is effective (i-j-k -> i*j-k)





Schedule of Public Release

(with English Documents, MIT License)

We are now focusing on MIC/Xeon Phi

- 4Q 2012 (Ver.0.1.0)
  - ppOpen-HPC for Multicore Cluster (Cray, K etc.)
  - Preliminary version of ppOpen-AT/STATIC
- 4Q 2013 (Ver.0.2.0)
  - ppOpen-HPC for Multicore Cluster & Xeon Phi (& GPU)
- 4Q 2014 (Ver.0.3.0)
  - Prototype of ppOpen-HPC for Post-Peta Scale System
- 4Q 2015
  - Final version of ppOpen-HPC for Post-Peta Scale System
  - Further optimization on the target system

ppOpen-HPC Ver.0.3.0

http://ppopenhpc.cc.u-tokyo.ac.jp/

- Released at SC14 (or can be downloaded)
- Prototype version (Flat MPI, OpenMP/MPI Hybrid) with documents in English
- Collaborations with scientists

Component	Archive	Flat MPI	OpenMP/ MPI	Fotran	C
ppOpen-APPL/FDM	ppohFDM_0.2.0	○	○	○	×
ppOpen-APPL/FDM-AT	ppohFDM_AT_0.2.0	○	○	○	×
ppOpen-APPL/AMR-FDM	ppohAMRFDM_0.3.0	○	○	○	×
ppOpen-APPL/FVM	ppohFVM_0.3.0	○	○	○	×
ppOpen-APPL/FEM	ppohFEM_0.3.0	○	○	○	○
ppOpen-APPL/BEM	ppohBEM_0.3.0	○	○	○	×
ppOpen-APPL/BEM-AT	ppohBEM_AT_0.1.0	○	○	○	×
ppOpen-APPL/DEM-util	ppohDEM_util_0.3.0	○	○	○	×
ppOpen-MATH/VIS	ppohVIS_FDM3D_0.2.0	○	×	○	○
ppOpen-MATH/MP-PP	ppohMATHMPPP_0.2.0	○	×	○	×
ppOpen-AT/STATIC	ppohAT_0.2.0	-	-	○	○

普及活動(1/2)

- ppOpen-AT関連共同研究
  - 工学院大学　田中研究室
    - 田中研究室開発のAT方式(d-spline方式)の適用対象としてppOpen-ATのAT機能を拡張
  - 東京大学　須田研究室
    - 電力最適化のため，須田研究室で開発中のAT方式と電力測定の共通APIを利用し，ppOpen-ATを用いた電力最適化方式を提案
- JHPCN共同研究課題
  - 高精度行列-行列積アルゴリズムにおける並列化手法の開発（東大，早稲田大）（H24年度）（研究としては継続）
    - 高精度行列-行列積演算における行列-行列積の実装方式選択に利用
  - 粉体解析アルゴリズムの並列化に関する研究（東大，法政大）（H25年度）
    - 粉体シミュレーションのための高速化手法で現れる性能パラメタのATで利用を検討

普及活動(2/2)

- JHPCN共同研究課題(続き)
  - 巨大地震発生サイクルシミュレーションの高度化(京大，東大他)(H24・25年度)
    - Hマトリクス，領域細分化
  - ポストペタスケールシステムを目指した二酸化炭素地中貯留シミュレーション技術の研究開発(大成建設，東大)(H25年度)
    - 疎行列ソルバー，並列可視化
  - 太陽磁気活動の大規模シミュレーション」(東大(地球惑星，情報基盤センター))(H25年度)
    - 疎行列ソルバー，並列可視化
- 講習会，講義
  - ppOpen-HPCの講習会を2014年3月から実施
  - 講義，講習会(並列有限要素法)でppOpen-MATH/VISを使用して可視化を実施する予定

ポストペタからエクサスケールへ

- ポストT2Kを念頭に置き，Intel Xeon/Phiなどメニィコアアーキテクチャ向け最適化を中心に研究開発を進める
- 「ポストペタスケール」から「エクサスケール」へ
  - エクサスケールシステムの詳細が具体的になりつつある
    - スーパーコンピュータシステムはより複雑化，巨大化
    - アプリケーション実行性能を引き出すためのプログラミングはより困難に
  - ppOpen-HPCのような環境の必要性はより高まる
  - エクサスケールシステム開発の動向も念頭に置いた開発を継続して実施して行くことで，エクサスケールシステムの利用に当たってもスムーズな移行を支援できると考えられる。
- エクサスケールシステムを念頭においた研究開発
  - 電力最適化
  - Communication/Synchronization Reducing Algorithms（通信・同期削減アルゴリズム）