

第3回計算科学技術推進WG
平成16年10月27日

資料3 - 3



将来(2010年前後を想定)の研究目標と スーパーコンピューティング環境について

東京大学 人工物工学研究センター
奥田 洋司

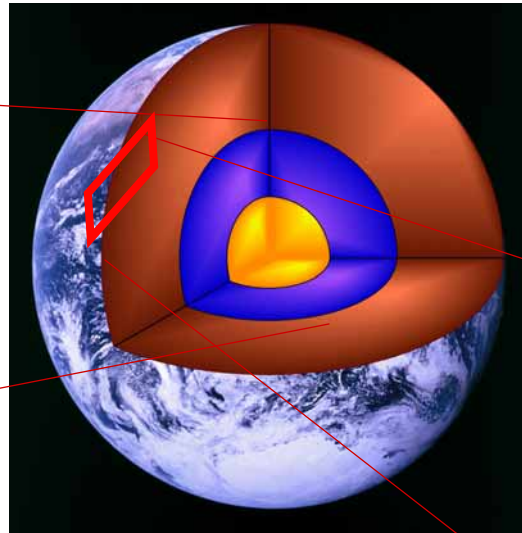
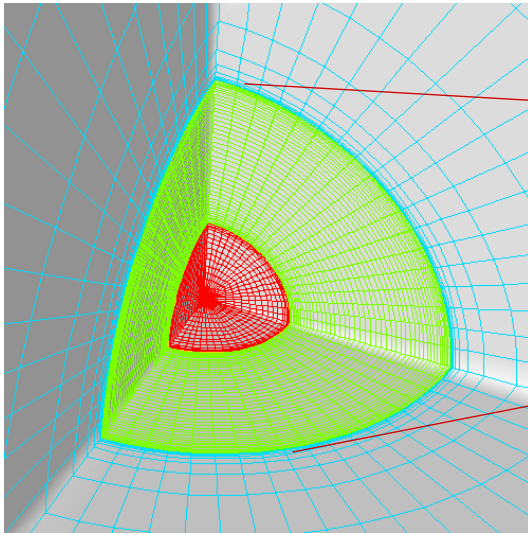
目次

- 現行のスーパーコンピュータシステム及び研究成果について
 - GeoFEM開発の経験から
 - グリッドコンピューティングの可能性
- 将来(2010年前後)の研究目標と期待される成果について
- 将来(2010年前後)のスーパーコンピュータシステムについて

GeoFEM開発の経験から

- 科学技術振興調整費総合研究(平成10～14年度)
「高精度の地球変動予測のための並列ソフトウェア開発」の一部
- GeoFEM: 固体地球分野を対象とした, 並列有限要素法解析システム
- 有限要素解析, 可視化・情報処理系, 特定の固体地球分野, という, 異なるバックグラウンドを有する3者の協力
- 地球シミュレータ共同プロジェクト, 固体地球分野
「固体地球シミュレーションプラットフォームの開発」

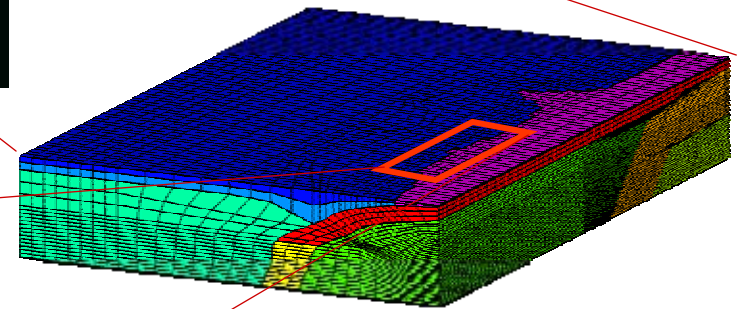
固体地球におけるマルチスケール・ マルチフィジクス



マルチフィジクス

Mantle-Core Dynamics & Plumes

Target : 10^7 nodes ($\Delta h = 10$ km order)



Crustal movements & tectonic deformation

Target : 10^9 nodes ($\Delta h = 1$ km)



Seismic wave generation & propagation

Target : 10^{11} nodes ($\Delta h = 20$ m)

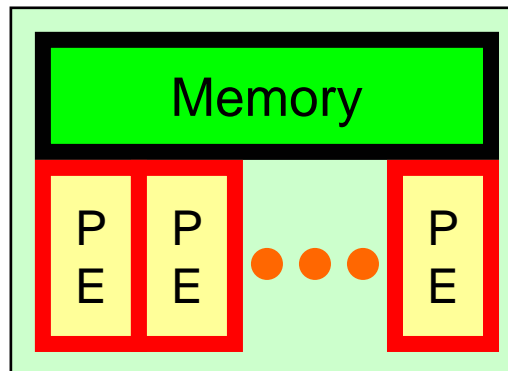
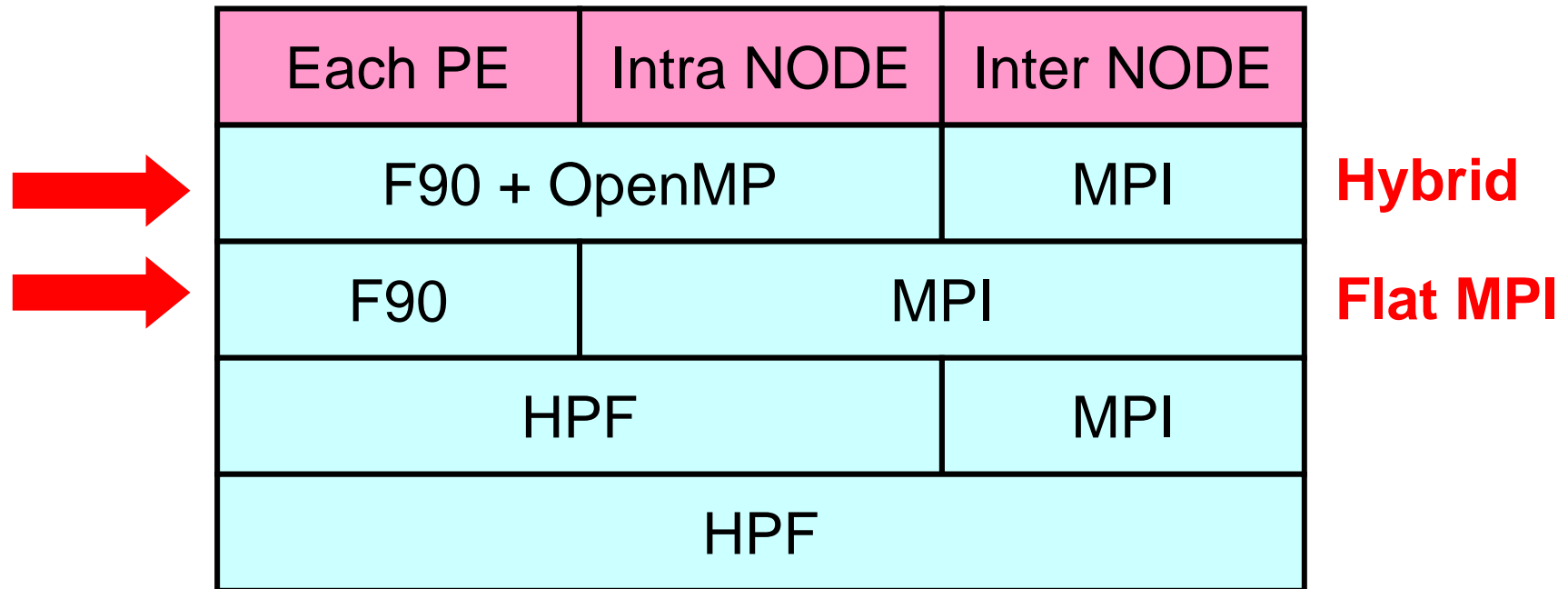
■ GeoFEM開発に用いた計算機環境

- 当時ESは設計中(大まかなスペック情報のみ)
 - PCクラスタ / MPP / SMPクラスタ, 最終ターゲットはES
 - 並列性能とベクトル性能
-
- Alphaクラスタ(東大, RIST)
 - 開発, デバッグ, 小規模ラン
 - SR2201(東大, 原研CCSE)
 - 1,024PEまでの大規模並列性能, 擬似ベクトル
 - SX4(東北大, 原研CCSE)
 - 単体ベクトル性能評価
 - SR8000(東大)
 - ノード間 / ノード内ハイブリッド並列, 擬似ベクトル, 128ノード(1,024PE)

最終的にはSR8000でES用を準備

途中, 適宜ES上での性能を予測(妥当性を後に確認)

Programming Models for Earth Simulator



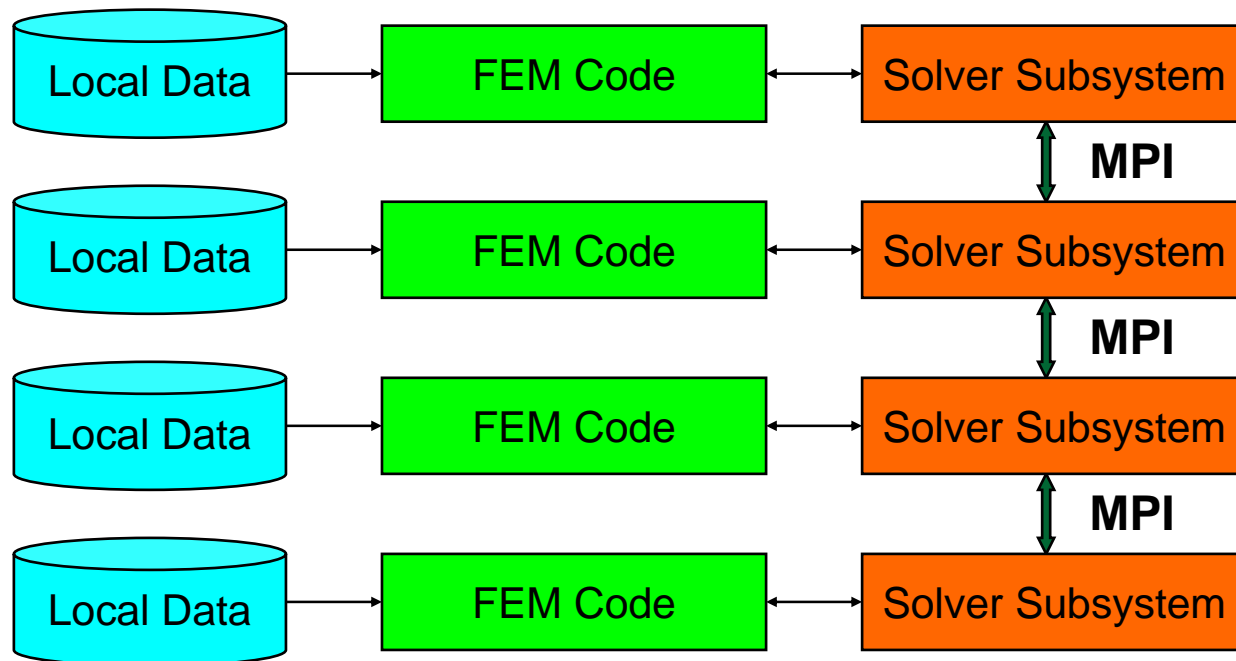
Hybrid Parallel / Vector

- Fortran90 (可視化部はC)
- Inter-node : MPI
 - 領域分割に基づくSPMD (Single-Program Multiple-Data) 並列
- Intra-node : OpenMP (可視化部はPthread)
 - オーダリング (演算の依存性をなくす)
 - 外側ループに対してOpenMP
- Each PE : Vectorization
 - オーダリング (演算の依存性をなくす)
 - 最内側のループを最も長くしてベクトル化

- 並列化(データの局所化, SPMD, 反復法ソルバの採用)
 - ロバストな反復法ソルバ
 - 並列計算を行った場合でも1領域での逐次演算となるべく等価なアルゴリズム(前処理を含む)
 - 領域分割(データシェアリング, ワークシェアリング)の解析者(開発者)からの隠蔽
 - 入力データに領域間隣接情報を含め, 通信はソルバのみで考慮する
 - 「FEMは本質的に並列向き」なる性質を堅持

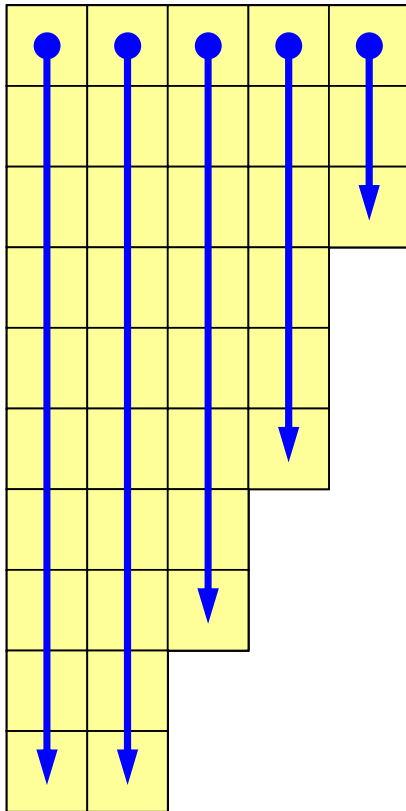
SPMD Programming Style

- Large file handling Local distributed data
- Fluid/structure analysis modules just consider local operation (element matrix assemble)
- Global operation occurs only in linear solver.

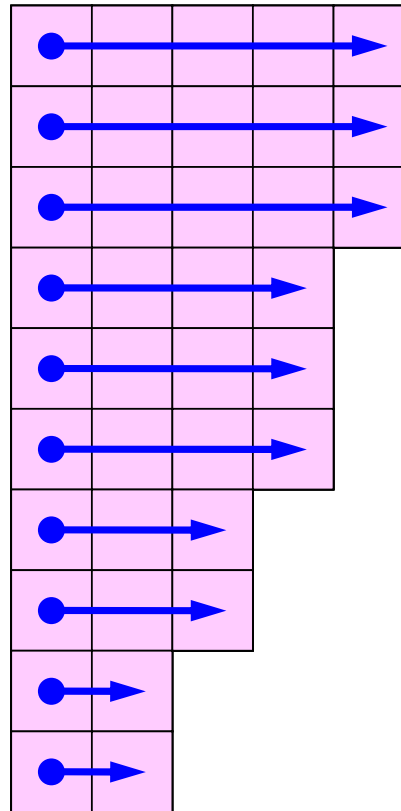


SMPクラスタ型ベクトル計算機用 オーダリング

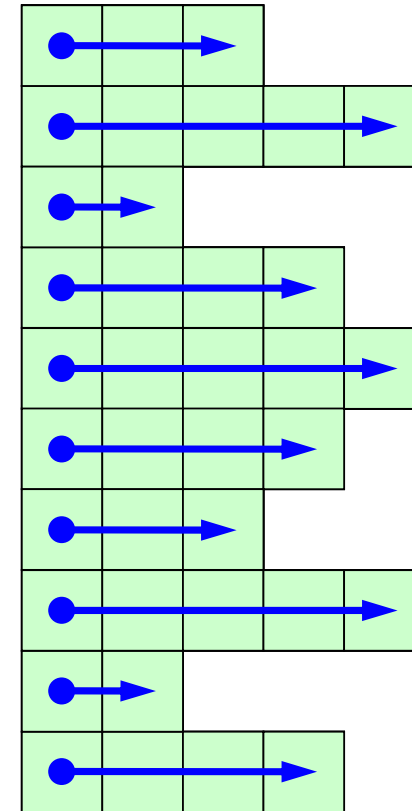
PDJDS/CM-RCM



PDCRS/CM-RCM
short innermost loop

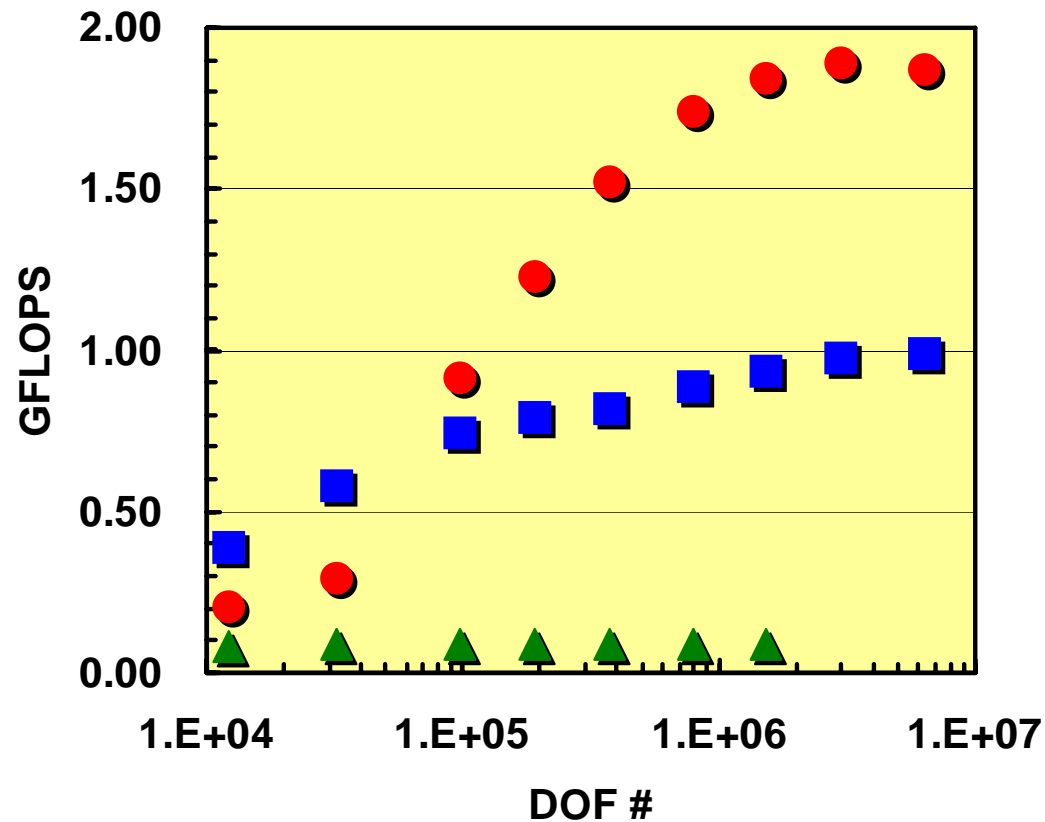


CRS no re-ordering



3次元弾性問題計算結果 (問題規模とGFLOPS値の比較)

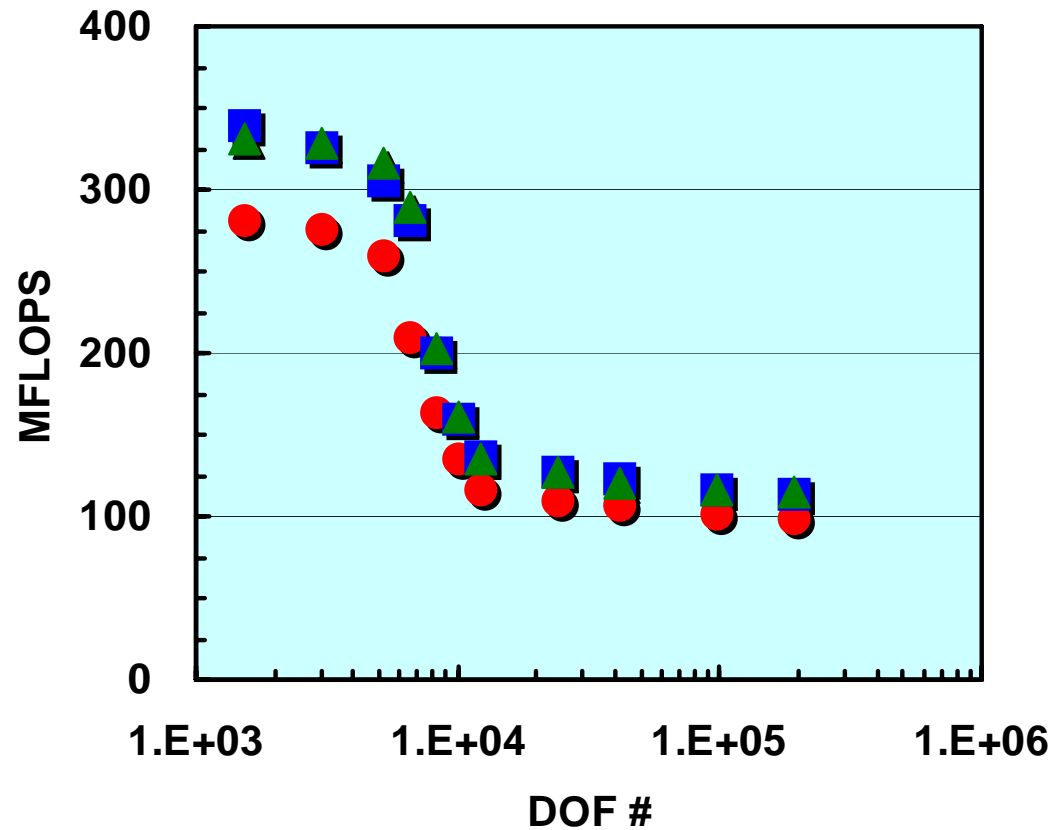
Hitachi-SR8000/SMP, Pseudo Vector



:PDJDS/CM-RCM, :PCRS/CM-RCM, :Natural Ordering


3次元弾性問題計算結果 (問題規模とMFLOPS値の比較)

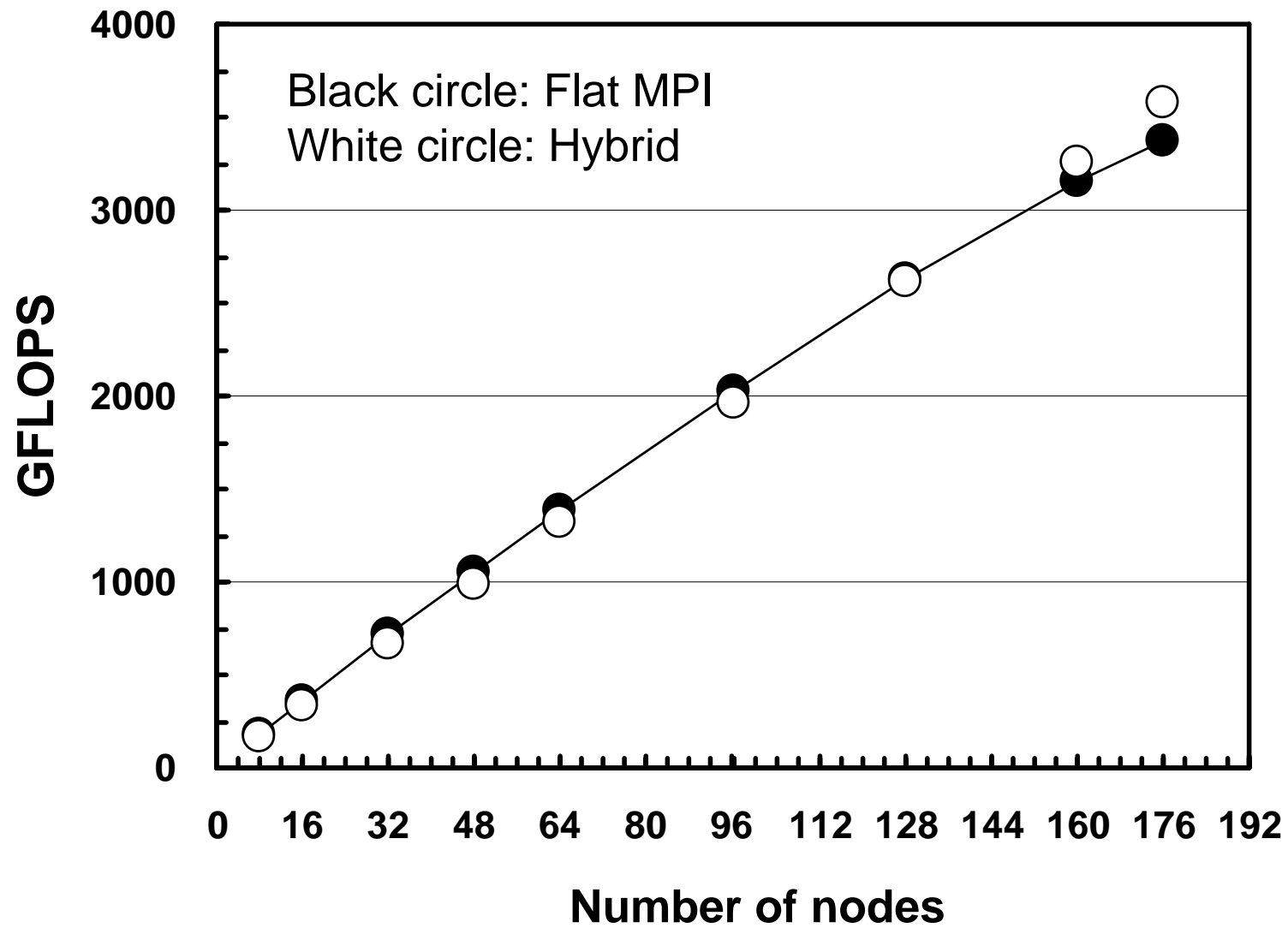
Compaq Alpha 21164, 599MHz



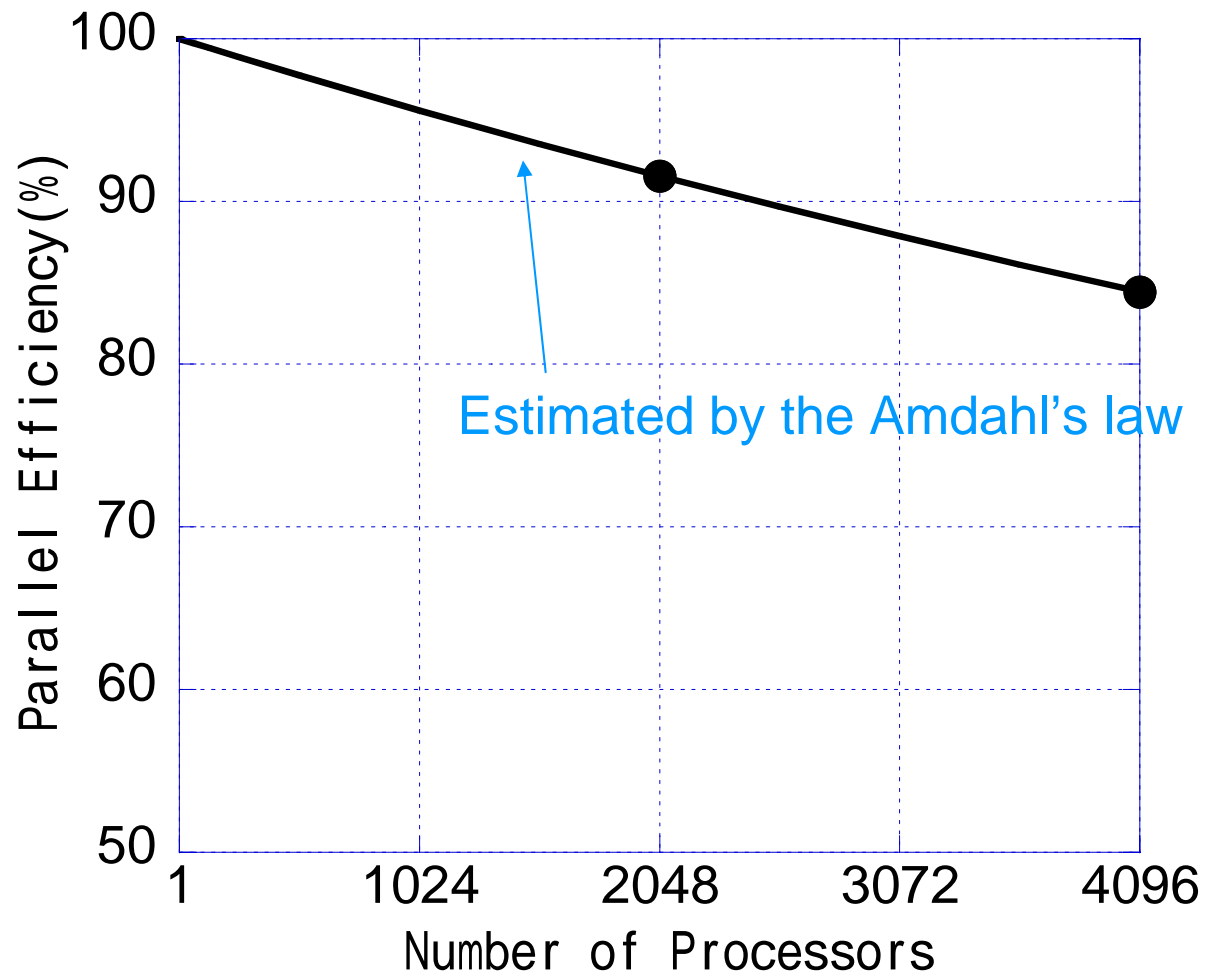
: PDJDS/CM-RCM, : PCRS/CM-RCM, : Natural Ordering

Elastic Analysis Benchmark

- Computational speed (Fixed problem size per node)
 - 3.8 TFLOPS (176 nodes , 33.7 % to peak, 2.2 GDOFs, work ratio 95%)
 - 10.4 TFLOPS (512 nodes , 31.8 % to peak, 6.4 GDOFs)
 - Parallel efficiency
 - Compute 0.8 GDOF problem using 256 nodes and 512 nodes, and estimate the parallel efficiency by the Amdahl's law
 - Parallel efficiency = 84% when using 512 nodes (parallelization ratio = 0.999955)
- max. number of nodes for users
- 



GFLOPS rate for 3D linear elastic problem



Parallel efficiency for 3D linear elastic problem
(estimated with the Amdahl's law)

Comparison with SR8000 GFLOPS rate

SMP node#	DOF	SR8000/mpp	Earth Simulator
8	50,331,648	21.3 GFLOPS (18.5%)	169.4 GFLOPS (33.1%)
16	100,663,296	42.4 GFLOPS (18.4%)	335.9 GFLOPS (32.8%)
128	805,306,368	335.2 GFLOPS (18.2%)	2611.8 GFLOPS (31.9%)



~ 8 times faster

	SR8000/mpp	E S
プロセッサ単体性能	1.8 GFLOPS	8 GFLOPS
ノード性能	14.4 GFLOPS	64 GFLOPS
ノードあたりメモリ	16 GB	16 GB
プロセッサ数	1,024	5,120
ノード数	128	640
ピーク性能	1.8 TFLOPS	40 TFLOPS

参考データ

ESでの計算性能予測 (昔のスライド)

Estimation of Computational Speed of GeoFEM on the Earth Simulator

40 Tflops (peak speed of ES)

x 0.801 (Inter-node parallel performance for 640 nodes, **estimated**)

x 0.875 (Intra-node parallel performance for 8 PEs, **measured**)

x 0.40 (1PE vector performance, **measured**)

$$0.801 \times 0.875 \times 0.4 = 0.28$$

= **11.2 Tflops**

実測 10.4 TFLOPS (512 nodes)

Grid Computing Environments

- Test Bed on ApGrid



- PC-clusters on WAN, $\sim 10^0$ - 10^1 km , ~ 10 hops

- F32** (Xeon, AIST/GTRC) ← CA

- OGT** (Alpha, UT)

- Skyraiders** (P4, RIST)

- Test Bed on SuperSINET

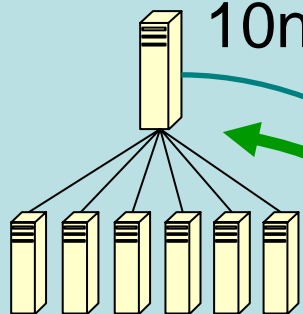
- Supercomputers on fast network , $\sim 10^2$ km

- SR8000-compact (UT) ← CA

- SR8000-compact, Onyx300 (Hokkaido Univ.)

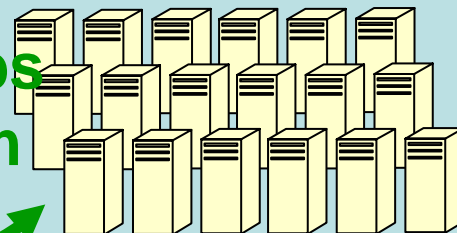
- PrimePower (Kyushu Univ.)

'OGT' Alpha21269, 667MHz
10nodes, 100baseTX



UT, Tokyo

'F32' Xeon, 3.0GHz
2x64nodes
Giga bit Ether



AIST, Tsukuba

12 hops
~50 km

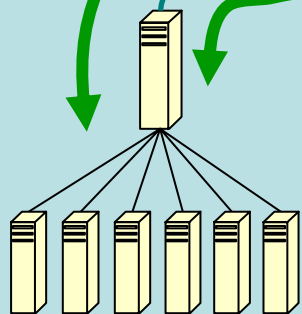


Internet

10 hops
~5 km

9 hops
~50 km

'Skyraiders'
P4, 2.5GHz
2x24nodes
Myrinet



RIST, Tokyo

traceroute from OGT to F32

1	0.535 ms
2	0.534 ms
3	0.384 ms
4	1.142 ms
5	0.813 ms
6	0.897 ms
7	2.251 ms
8	3.133 ms
9	2.997 ms
10	2.971 ms
11	3.516 ms
12	3.747 ms
13	2.901 ms

OS, Grid Middleware

- OS
 - Redhat Linux 7.3 (F32, SKR)
 - Kondara MNU/Linux (OGT)
- Grid Middleware
 - Globus 2.2.2 MPICH does not catch up with ver 3.
 - MPICH-G2 (mpich-1.2.4)

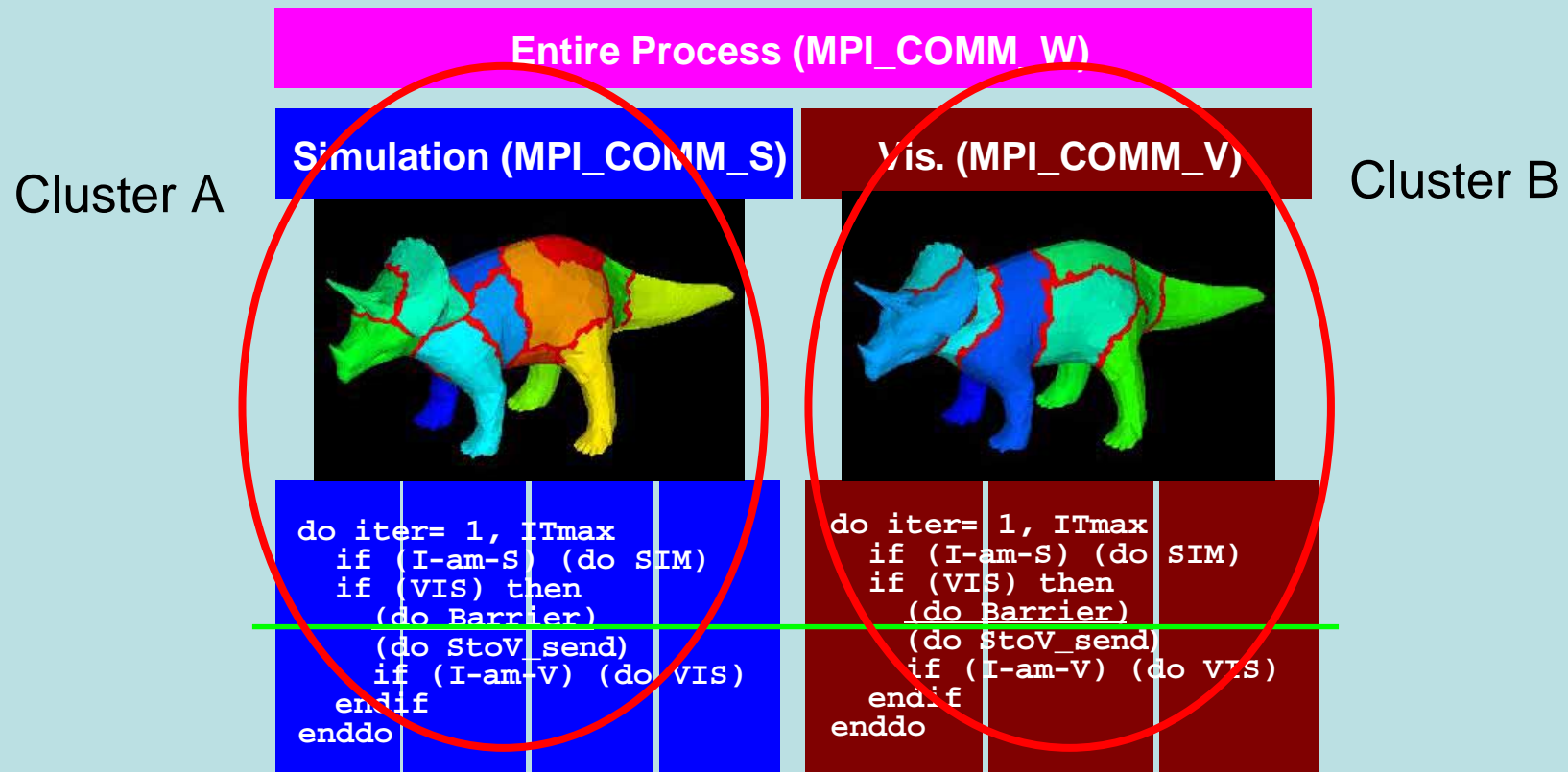
Note :

Currently, to run MPI programs under the Globus service, all the hosts need to have [public IP addresses](#).

Applications considered in this study

- Target
 - Parallel finite element analysis (FEA)
 - Contains frequent communications
- Two Types of Applications
 - Tightly connected application
 - Frequent communications among clusters
 - i.e. **FEA for one domain**
 - Loosely connected application
 - Infrequent communications among clusters
 - e.g. Fluid-Structure Interaction
 - Steering/Tracking (Vis. and FEA)**

Steering (FEA and Visualization) on Cluster-of-Clusters



Loosely connected SPMD programs

Inside each program : Frequent communication

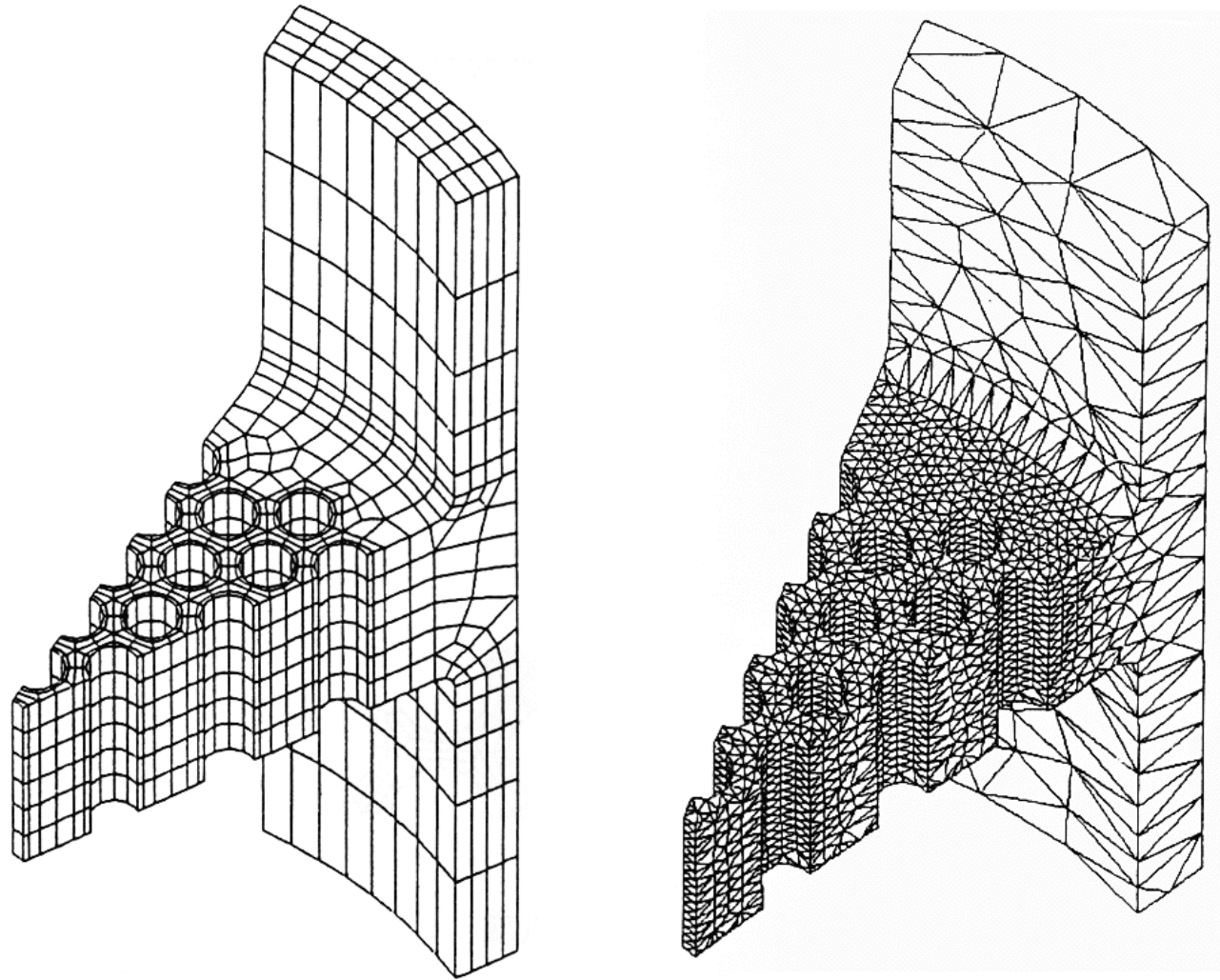
In between the two : Once every several time steps

目次

- 現行のスーパーコンピュータシステム及び研究成果について
- 将来(2010年前後)の研究目標と期待される成果について
 - HPC-MW
 - HLW処分シミュレータ(地球シミュレータプロジェクト 原子力分野)
- 将来(2010年前後)のスーパーコンピュータシステムについて

大規模解析の恩恵と課題

- 事前誤差 (モデリング段階の誤差) の低減化
 - 「対称性の仮定」から「まるごと解析」へ
 - 良質なメッシュ
- モデリング, 解析の容易さ
 - しかしながら:
 - ・ プレプロセッサ能力との乖離
 - ・ メッシュの階層性の考慮、アダプティブ解析がより重要に
- 最適化の必要性 vs 日進月歩の計算機環境
 - 「大規模解析」よりも「速いスループット」を重視という声

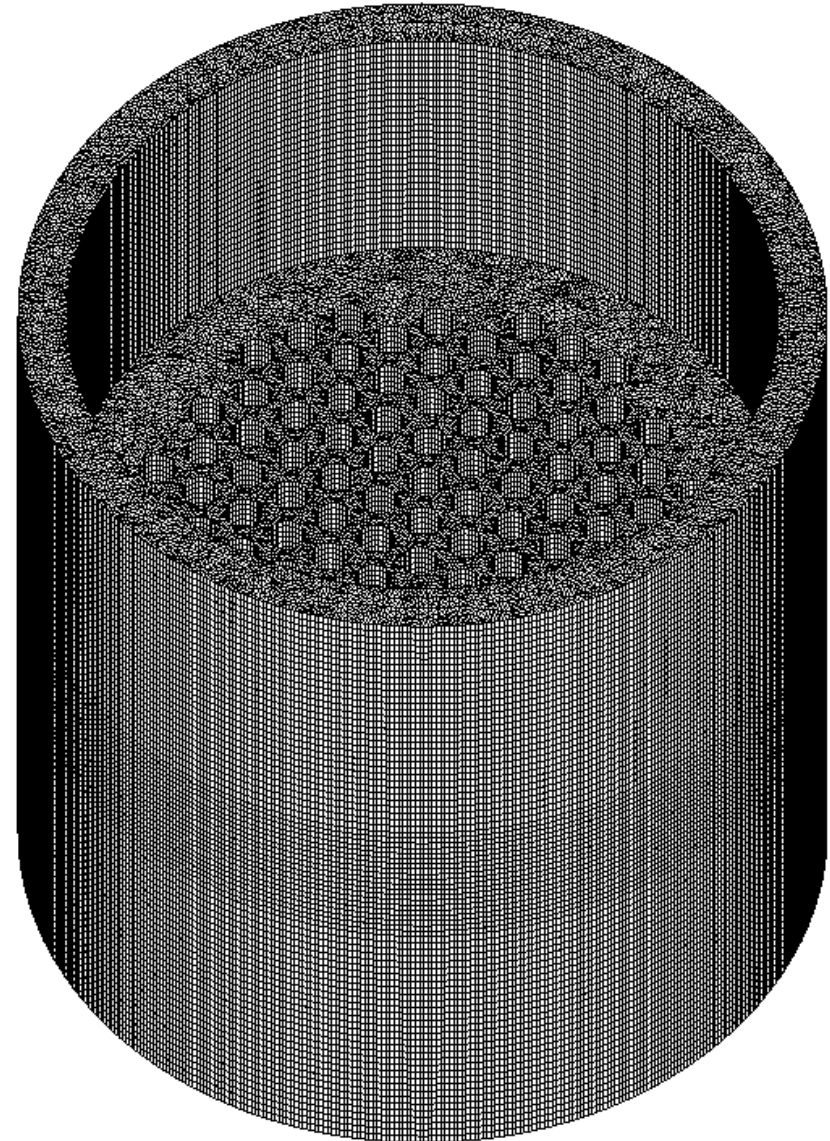
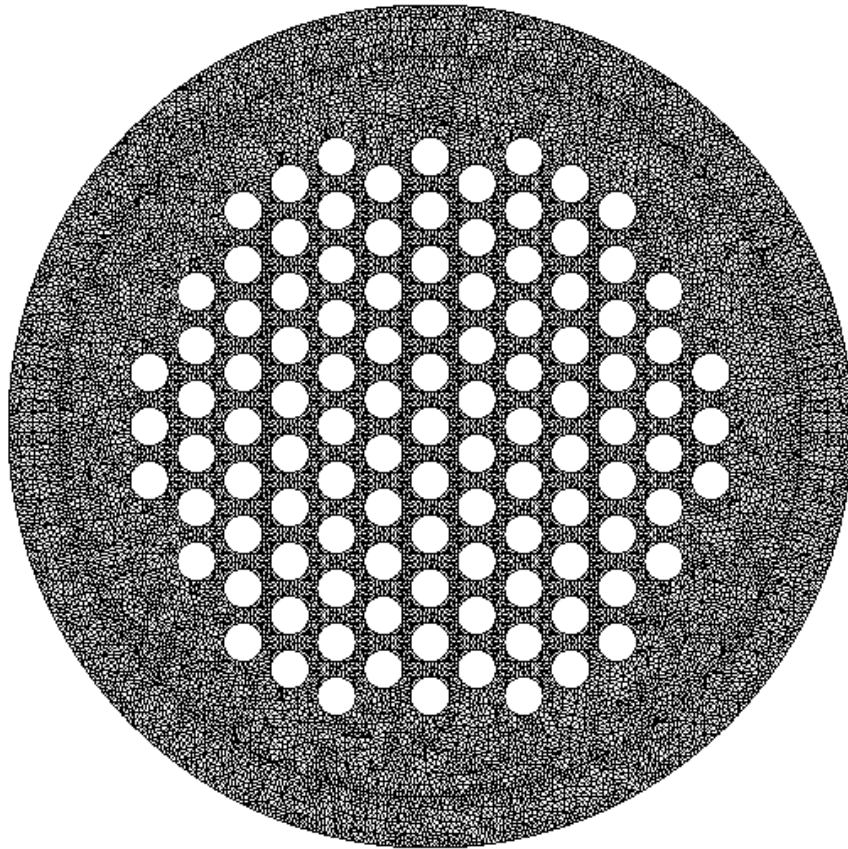


Examples of conventional FE modeling
(30° symmetry)

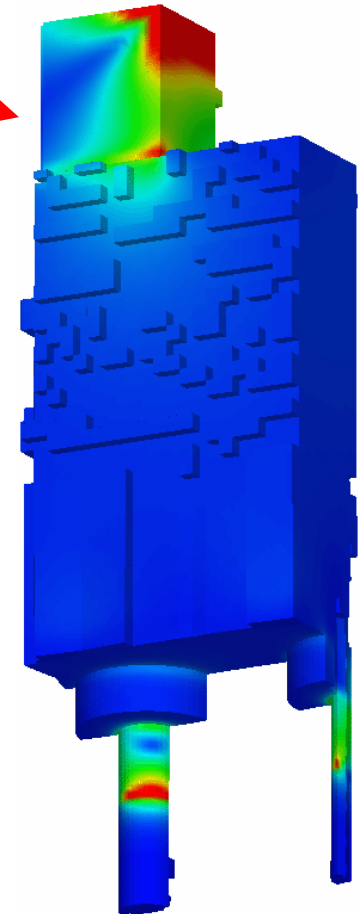
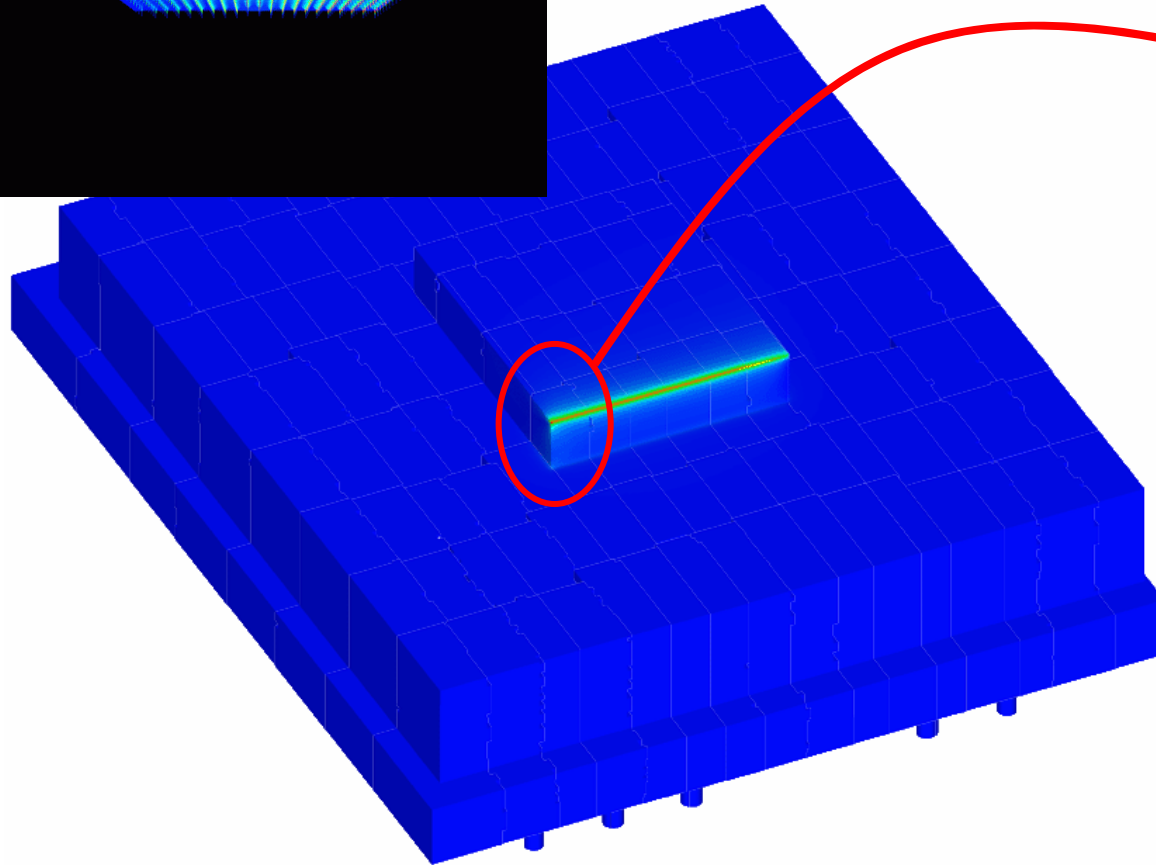
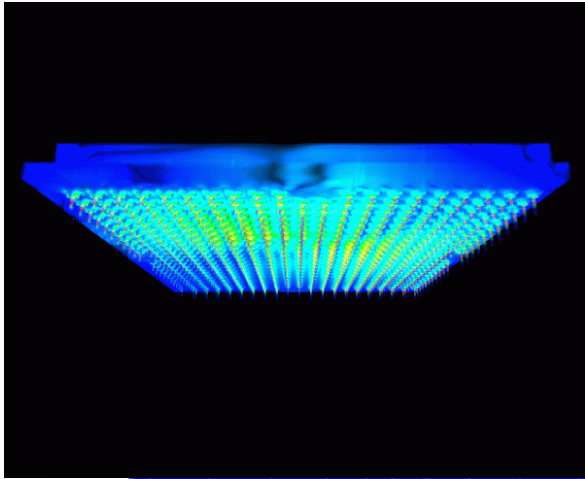
Tubesheet : Large_Model

1,053,906 nodes

949,512 elements



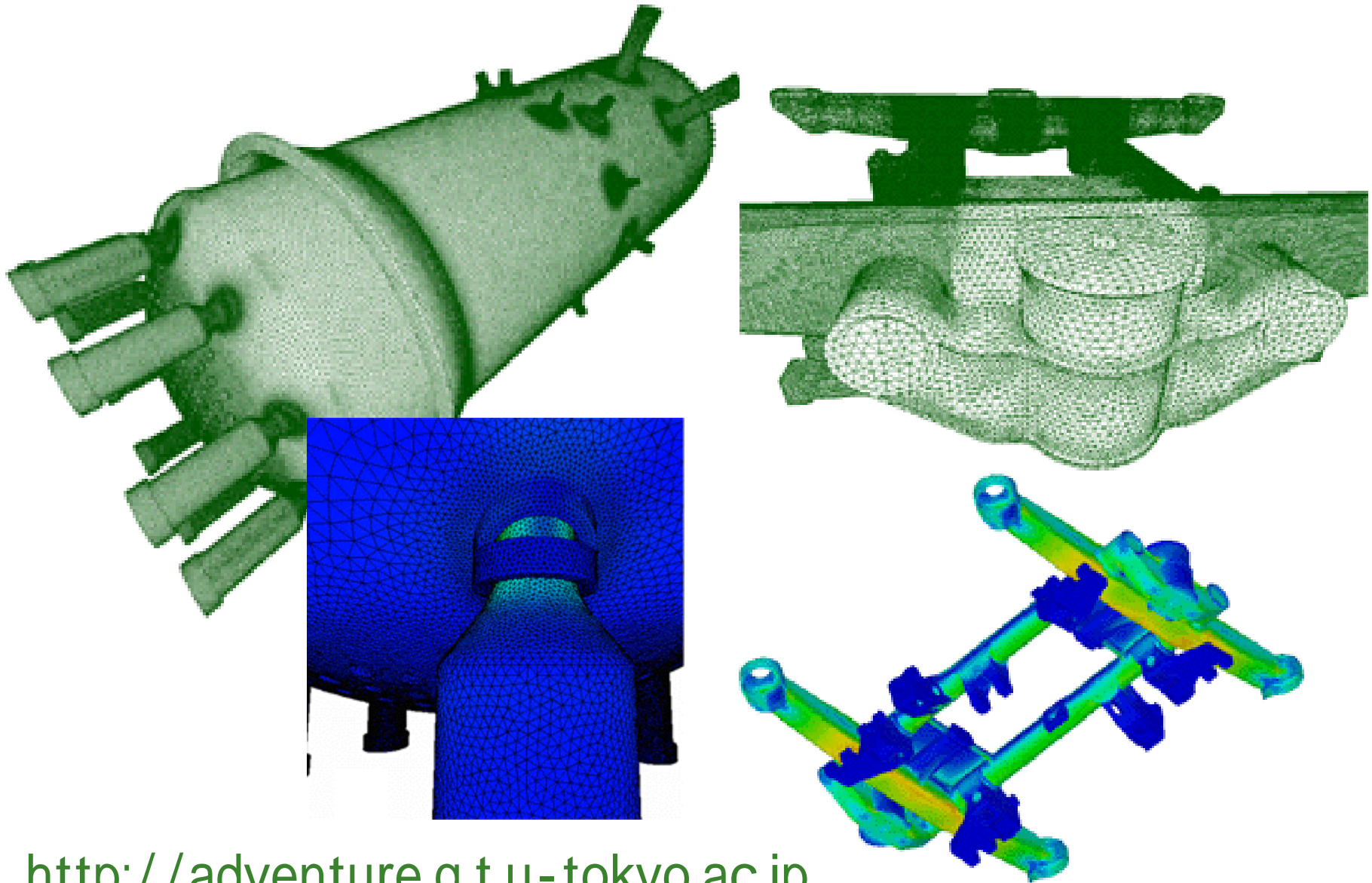
495-pin Micro-PGA package



- 7.8 M nodes, 7.6 M elements
- Mises stress

Sub-domain
(PE #76)

ADVENTUREの解析事例より



<http://adventure.q.t.u-tokyo.ac.jp>

ハイエンド構造解析に必要な ブレイクスルー

- 「線形・単一事象」から「非線形・連成・最適化」へ
 - 解析対象に応じたアルゴリズムの選択
 - PSEツールの重要性
- 「大規模丸ごと解析」から「アセンブリ構造解析」へ
 - 境界非線形
 - モデリングツールとの統合化がますます重要に
- 「汎用コード」vs「専用コード」
- 「コードの最適化」vs「専用計算機」

→ HPC-MWによる機能
の多チャンネル化へ

最適化の重要性と負担

- 様々なHPC環境
 - PCクラスタ / 分散メモリ型超並列計算機 / SMPクラスタ (8-way, 16-way, 256-way)
 - Power, HP-RISC, Alpha/Itanium, Pentium, Vector PE
 - Grid環境
- 最適化 (単体CPU, 並列) は非常に重要
 - Grid環境での移植性 (portability)
 - H/Wに応じた最適化が必要
 - ただ動く・・・ことは可能
 - アルゴリズム, データ構造の変更が必要な場合もある。
- 重要だがアプリケーション開発者にとって大きな負担

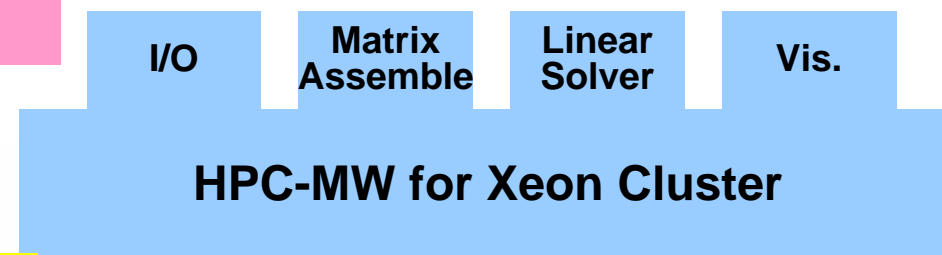
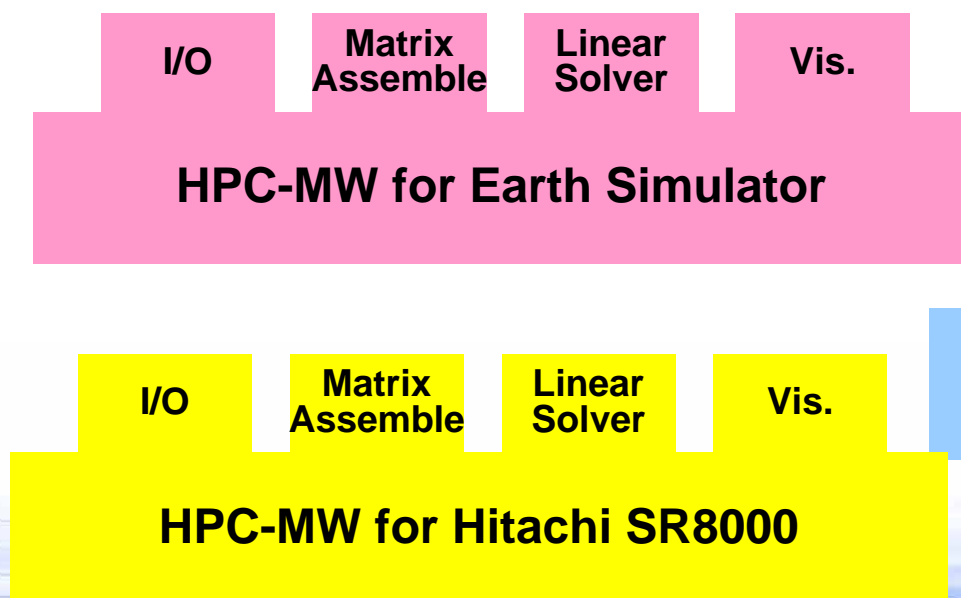
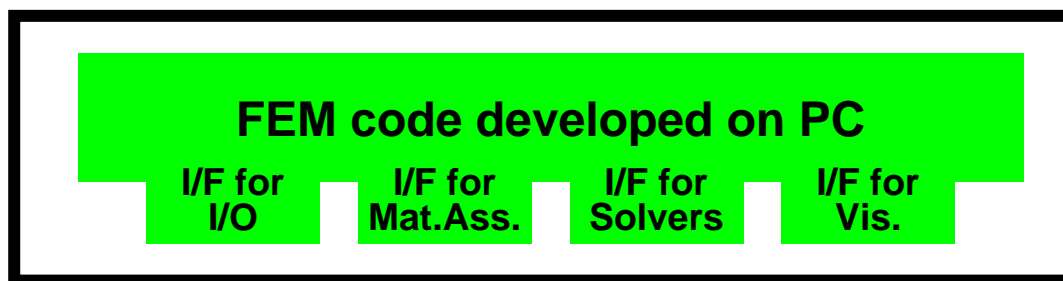
戦略的基盤ソフトウェアの開発

<http://www.fsis.iis.u-tokyo.ac.jp/>

- 文部科学省「ITプロジェクト」の一部
- 東京大学生産技術研究所 計算科学技術連携研究センター（平成14年度～18年度）（平成17年度より体制変更予定）
- 実用ソフトの開発
 - 単なる研究開発ではない
 - ソフトウェアの公開, 商用化
- 7サブプロジェクト
 - 量子化学, タンパク質, ナノテクノロジー
 - 次世代流体解析, 構造解析
 - PSE, HPC-MW

HPC Middlewareの利用イメージ

各ライブラリ (HPC-MW) に対して
同じインタフェースを使える



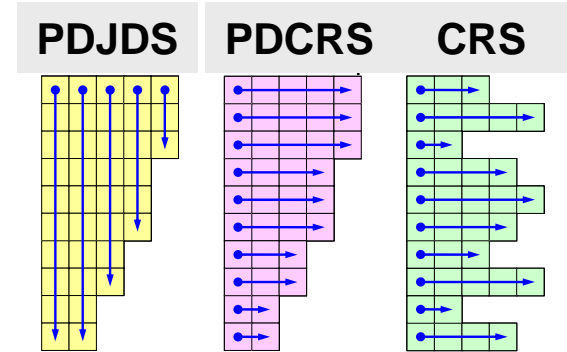


ライブラリ型HPC-MW のサポート機能

- データ入出力
- 適応格子
- 動的負荷分散 (+pMETIS)
- 並列可視化
- 線形ソルバー (反復法, 直接法)
- 有限要素処理 (コネクティビティ処理, 係数行列生成)
- カップリング
- メッシュ関連Utility

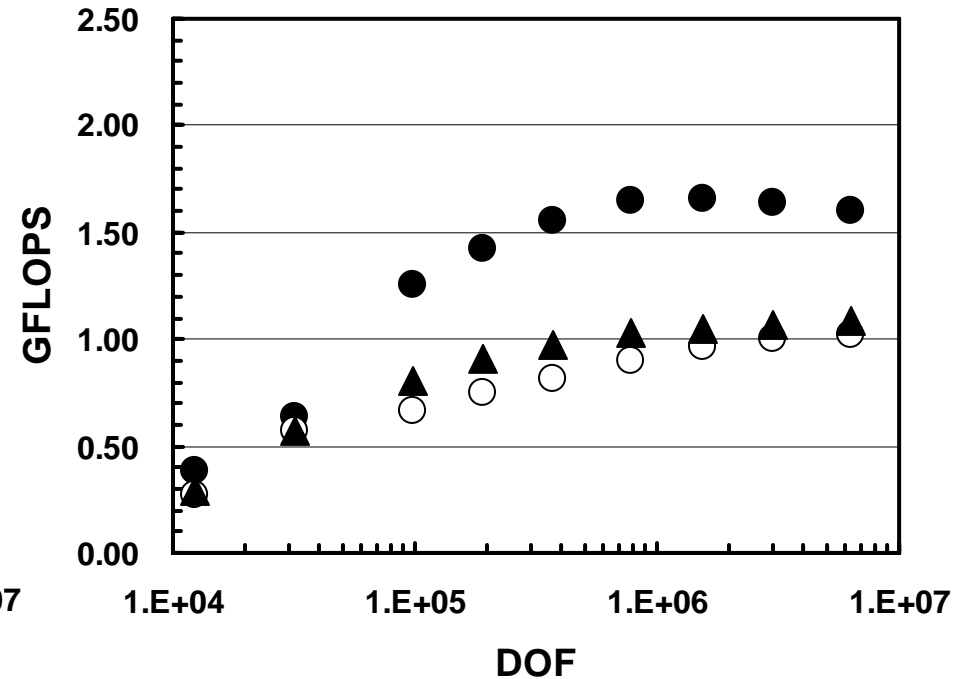
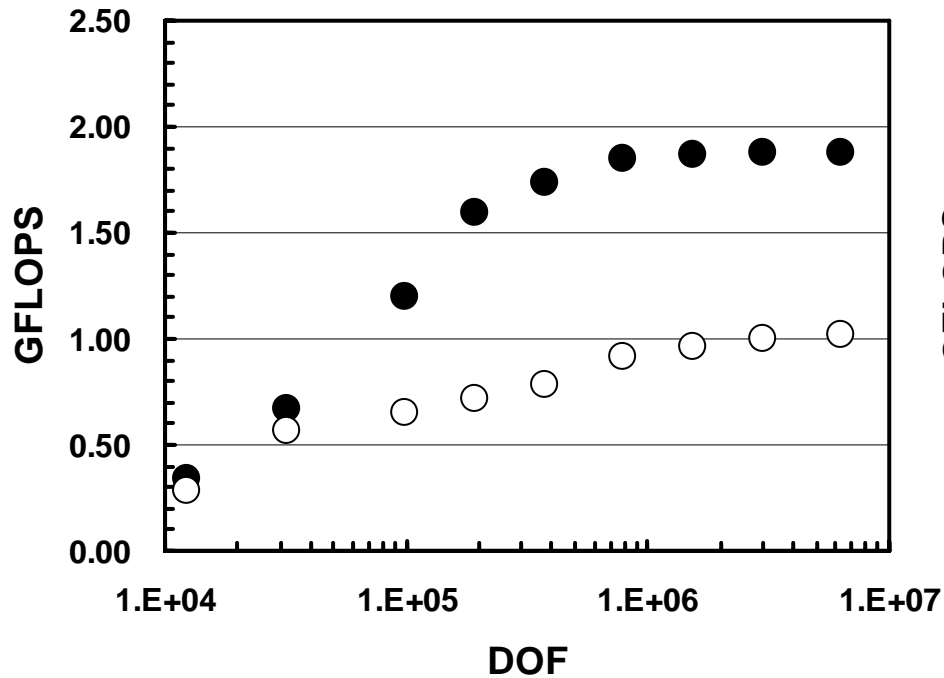
Hitachi SR8000/128

8 PEs/1-SMP node



SMP

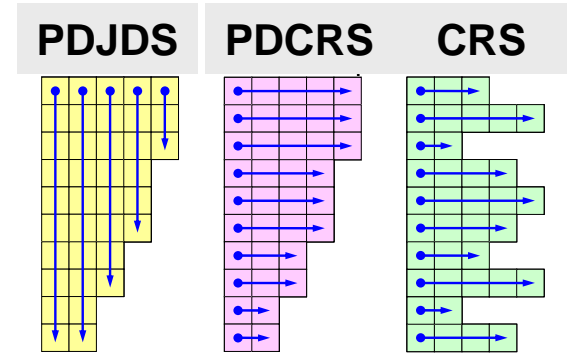
Flat-MPI



: PDJDS, : PDCRS, : CRS-Natural

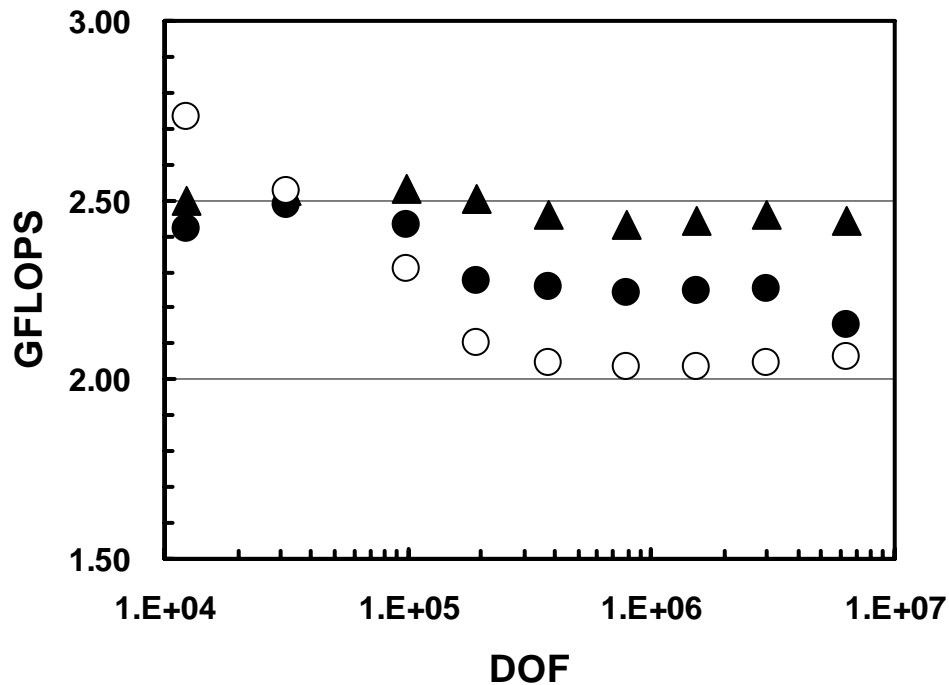
Xeon & SR2201

8 PEs

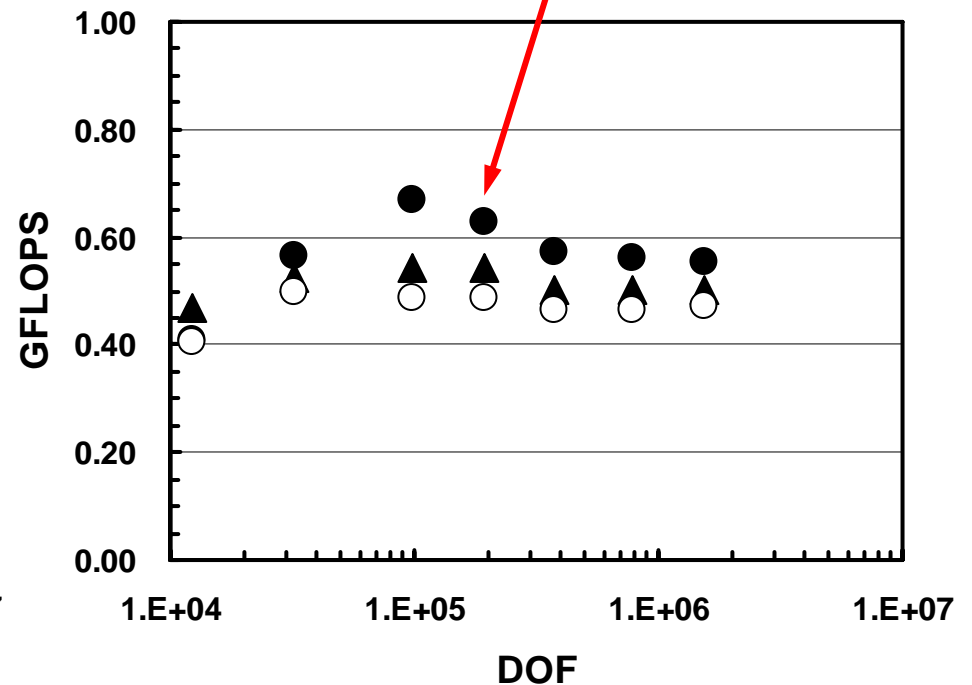


: PDJDS, : PDCRS, : CRS-Natural

Xeon



SR2201

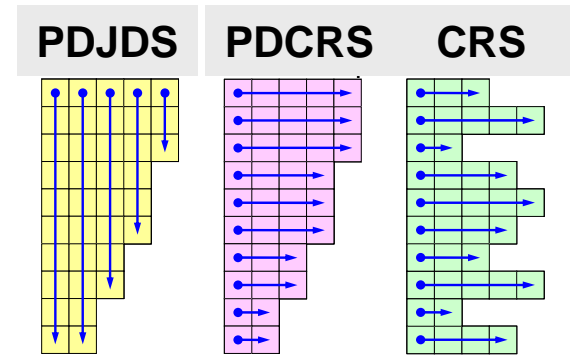


Not so significant deterioration due to pseudo-vector.



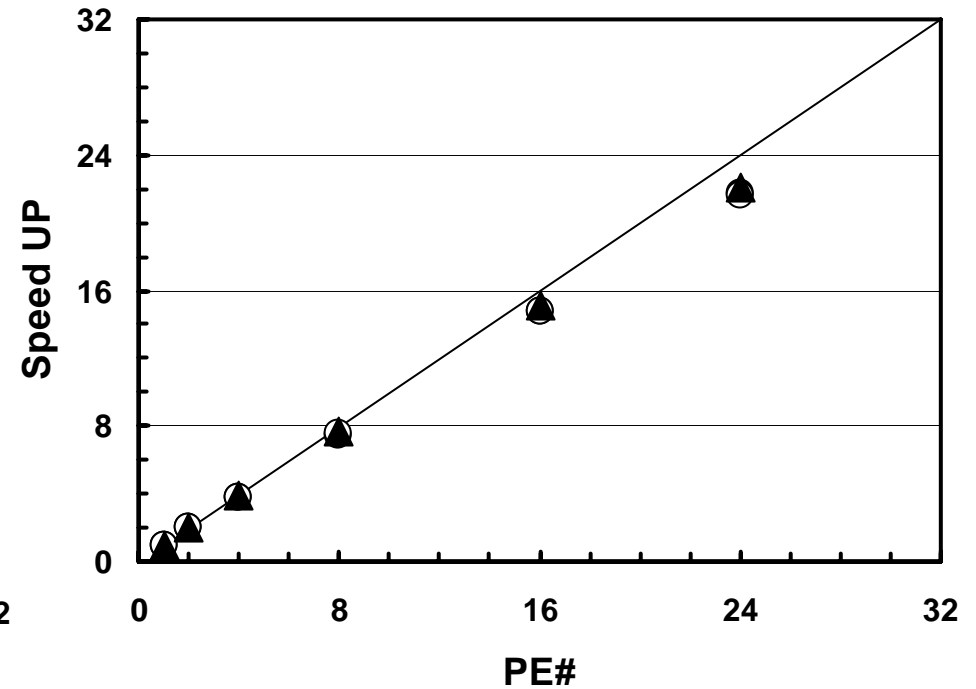
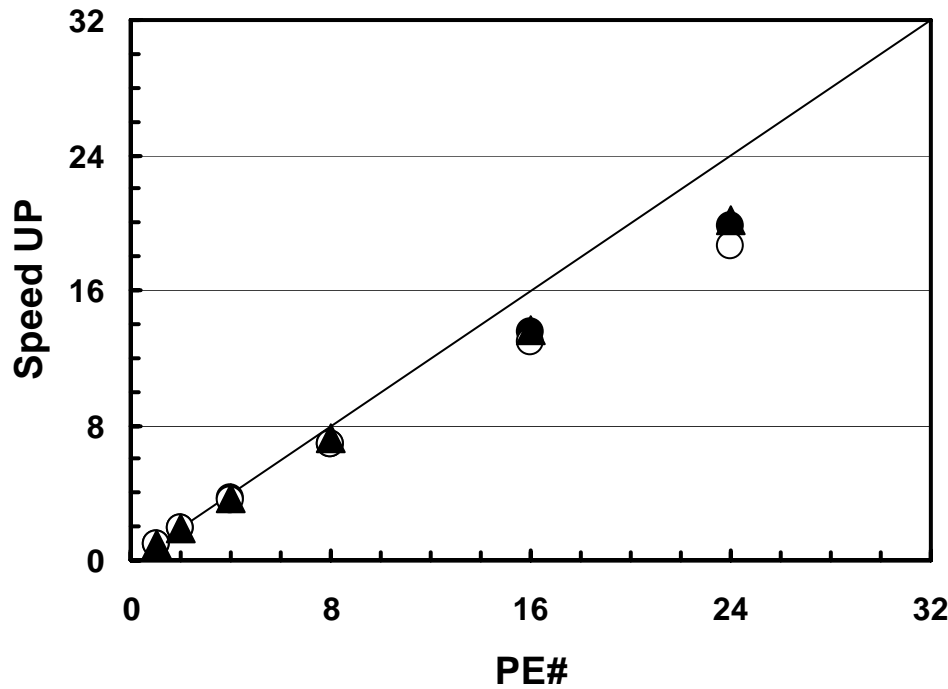
Xeon: Speed UP

1-24 PEs



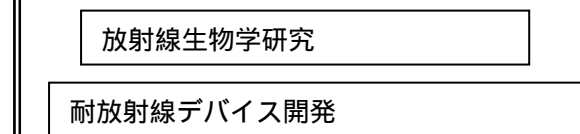
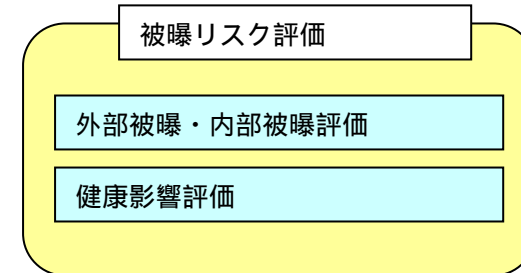
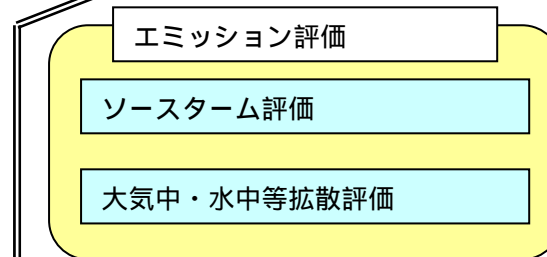
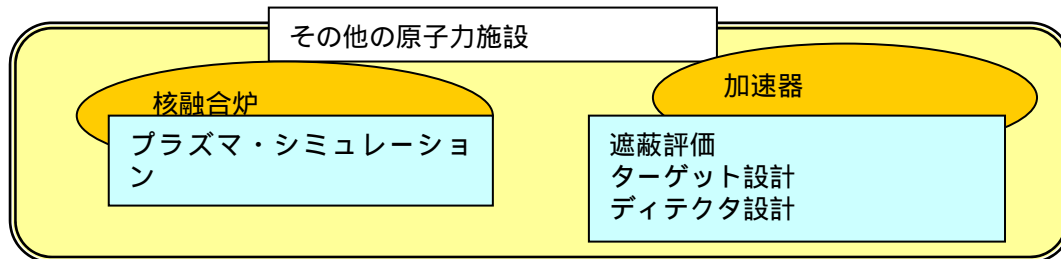
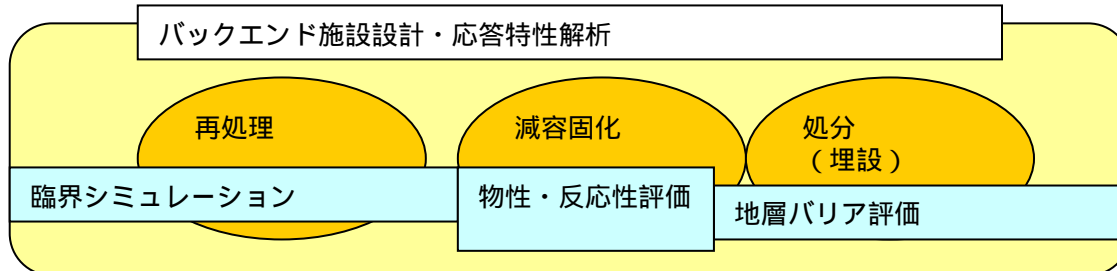
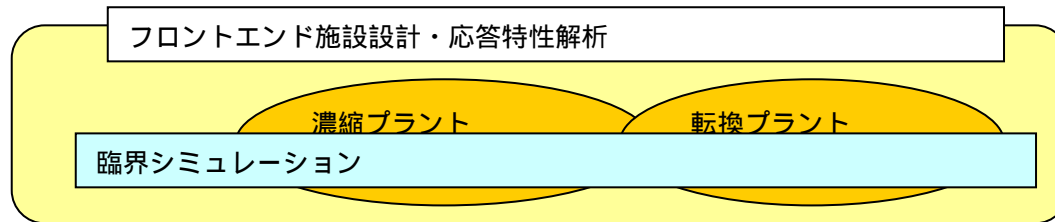
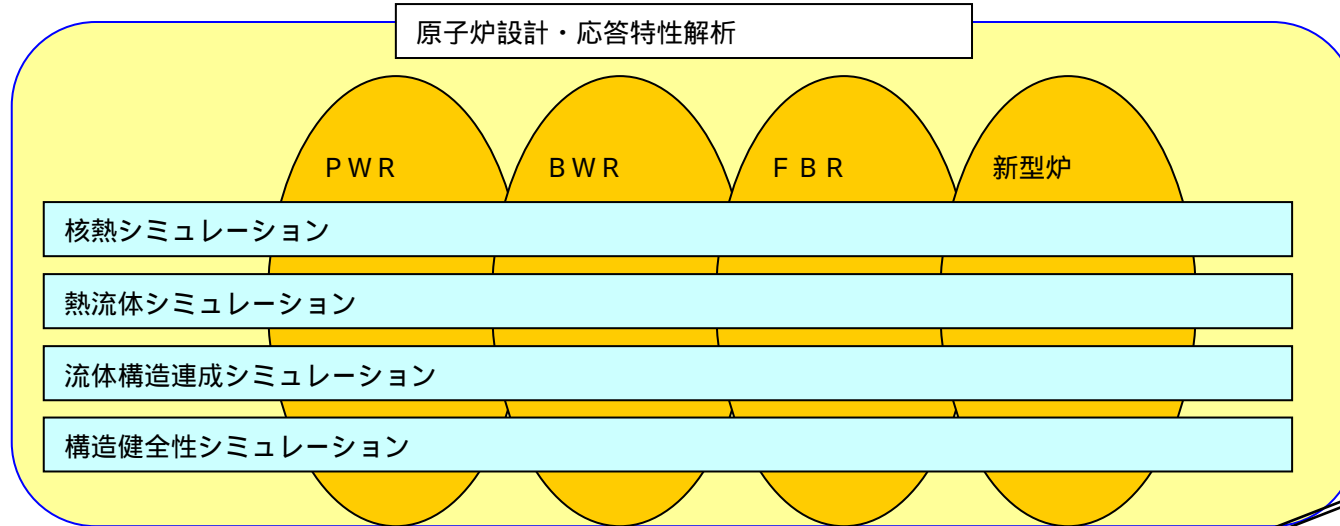
16³ nodes/PE

32³ nodes/PE



: PDJDS, : PDCRS, : CRS-Natural

原子力分野テーマ の位置付け



34	原子力関係の大規模シミュレーション研究*	奥田 洋司	(社)日本原子力学会 大規模シミュレーション研究 専門委員会
----	----------------------	-------	--------------------------------------

*原子力関係の大規模シミュレーション研究 サブテーマ

プロジェクト名	責任者氏名	所属
直接解析手法による原子炉内複雑熱流動挙動の大規模数値シミュレーション	高瀬 和之	日本原子力研究所 東海研究所
溶液の第一原理分子動力学シミュレーション	平田 勝	日本原子力研究所 東海研究所
多階層ダイナミクスが支配するプラズマの構造形成に関する研究	岸本 泰明	日本原子力研究所 那珂研究所
水銀ターゲットにおける液体水銀の圧力波伝播と容器壁の変形挙動と気泡成長の相互作用のシミュレーション	荒川 忠一	日本原子力研究所 計算科学技術推進センター
超伝導ナノファブリケーションによる新奇物性と中性子検出デバイス開発のための超伝導ダイナミクスの研究	町田 昌彦	日本原子力研究所 計算科学技術推進センター
放射線照射に伴う材料の物性変化と破壊の微視的シミュレーション	蕪木 英雄	日本原子力研究所 計算科学技術推進センター
地下空間における放射性核種移行と地下水挙動の大規模シミュレーション技術に関する研究	奥田 洋司	東京大学 人工物工学研究センター
耐放射線性SiCデバイス用酸化膜の第一原理分子動力学シミュレーション	宮下 敦巳	日本原子力研究所 高崎研究所
稠密格子燃料集合体サブチャンネル内冷却材直接乱流シミュレーション	二ノ方 寿	東京工業大学 原子炉工学研究所

平成16年度 地球シミュレータ プロジェクト 先進・創出分野

原子力関連

地球シミュレータセンター
ホームページより

<https://www.es.jamstec.go.jp/>

HLW処分場設計の為の 大規模シミュレーション

目的: 廃棄体4万體から構成される処分場モデル上で、**熱-水-応力-物質移行-地球化学現象の連成解析、**
ならびに、**処分場設計パラメータの不確実性を考慮**
した解析を実施する。

効果: HLW処分場のサイト選定に向けて、従来の性
能評価から保守性を排除した、現実的な処分場モ
デルの性能評価、また地下空間の非均質性などに
由来する処分場設計パラメータの不確実性解析が
可能となる。

目次

- 現行のスーパーコンピュータシステム及び研究成果について
- 将来(2010年前後)の研究目標と期待される成果について
- 将来(2010年前後)のスーパーコンピュータシステムについて
 - 背景となるいくつかの考え方
 - どのようなシステムがよいか？
 - システムの運用体制について

ハード開発とソフト開発の スパイラル

- フロンティア領域ではアプリケーションへのニーズがハード開発を牽引
- 専用機の適用分野拡大
 - 静的なデータ構造 専用プロセッサ
 - 動的なデータ構造 F P G A
- ソフトウェア側からのハードウェアの多様性の吸収
- プアーなコンパイラをアプリ側から補完
- 大量生産プロセッサ (GPUなど) のプログラマブル演算機能の利用 差分演算, 行列演算

ソフトウェアの多チャンネル化

- 産業界におけるシミュレーション：
 - 許されるコスト, 時間に制約
 - 設計現場というよりは研究所での利用がほとんど
 - これまではH/Wの機能向上, 低廉化に負うところが大きい
- ハイエンドのコンピュータが産業界で有効に利用されるには, 衝突解析の例に見られるように, 個々の細かなニーズに応じたソフトウェアが必要.
- 多様なコンピュータを, 様々なニーズに特化した使い方ができ, 同時にそれを可能とする応用ソフトウェア, すなわち多チャンネル化したシミュレーションソフト, また, そのための開発環境の整備が課題.

どのようなシステムがよいか？ (1/3)

- パフォーマンス達成ひいては効果的に成果を生み出すには、ある程度アプリケーションのタイプを絞り込む必要
 - 次期システムの必要性と効果は多々あるが...
 - ペタがあれば現在の物理モデルで有意な結果が得られる
 - ペタがあれば現在の物理モデルを改善するデータが得られる
 - × 全ノードを使う必要がない
 - × 物理モデルが成熟していない
 - 大規模性を重視 / 高速性を重視 の2通りの使い方が可能
 - 構造格子 / 非構造格子 / 粒子系 のデータ構造に配慮
 - データの近接性 / 大域性 / 階層性 を配慮
 - 使うときは一人で使う(意味ある使い方をする)

どのようなシステムがよいか？ (2/3)

- コンパイラ、ミドルウェアの充実
 - 実効性能を左右
 - コード開発の効率、信頼性の向上
- データ生成(モデリング)、可視化の支援ツールの充実
 - いつも後まわしになっている
- PCクラスタなどでは、基本的にはPE性能と通信性能(メモリを含む)のバランスのとれたシステムが使いやすい
- しかし、実効ピークが高ければ「くせ」があってもよい

どのようなシステムがよいか？ (3/3)

- 手をかければ高いパフォーマンスが引き出せるシステムが必要
 - 少数の特定アプリには手をかけてよいはず。
 - 多数の多様なアプリ 手をかけず、実効効率は気にせずにスーパークラスタで実行すればよい
- アプリ側にとって都合がよいのは：
 - アルゴリズムに修正が少ないために ノード数が少なく(メモリが大きい) **SMPクラスタ**
 - 高い実効性能を得る 最適化により高いIPE単体性能が得られる **ベクトルプロセッサ** スカラは対ピーク数%

システムの運用体制について

- 特定ユーザーへの権利と責任
 - いまのES方式には緊張感があってよい
 - 課金なし
 - ユーザーへの研究費手当ての制度化(アカウントだけでは動けないケースがある)
- LINPACKに一喜一憂しない、もしくは新たなベンチマーク問題の設定
 - 「ピーク性能の足し算」は無意味
 - アプリケーションの実効性能と研究成果で勝負
- 遠隔アクセスの許可