ポストペタスケール時代のメモリ階層の深化に対応するソフトウェア技術

研究代表者
遠藤敏夫
東京工業大学 学術国際情報センター 准教授

JST-CREST「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」

研究計画概要

- ポストペタ時代の気象・医療・防災シミュレーションの大規模化・高性能化実現に向けて、メモリウォール問題の悪化が妨げとなる
- システムソフトウェア・アーキテクチャ・アプリにまたがったco-designにより問題解決を図る

異種メモリ階層アーキテクチャを想定

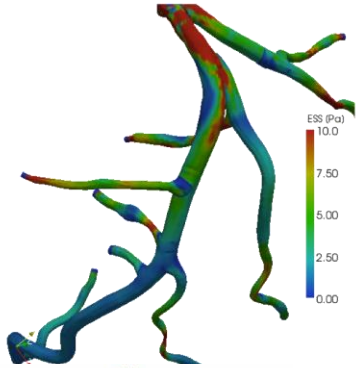
- HMC, NVRAMなど次世代メモリ技術含む

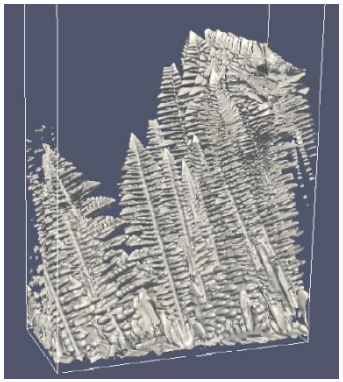
階層活用システムソフトウェアの研究

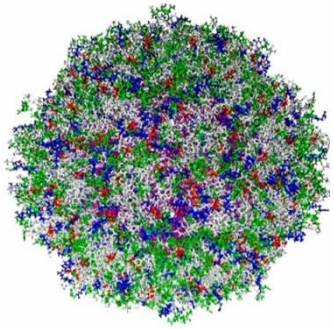
- コンパイラ・ランタイム・メモリ管理の連携によりアプリの局所性向上を自動/半自動で実現

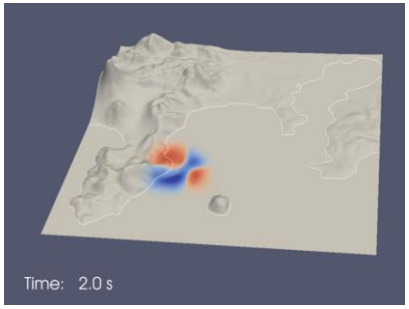
次世代大規模・高性能シミュレーションの実現

ポストペタ時代に実現が求められる安全・安心社会のためのシミュレーション

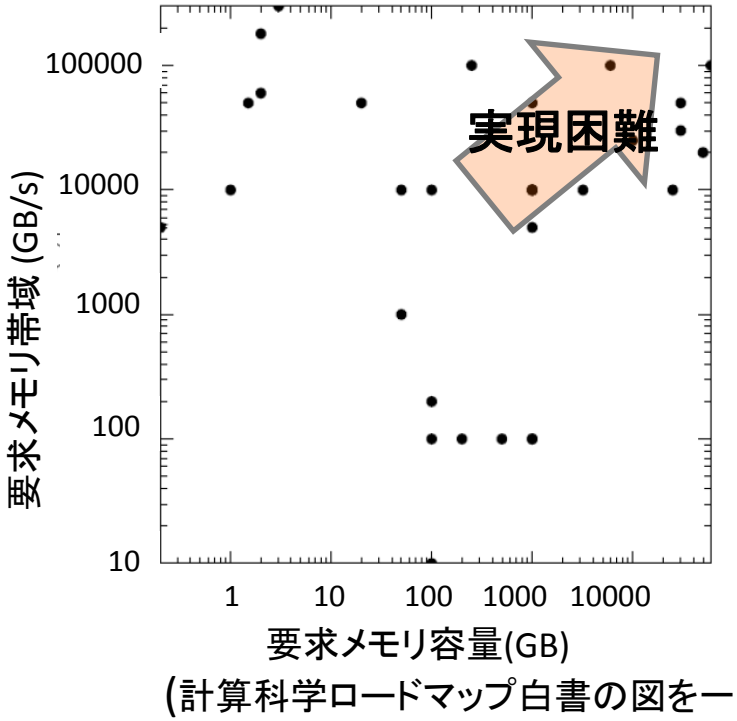








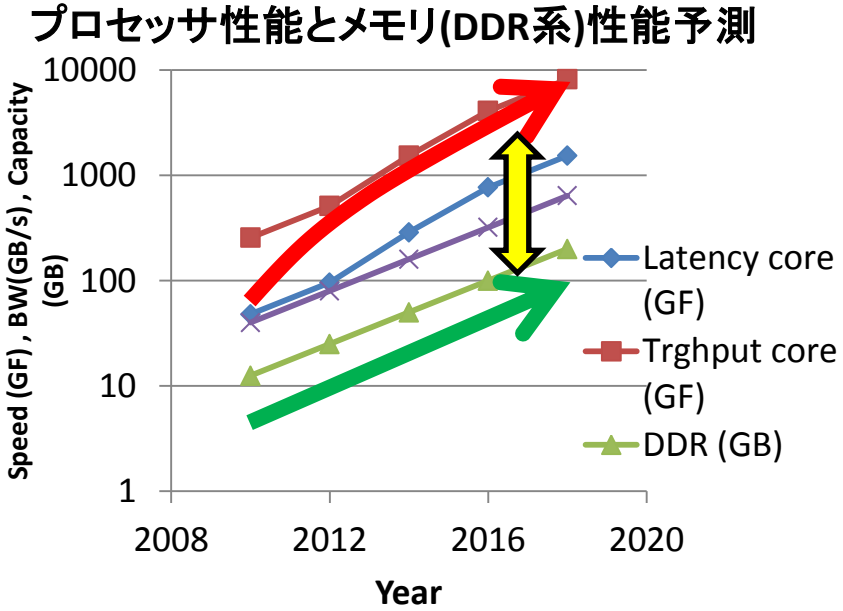
約50種のシミュレーションが要求する
メモリ容量・帯域見積もり (100TFlopsあたり)



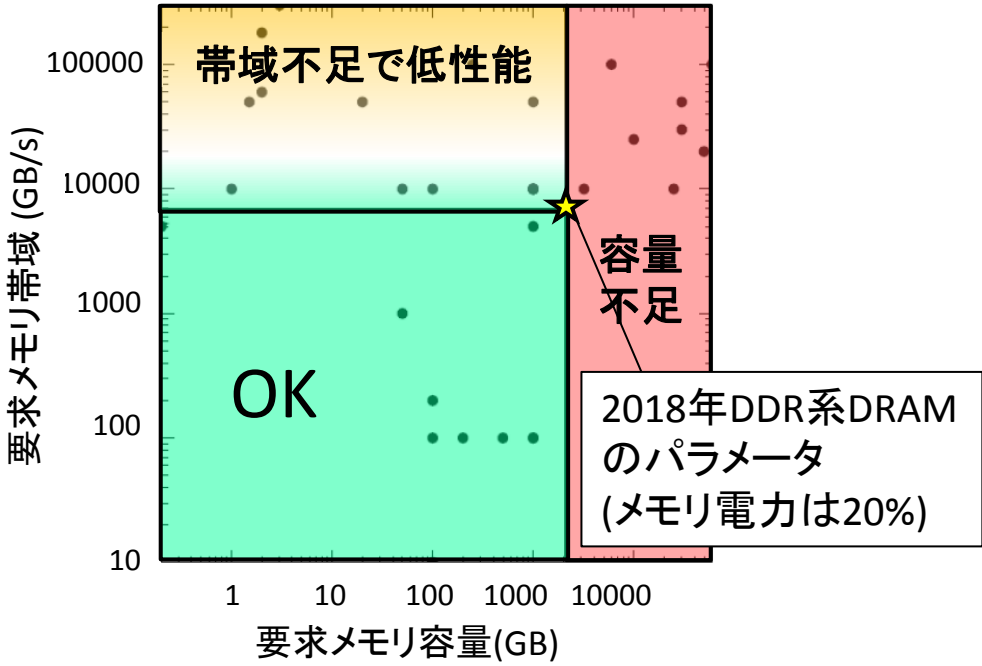
(計算科学ロードマップ白書の図を一部変更)

既存HW/SW技術の延長で
多様な要求に応えられるか？

ポストペタ時代におけるメモリウォール問題の悪化



※ メモリ電力がシステムの20%と仮定



現在のHW技術(DDR系DRAM) の延長では、提供可能な

- **演算速度あたりの帯域** (B/F)
- **演算速度あたりの容量** (B/FLOPS)

が**低下**してしまう

OK領域に含まれるアプリを増やす必要!!

単純なアプローチ:
多数ノードの利用、別スパコンアーキ利用
→ エネルギー・計算資源利用効率に難

既存のアプローチと本研究のアプローチ

アプリの「手動」局所性向上改良

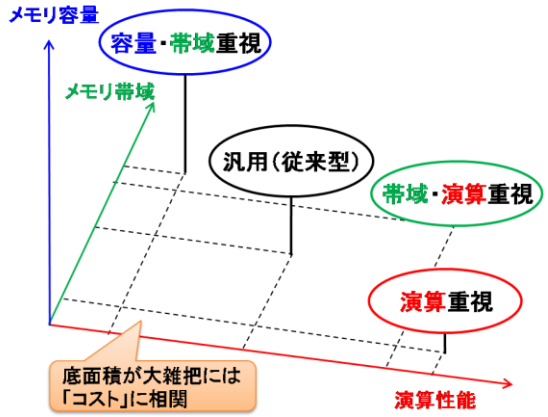
- Level3 BLAS

→ カーネルをライブラリ化できる計算だけはない

- ステンシル計算の空間ブロッキング・時間ブロッキング
- 京のRSDFTの局所性向上アルゴリズム

→ アプリ開発コストを押し上げる;さらに複雑化するアーキにすぐ対応が困難

アプリ特性に合わせたスパコンアーキの開発



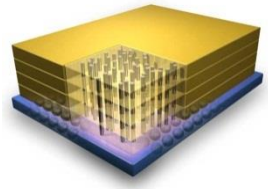
→ 設計パラメータはメモリだけではなく、無尽蔵に大規模システムは構築できない

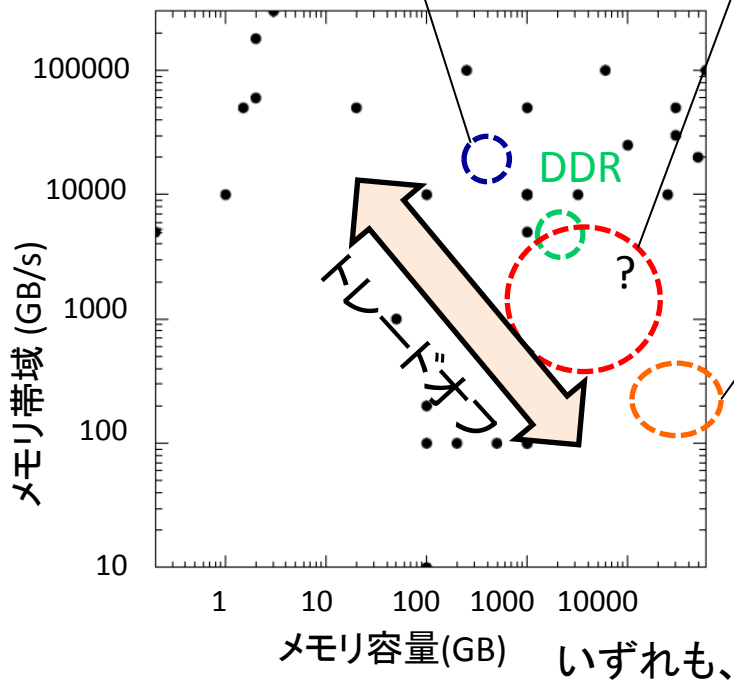
本研究のアプローチ:
異種メモリ階層＋階層利用システムソフトウェア

多様化するメモリアーキテクチャ技術の活用

Hybrid Memory Cube (HMC)

- DRAMチップの3D積層化による高帯域化
- DDRより電力あたり容量は不利
- Micron/Intelなど



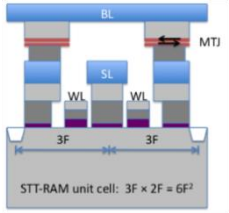


いずれも、2018年ごろの見積もり

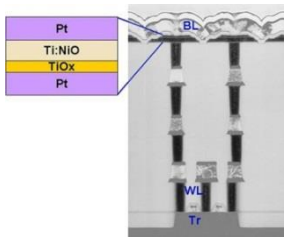
次世代 不揮発メモリ(NVRAM)

- DRAMと異なる記憶方式
- アクセス速度・密度・write耐性まちまち

STT-MRAM



ReRAM



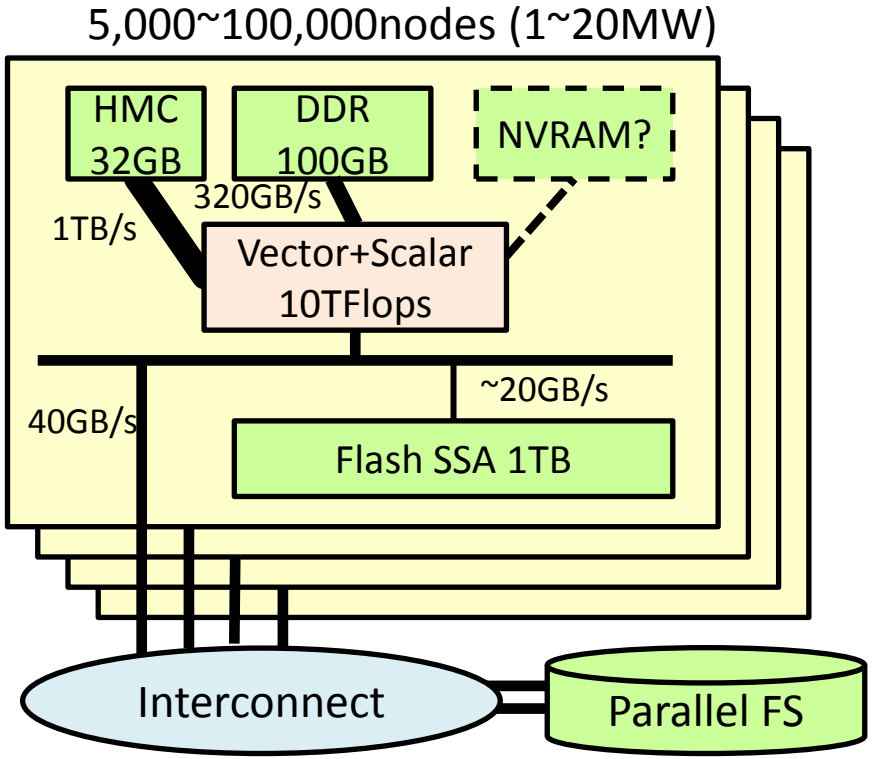
他、PCM, FeRAM・・・

高速Flashメモリ

- PCI-Express直接接続・デバイス並列化によりO(GB/s)の帯域
- Solid State Accelerator(SSA)とも

Fusion-io社ioDrive

想定するスパコンアーキテクチャ



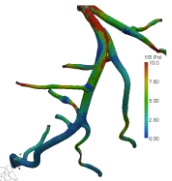
[ノード]

- 10TFlops, 200W
- HMC, DDR, Flash SSAなどの**異種メモリ階層**を持つ
  - 図は、メモリ電力20%を仮定した例であり、メモリ構成自体も研究対象
- リモートノードのメモリも階層に含まれる

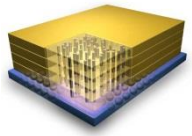
[ネットワーク]

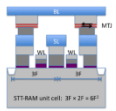
例: ショートカット技術を利用した広帯域・低直径ネットワーク

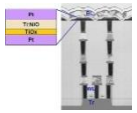
研究のねらい:
トレードオフを持つ異種メモリを有効活用し、大規模・高性能シミュレーションを実現するシステムソフトウェアの研究開発



メモリ階層対応
システムソフトウェア







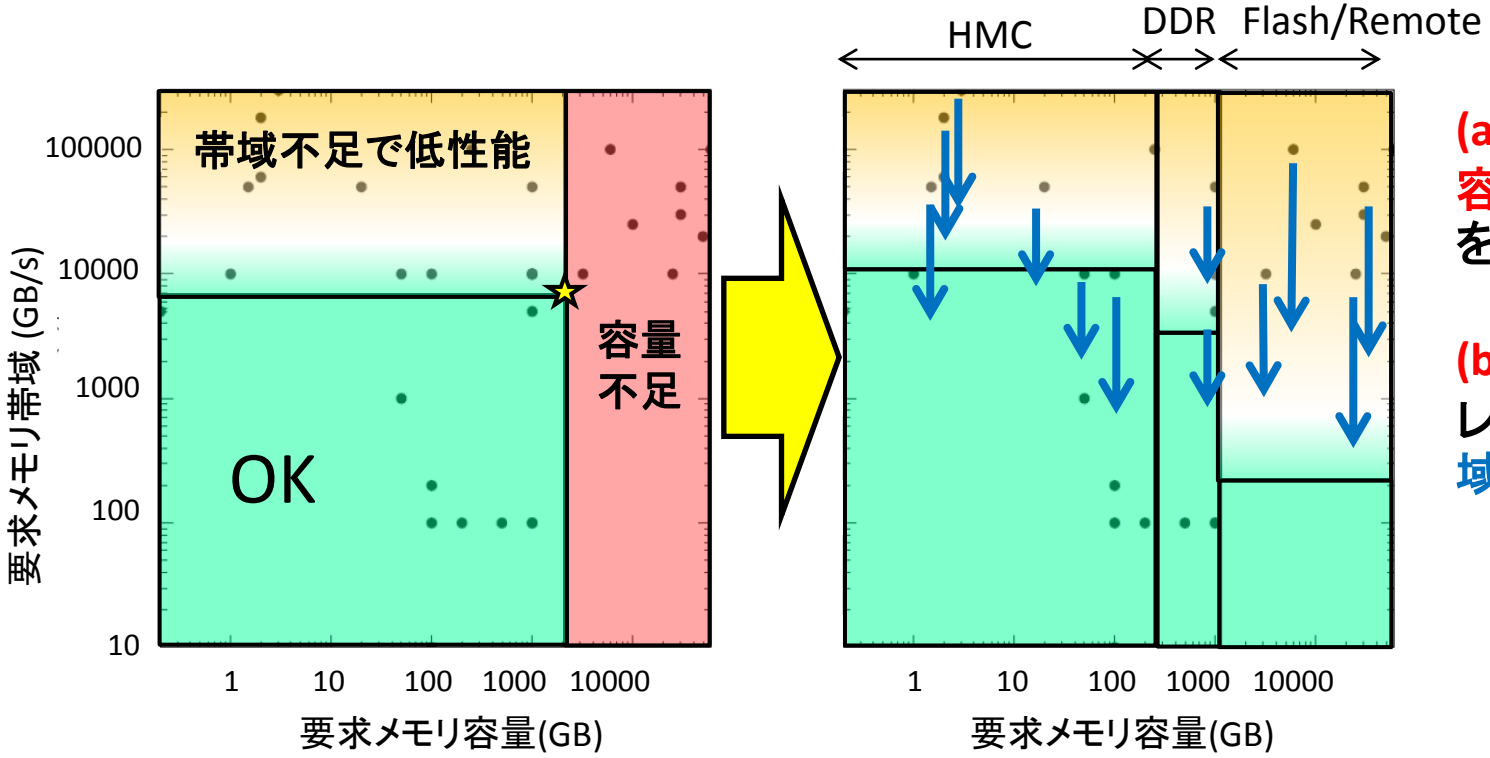
本研究で取り組む技術的課題

メモリ階層の有効活用のため、下記課題の実現に取り組む

(a) データの階層間配置・移動
(b) アプリの局所性向上
(c) 強スケーラビリティ確保

システムソフトウェアのレイヤで、アプリ開発コストを抑制しつつ実現

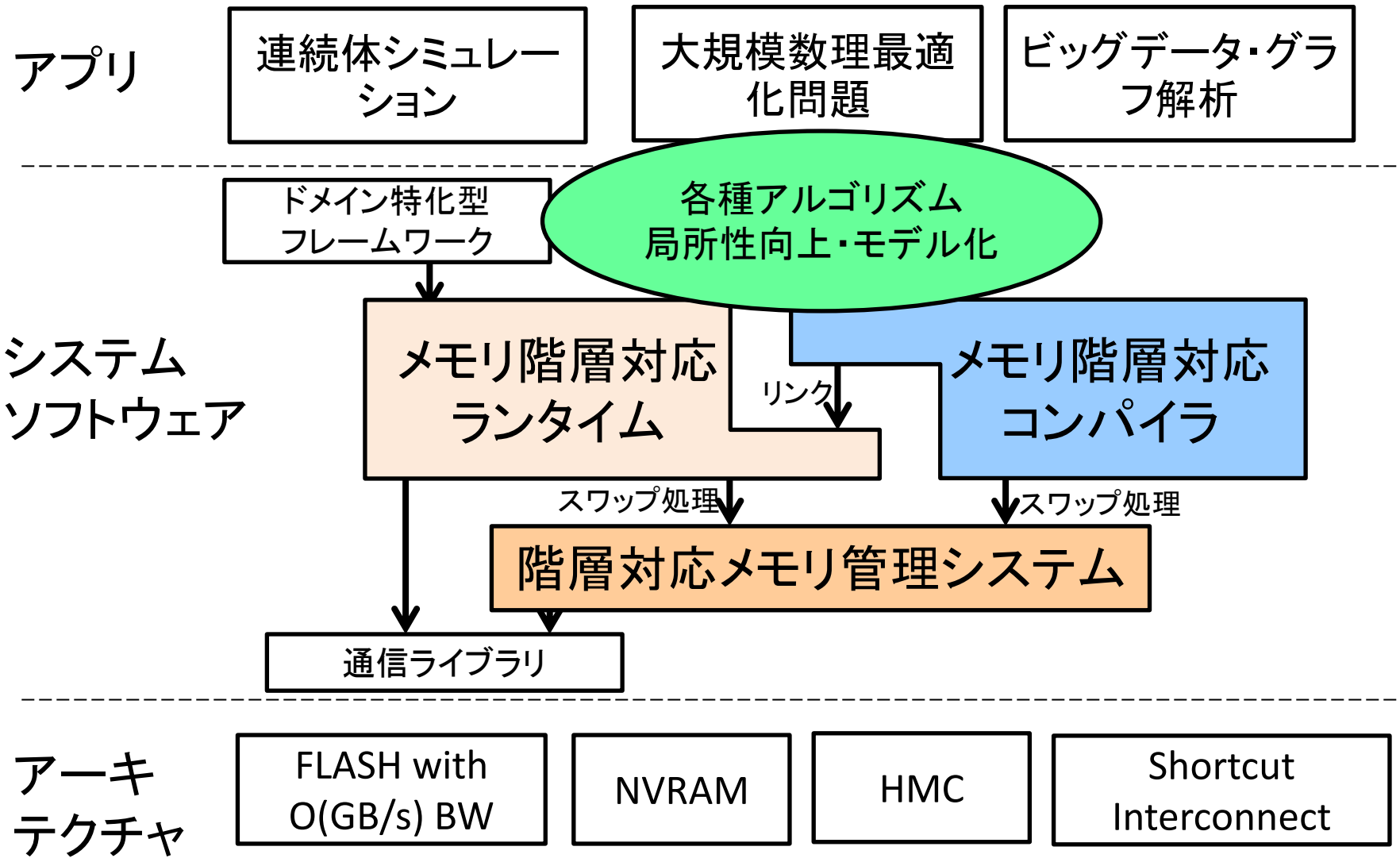
これらの実現により、**OK領域**に入るアプリを増やす



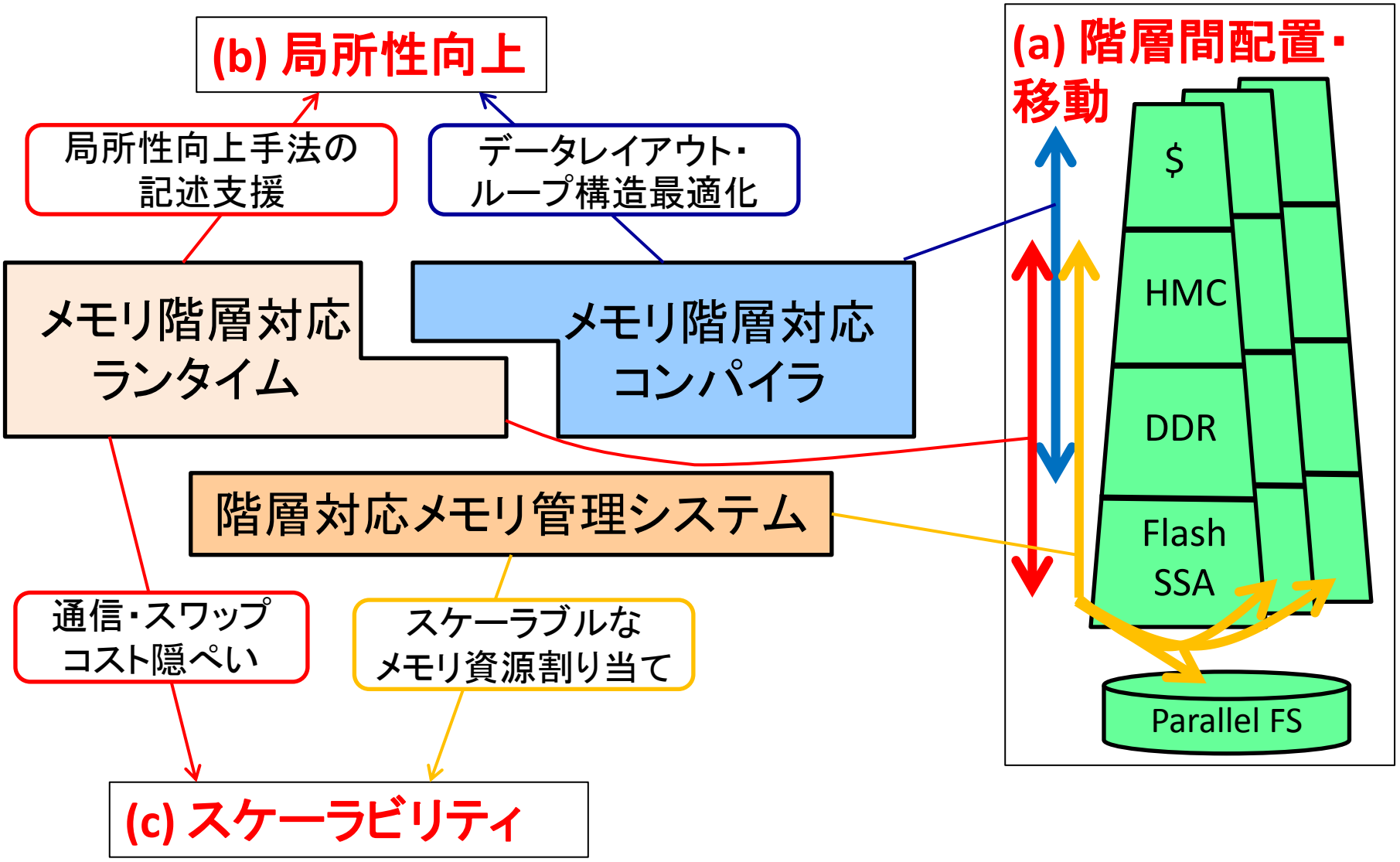
(a)により、原理的に**容量不足**のケースを排除

(b)により、各シミュレーションの**要求帯域を削減**

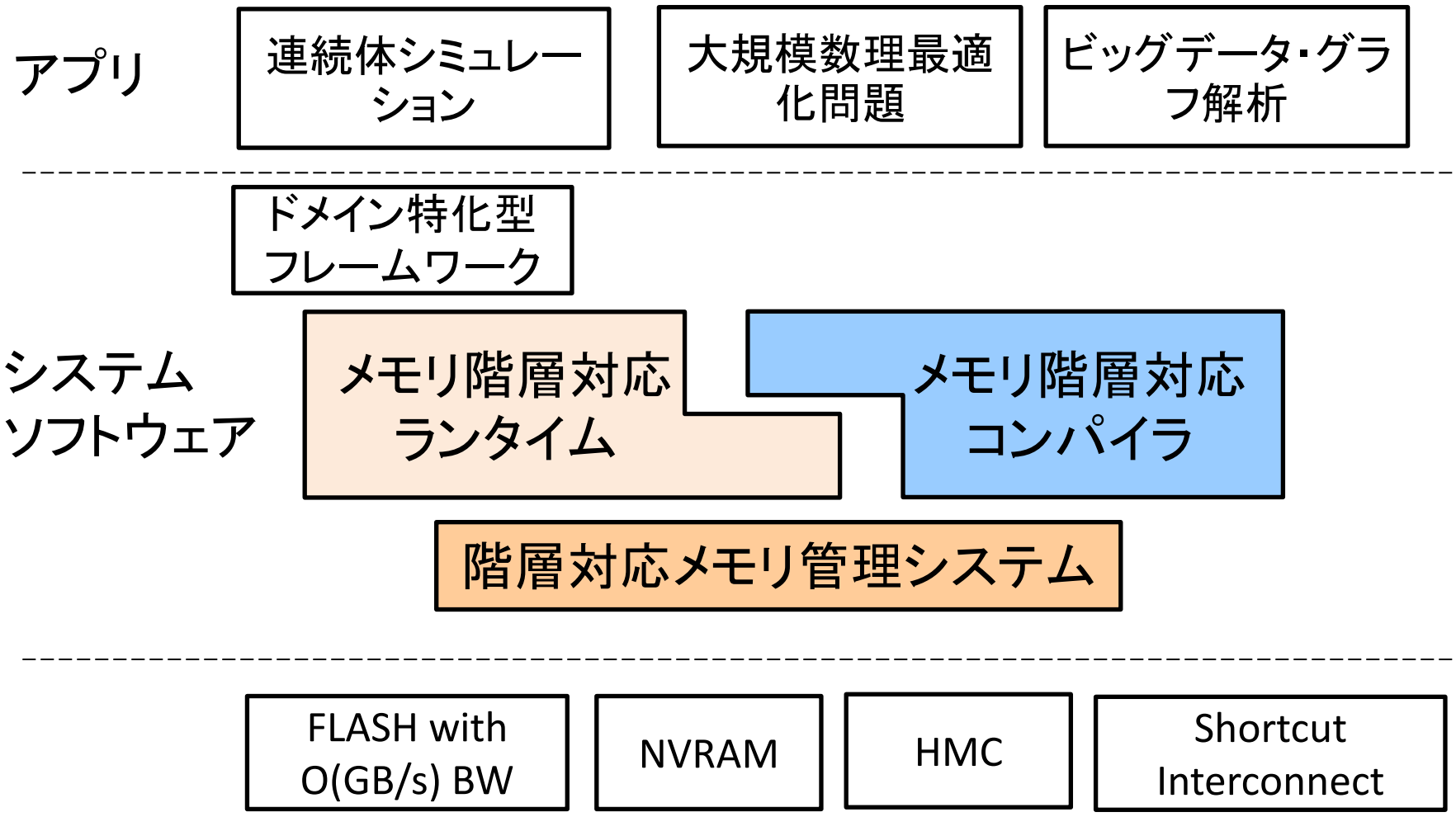
研究開発するソフトウェア要素



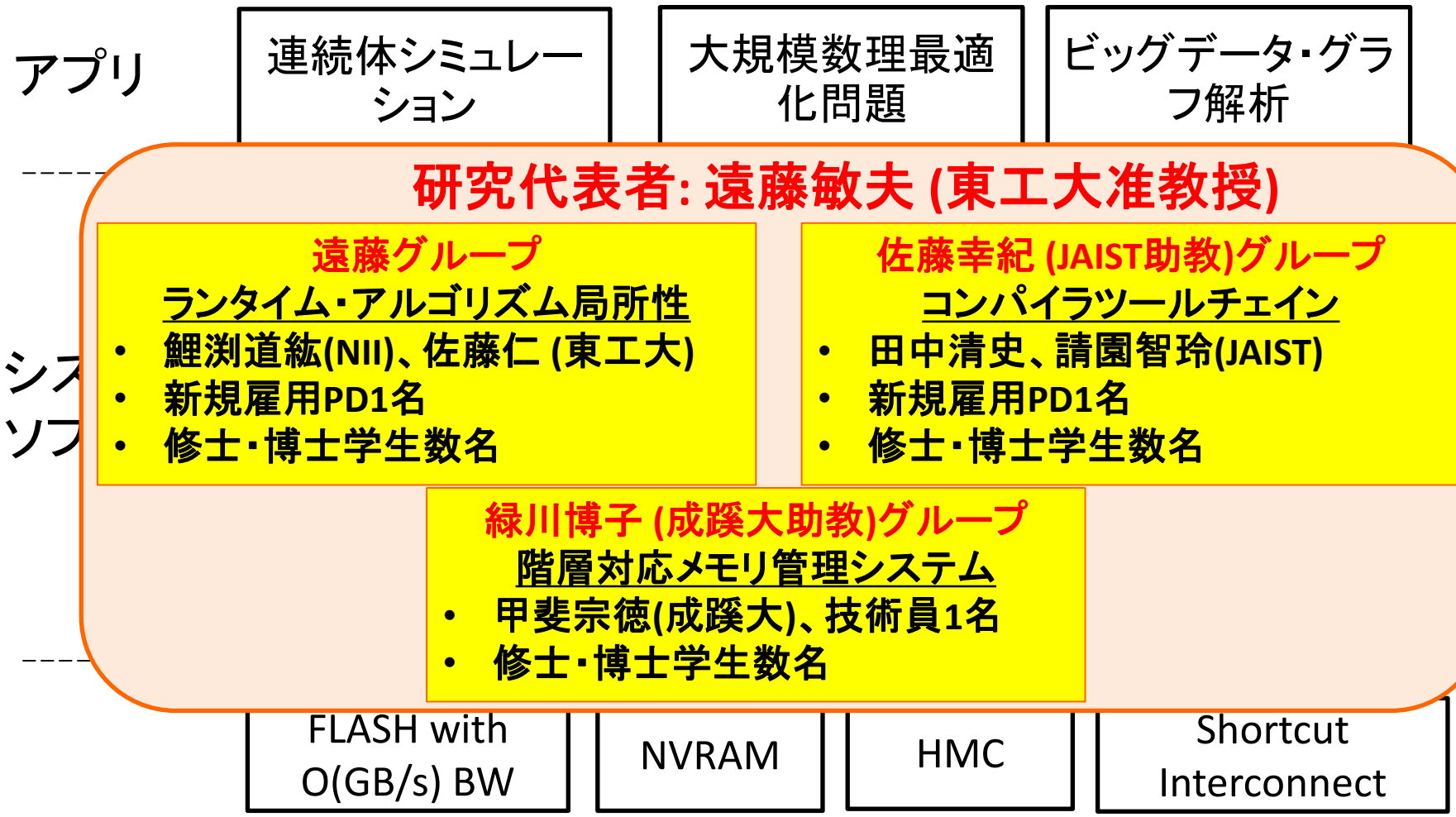
各ソフトウェア要素が解決する技術的課題



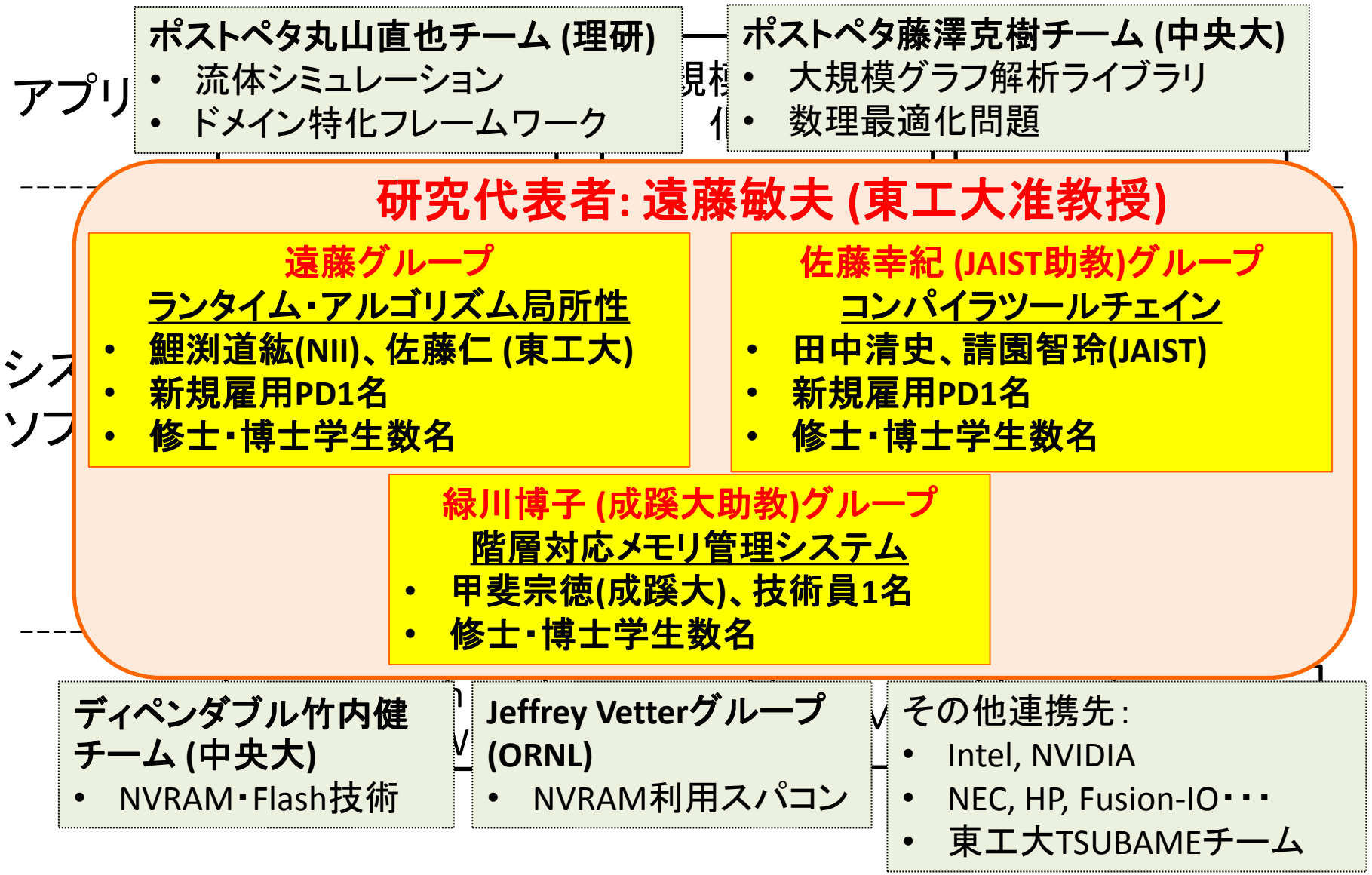
研究実施体制


アプリ
連続体シミュレーション
大規模数理最適化問題
ビッグデータ・グラフ解析
ドメイン特化型フレームワーク
システムソフトウェア
メモリ階層対応ランタイム
メモリ階層対応コンパイラ
階層対応メモリ管理システム
FLASH with O(GB/s) BW
NVRAM
HMC
Shortcut Interconnect

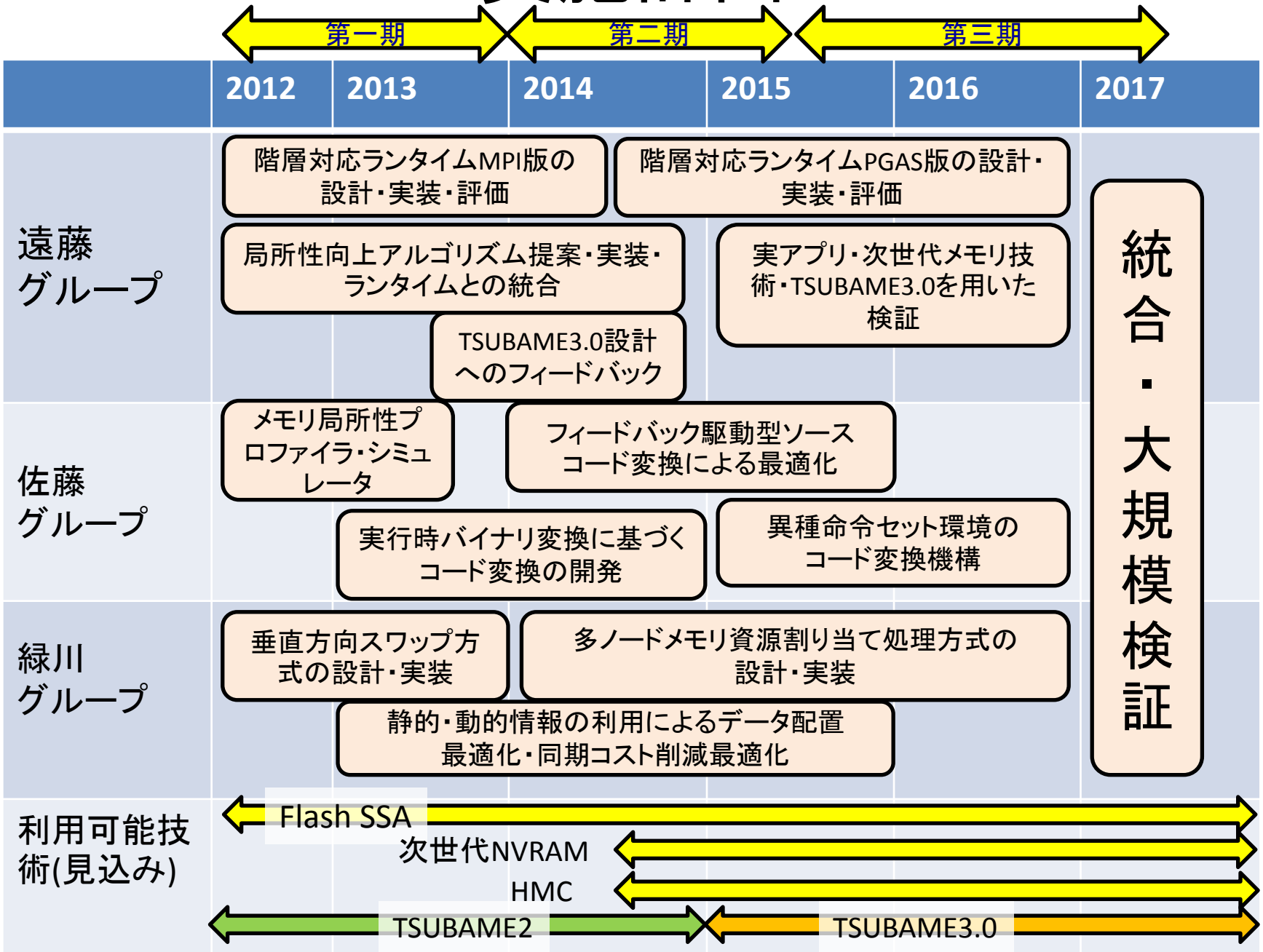
研究実施体制


アプリ
連続体シミュレーション
大規模数理最適化問題
ビッグデータ・グラフ解析
研究代表者: 遠藤敏夫 (東工大准教授)
遠藤グループ
ランタイム・アルゴリズム局所性
• 鯉渕道紘(NII)、佐藤仁 (東工大)
• 新規雇用PD1名
• 修士・博士学生数名
佐藤幸紀 (JAIST助教)グループ
コンパイラツールチェイン
• 田中清史、請園智玲(JAIST)
• 新規雇用PD1名
• 修士・博士学生数名
緑川博子 (成蹊大助教)グループ
階層対応メモリ管理システム
• 甲斐宗徳(成蹊大)、技術員1名
• 修士・博士学生数名
FLASH with O(GB/s) BW
NVRAM
HMC
Shortcut Interconnect

他チームとの連携研究体制



実施計画


第一期
第二期
第三期
2012
2013
2014
2015
2016
2017
遠藤
グループ
階層対応ランタイムMPI版の
設計・実装・評価
階層対応ランタイムPGAS版の設計・
実装・評価
局所性向上アルゴリズム提案・実装・
ランタイムとの統合
実アプリ・次世代メモリ技
術・TSUBAME3.0を用いた
検証
TSUBAME3.0設計
へのフィードバック
佐藤
グループ
メモリ局所性プ
ロファイラ・シミュ
レータ
フィードバック駆動型ソース
コード変換による最適化
実行時バイナリ変換に基づく
コード変換の開発
異種命令セット環境の
コード変換機構
緑川
グループ
垂直方向スワップ方
式の設計・実装
多ノードメモリ資源割り当て処理方式の
設計・実装
静的・動的情報の利用によるデータ配置
最適化・同期コスト削減最適化
統合・大規模検証
利用可能技
術(見込み)
Flash SSA
次世代NVRAM
HMC
TSUBAME2
TSUBAME3.0

メモリ階層対応ダイナミックコンパイレーション(佐藤G)

アプリのデータレイアウトとループ階層構造をランタイムに変換し、自動的に局所性向上を行うコンパイラツールチェーンの研究開発

データの配置・レイアウト最適化

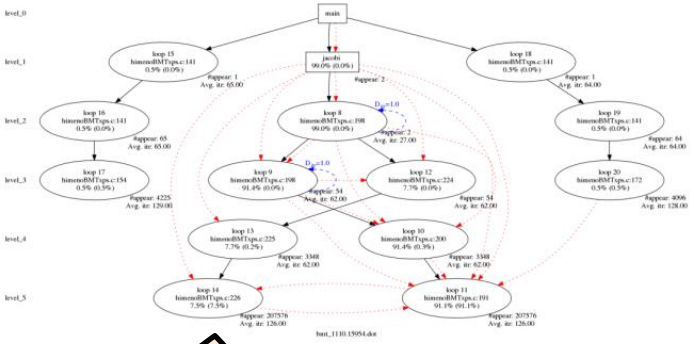
- 初期配置決定・階層間移動
- タイリング、ループ構造に合わせたデータ分割
- Structure of array 対 Array of structure

ループ階層構造の最適化

- loop unrolling, loop exchange
- loop fusion
- loop distribution blocking

- **膨大なソース中の関数にまたがった局所性**
- **動的なワーキングセット変化**
- **キャッシュ＋各メモリ階層の性能特性**

ソースコード情報 and/or
実行中の動的な情報

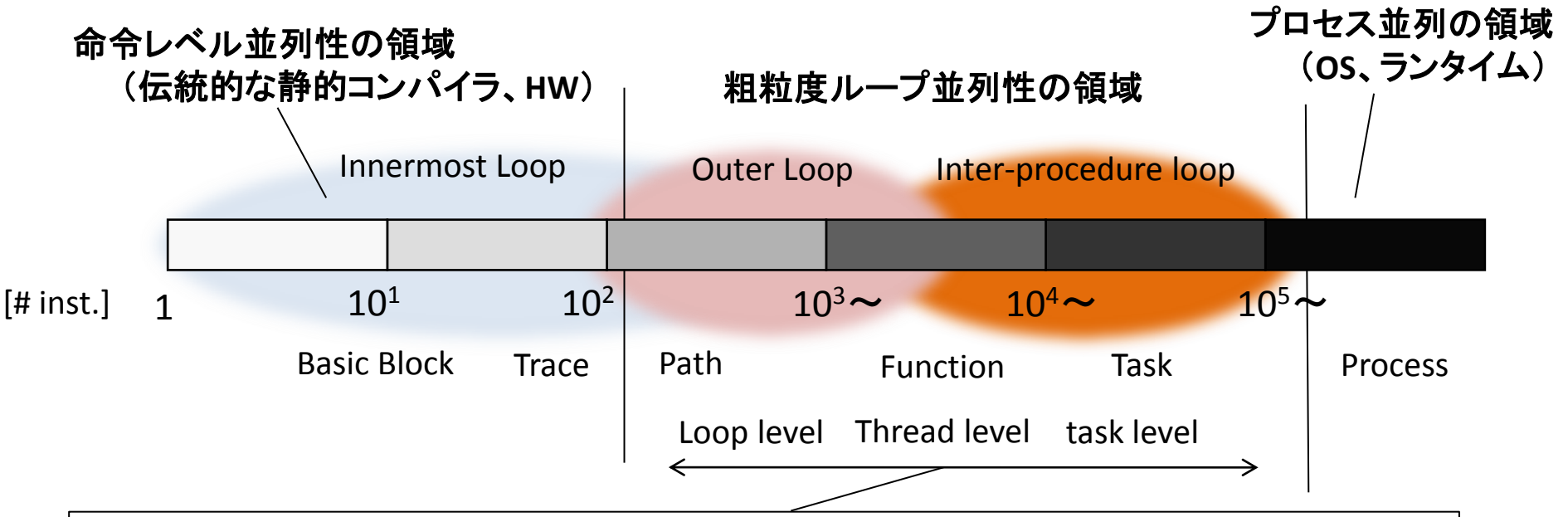


姫野ベンチを佐藤らのデータ依存解析器にて動的解析した結果。
メモリ参照局所性・データ依存関係・ループ情報を含む

1. **メモリ局所性プロファイラ:**
2. **メモリモデルを用いたコード最適化計画:** 想定するメモリ階層における性能予測を行いコード最適化のプランを作成
3. **実行時バイナリ変換・ソース変換によるコード最適化:**

- 動的コンパイル技術に基づく実行時バイナリ変換
- フィードバック駆動型ソースコード変換

多階層の並列性のマッピング



インナーループを超えて多階層の並列性を各種HWのメモリ階層特性にマッピングするためには、高度なチューニング技術が必要である

現状はプログラマによる人力のチューニング

静的なソースを対象とするコンパイラでは効率的マッピングは達成されない

既存のプロファイラは関数単位の実行時間の解析にとどまる

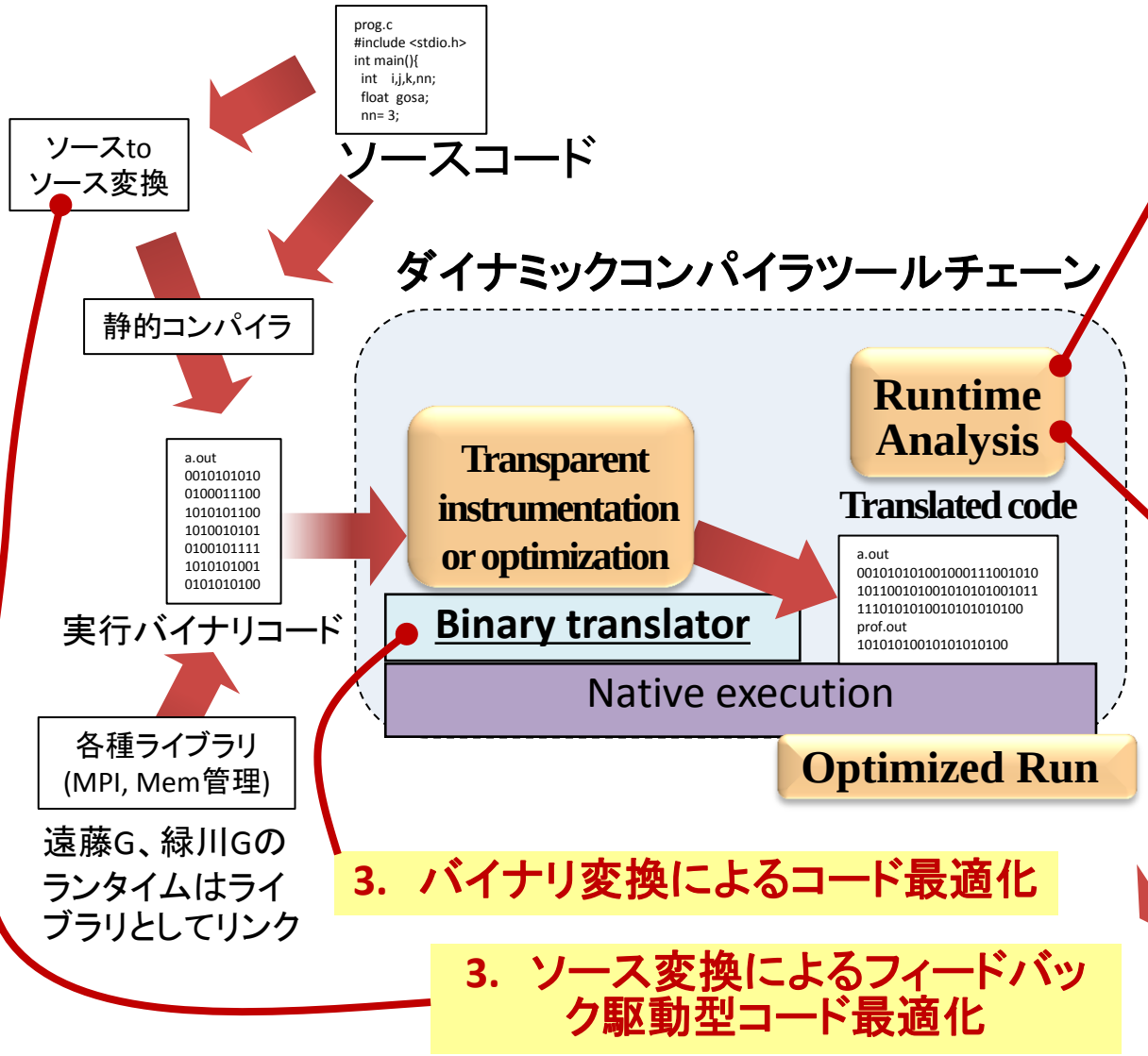
- gprof、Intel Vtune
- HPCToolkit [Rice Univ.]
- CRAY PAT

佐藤グループにて開発したループプロファイリングやデータ依存プロファイリングなど動的解析技術を駆使し、コンパイラ技術を高度化

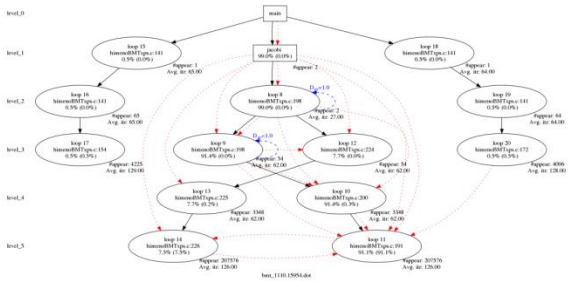
プログラム実行のダイナミックな特性を抽出し、**動的にコードを変換（ダイナミックコンパイレーション）**することにより、これまでの静的コンパイラでは達成できなかった次元である多階層の並列性をHWのメモリ階層特性に自動／半自動で効率的にマッピングすることを試みる

ダイナミックコンパイレーションツールチェイン

バイナリトランスレータを用いたコード変換を利用して、プロファイリングや最適化コード生成を行うツールチェインを開発する



1. メモリ局所性プロファイラ



メモリ参照局所性、データ依存関係やループに関する情報

2. メモリモデルを用いたコード最適化計画

アクセスパタンをキャッシュシミュレータに入力し、メモリ性能見積りモデルを構築。本モデルに基づきコード最適化計画を作成。

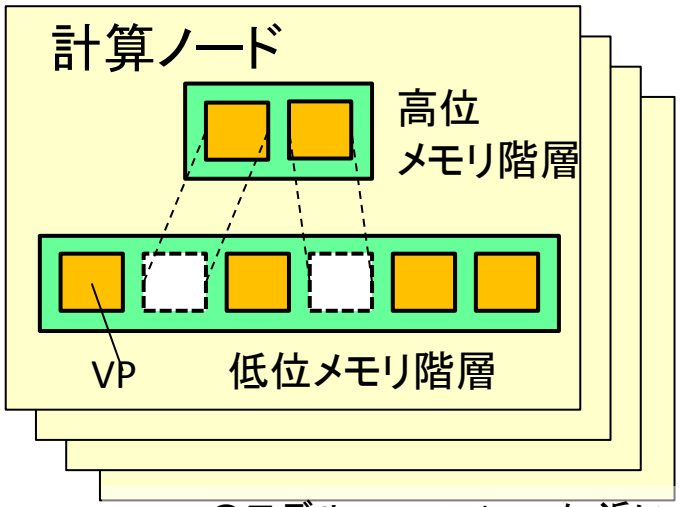
メモリ階層対応ランタイムライブラリの開発 (遠藤G)

既存の並列アプリ (例: MPI+CUDA)を少ない変更でメモリ階層に対応させるランタイムライブラリと、それを活用したアルゴリズム局所性向上手法の研究

1. 階層対応ランタイムHHRTの開発

下記のようなモデルに基づくアプリを効率的に実行するランタイムの開発

- モデル：並列ソフトウェアは必要十分な細粒度性を持った仮想プロセス(VP)達から成りたつ
- VP単位で階層間を**半自動でスワップ**しながら動作させるランタイムを開発。スワップコスト・通信コスト隠ぺいも



HHRTのモデル。UIUC Charmに近いが半自動スワップに対応

研究期間前半：MPI+CUDAベース ⇨ **研究期間後半：PGAS+OpenACCベース (検討項目)**

2. アプリアルゴリズム局所性向上の支援

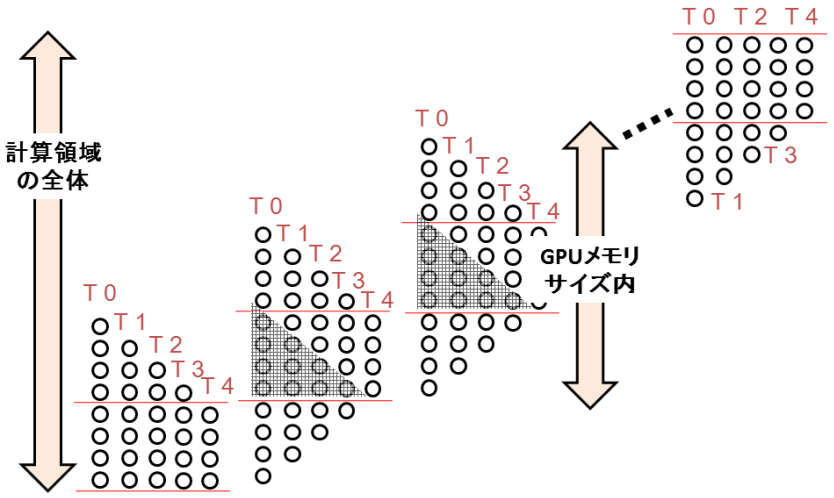
例) 格子系シミュレーションの時間ブロッキング

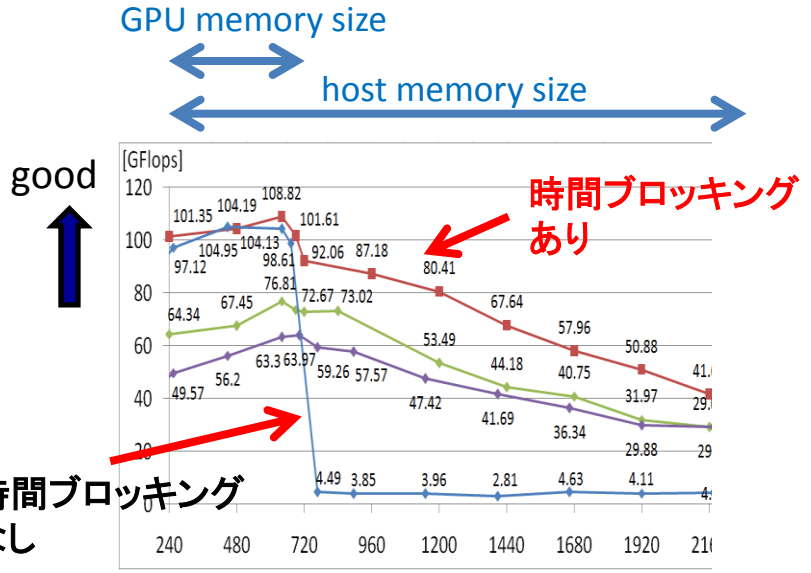
- 実装が複雑となり実アプリへの取入れはプログラミングコスト高
- 冗長計算の存在のためコンパイラ対応は比較的困難

→ HHRTの半自動スワップ機能との組み合わせにより、簡便に大規模・高性能なシミュレーションを実現

アルゴリズム局所性向上手法の予備評価

- GPU上の三次元ステンシル演算を、GPUメモリサイズを超える問題規模で実行
  - GPUデバイスメモリからの「手動スワップアウト」あり
    → ナイーブな実装では、速度1/30に!
  - 時間ブロッキング手法の利用により、良好な速度をキープ
    - キャッシュを対象とすると2~4段だが、約60段ブロッキング
    - キャッシュ効率の向上手法と、冗長計算削減手法も組み込み [Hokke12]





階層対応メモリ管理システムの開発 (緑川G)

省電力、大容量の次世代不揮発性メモリ特性を生かしながら、性能向上を実現する。1ノード内、垂直方向記憶階層（メモリ,SSA,SSD, HD) と、多ノード連携における水平方向記憶（局所・遠隔メモリ、共有メモリ・プール）を統一的記憶システムとして利用するためのシステムソフトウエアの試作、評価を行う。次世代ファイル・メモリ管理方式、マルチスレッド対応などをOSにフィードバックする。

1. **単一ノード内、垂直方向記憶階層間のキャッシュ/スワップ方式の試作と評価**

   SSD、遠隔メモリなど各記憶階層間での性能・容量を考慮、OS成熟度に応じた実装方式、

   次世代Flashを省電力・大容量・遅いメモリとして利用、寿命確保のための書き込み粒度制御など

2. **静的情報利用**による**単一ノード/複数ノード**における**データ配置と限定的メモリ同期方式**の実現

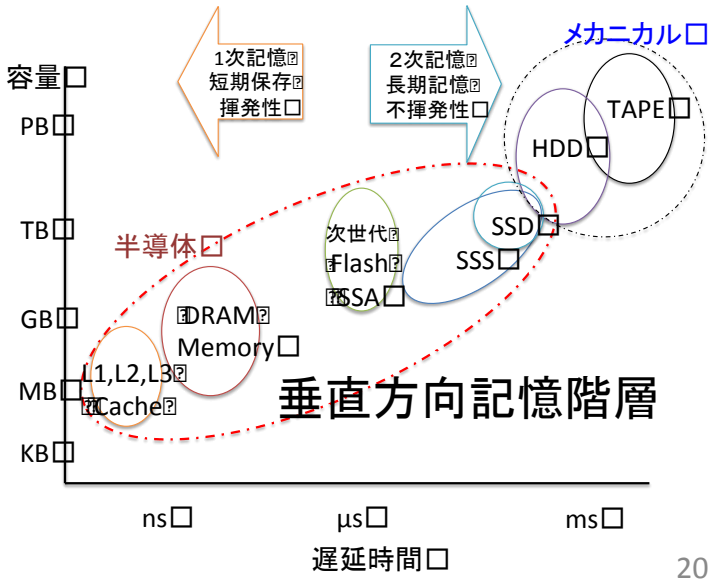
   コンパイラ、ユーザによるヒント（API)による、初期データ配置、規則的繰り返し処理などにおける限定的データ同期方式

3. **実行時情報利用**による**単一ノード/複数ノード**処理における**データ配置と限定的メモリ同期方式**の実現

   実行時メモリアクセス、ワーキングセット見積もり方式の設計、実行環境（各記憶階層の性能と容量）に応じたデータの移動,の次回繰り返し時の事前データ配置、データ転送サイズ（粒度）の自動調整

4. **多数ノードに分散するメモリを有効利用するためのメモリ資源割当方式**の実現

   複数ユーザ間での多数ノードに分散する記憶装置を有効利用するための方式



ローカルメモリの下位メモリ階層として遠隔メモリ利用

ローカルメモリサイズを超えるデータの処理（1ノードによるOpenMPプログラム）

逐次実行（ローカルメモリ率100%1スレッド）に対する性能向上比

DLM(Distributed Large Memory)

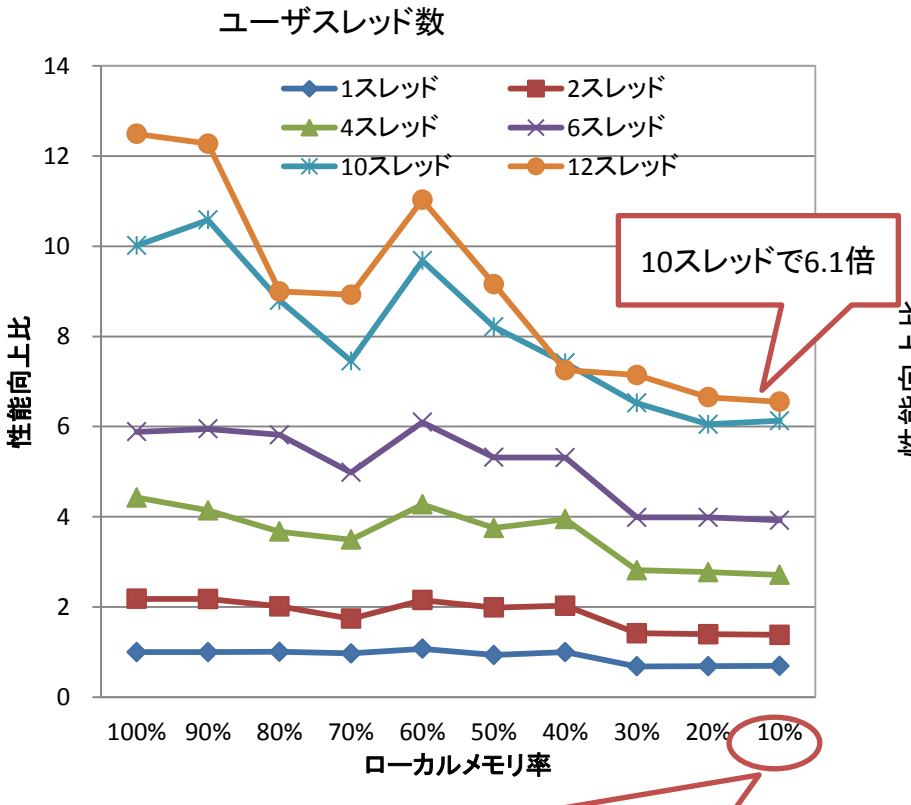
正方行列積(4096x4096，ブロック32x32)

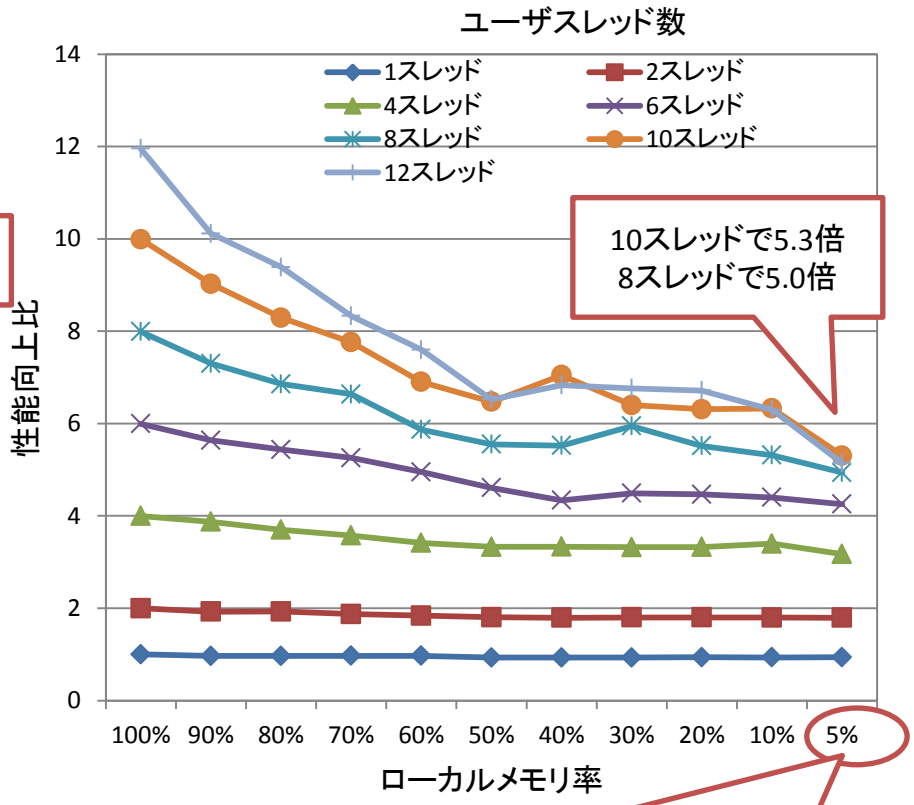
FDA cluster (IPoIB-QDR, 12cores)

ステンシル計算（8192x8192，15x15マスク処理）

FDA cluster (IPoIB –QDR,12cores)

使用スレッド ＝ ユーザスレッド数 ＋ DLM通信スレッド





ローカルメモリ利用サイズの10倍のデータ処理
ローカルメモリ10%, 遠隔メモリ90%利用

ローカルメモリ利用サイズの20倍のデータ処理
ローカルメモリ5%, 遠隔メモリ95%利用

ローカルメモリの下位メモリ階層として遠隔メモリ利用

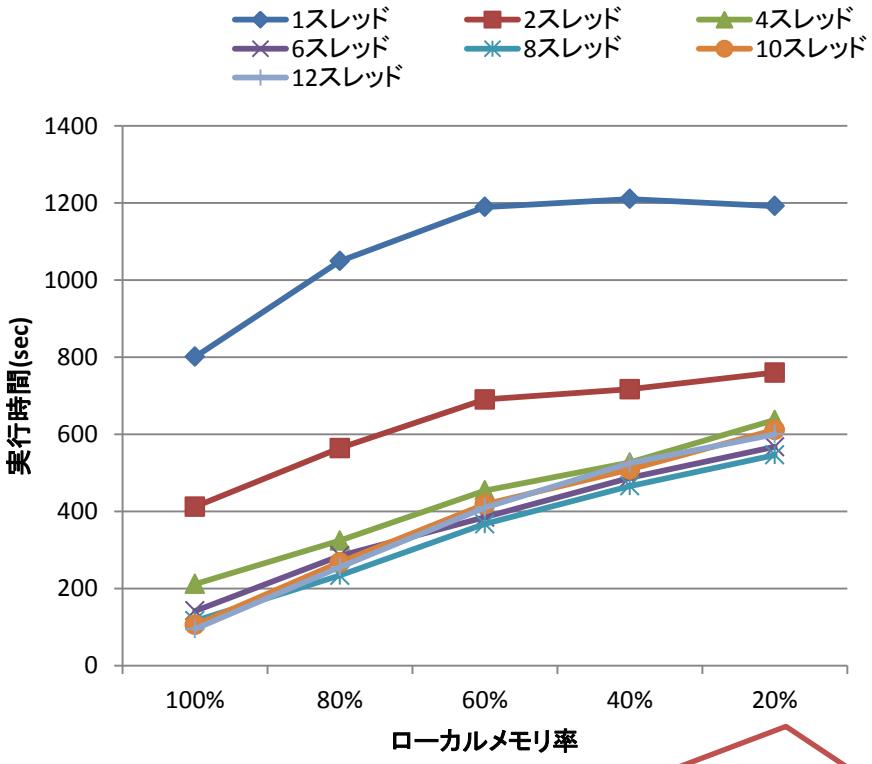
ローカルメモリサイズを超えるデータの処理

（1ノードによるスレッド実装ライブラリ使用 3D-FFT逐次プログラム）

FFTW 64GB （2048x2048x1024 要素） 実行時間

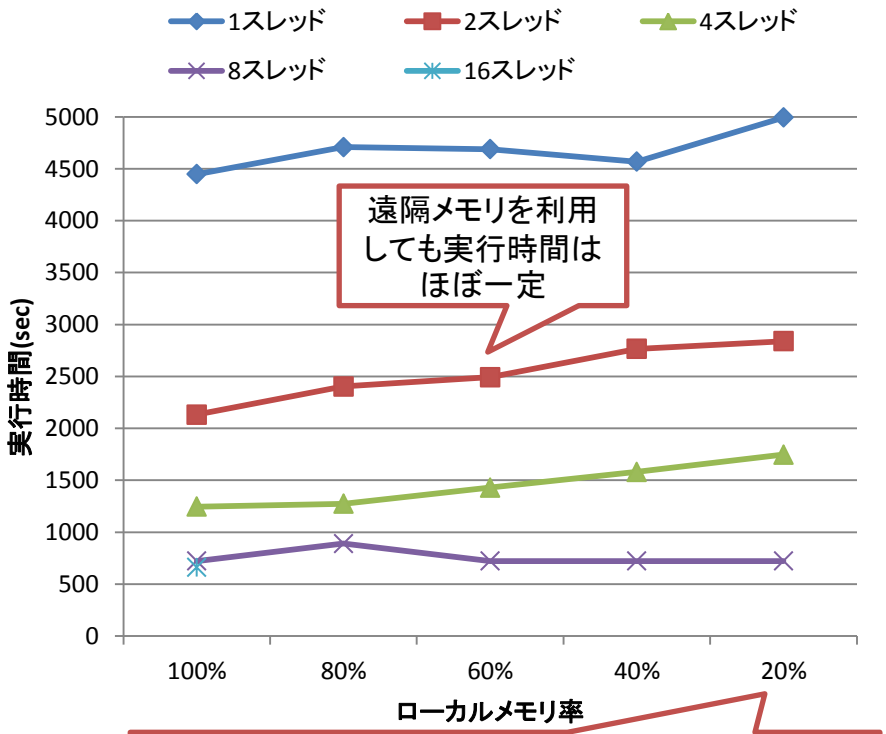
DLM(Distributed Large Memory)

FDA cluster (IPoIB-QDR, 12cores)



FDA Cluster: ローカルメモリ率低下によりスレッド効果が鈍るが，実行時間の低下はない（LMB/RMB比25倍）

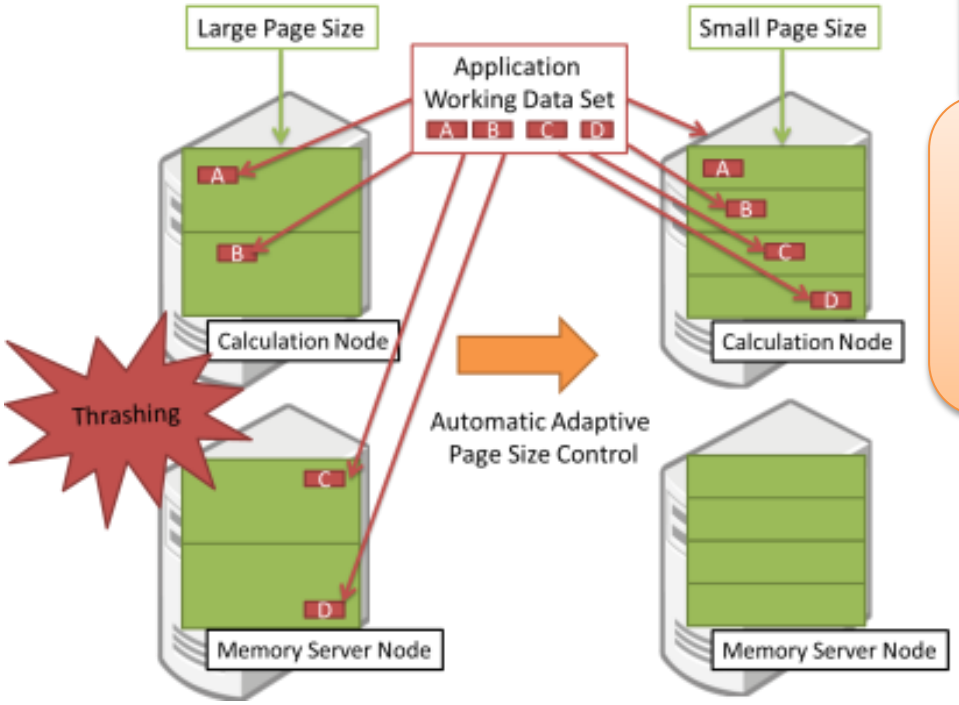
T2K HA8000 (Myri10Gx2, 16cores)



T2K：ローカルメモリ率による性能低下がほとんどない（LMB/RMB比6倍弱）

遠隔メモリページングのための　自動適応型ページサイズ制御機構

応用実行中にWorking Set 推定・ページ通信効率化とスラッシング抑制



コンパイラ
組み込み予定

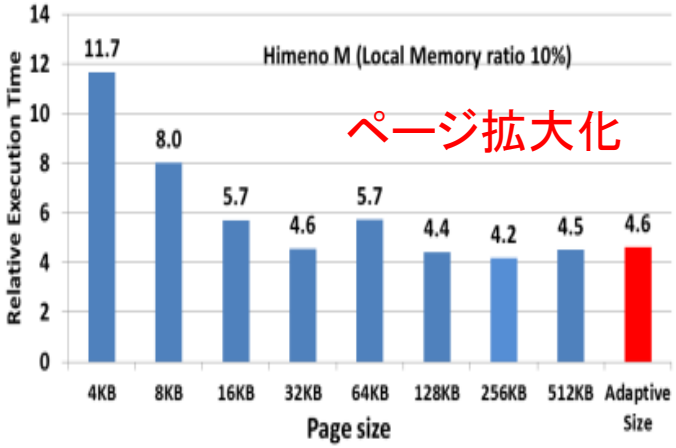
ループ文をヒントにWS推定とページサイズ変更関数を自動挿入、入れ子処理から最適部分処理を自動抽出して制御

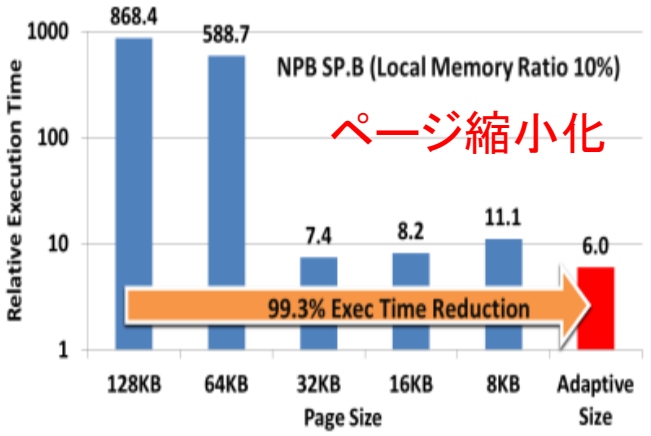
```
for( ite=0; ite<N; ite++){
  Measure_Start(0)
    処理部分0
    Measure_Start(1)
      処理部分1
    Measure_End_Resize(1)
  Measure_End_Resize(0)

  Measure_Start(2)
    処理部分2
  Measure_End_Resize(2)
}
```

応用の部分処理の前後にWS推定とページサイズ変更関数を挿入

繰り返しをもつ応用を対象





まとめ

- メモリウォール問題の悪化に対し、システムソフトウェア・アーキテクチャ・アプリ分野にまたがったco-designにより問題解決を図る
- **次世代気象・医療・防災シミュレーション**と**次世代メモリ技術**とのギャップを埋め、安心安全社会の実現に貢献

  - HMC・NVRAMなどアーキテクチャ分野への要件のフィードバック
  - 局所性向上の自動化・パッケージ化によりアプリ・シミュレーション分野へのフィードバック
  - TSUBAME3.0などポストペタスパコンのデザインへのフィードバック