「演算加速機構を持つ将来のHPCIシステムに関する調査研究」状況報告

主管事業実施機関：筑波大学計算科学研究センター

共同事業参画機関：東京工業大学，理化学研究所，
会津大学，日立製作所
協力機関：東京大学，広島大学，
高エネルギー加速器研究機構

「演算加速機構を持つ将来のHPCIシステムに関する調査研究」

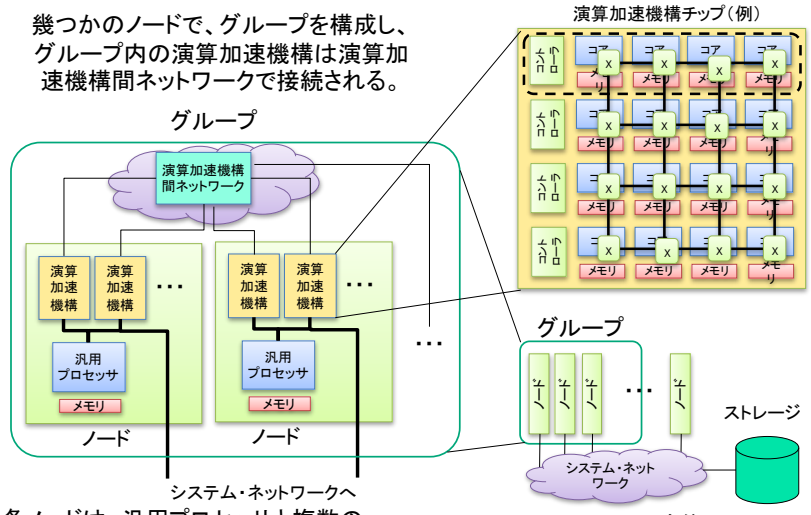
- ナノテクやライフサイエンスの進歩、気候気象予測や地震・防災への対処には計算科学は不可欠かつ有効な手段
  - そのためにはさらなる計算能力が要請されている。
  - 設置面積、消費電力等の制限からノード数の増加による並列システムの性能向上には限界
- ライフサイエンスの分子シミュレーション等、多様な分野で比較的小さい一定サイズの問題の高速化が望まれている(いわゆる強スケーリング)
  - 対応した研究開発の例: ANTON, MDGRAPE-4

電力効率の大幅な効率化と強スケーリング問題の高速化による新たな計算科学の展開を目指して、演算加速機構による並列大規模システムについて調査研究を行う。

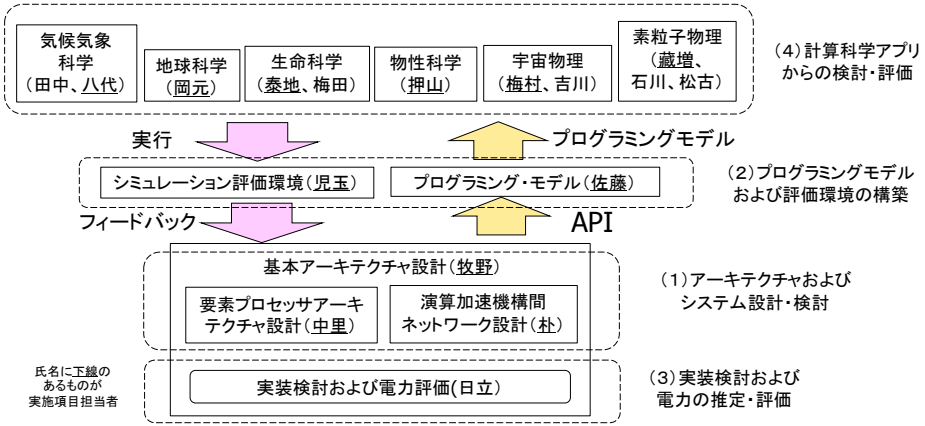
平成23年度文部科学省アプリケーション＆コンピュータアーキテクチャ・コンパイラ・システムソフトウェア合同作業部会において、まとめられた「今後のHPCI技術開発に関する報告書」の中で、分類されたシステム構成のうち**「メモリ容量削減」**および**「演算重視」**のシステムを主な調査研究の対象とする。

多数の演算コアを内蔵したチップによる演算加速機構が汎用プロセッサで構成された並列システムの各ノードに接続もしくは内蔵されているヘテロジーニアスな並列システムを想定

演算加速機構は、多数のスループットコアにより構成。スループットコアは、チップ内ネットワークにより結合される。図に示したものは一つの例。



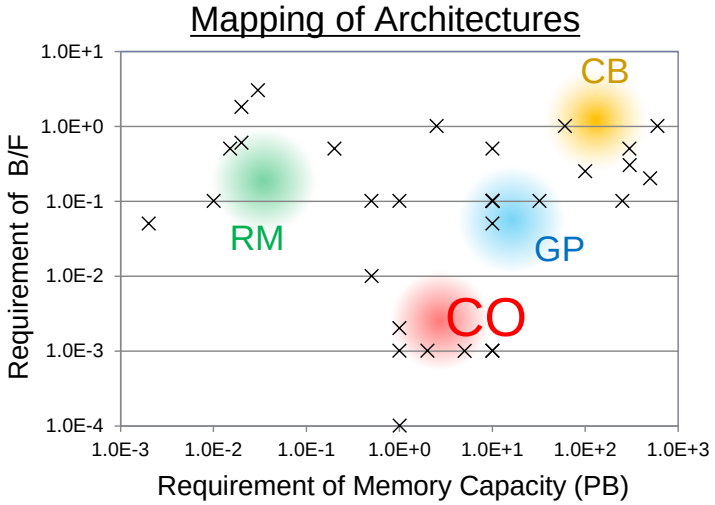
- 主管事業実施機関： 筑波大学　計算科学研究センター
- 共同事業参画機関：東京工業大学、理化学研究所、会津大学、日立製作所
- 協力機関：東京大学、広島大学、高エネルギー加速器研究機構
- 調査研究を、以下の4つの項目に分けて実施
  - （1）アーキテクチャおよびシステムの設計・検討
  - （2）プログラミング・モデルおよび評価環境の構築
  - （3）実装検討および電力の推定・評価
  - （4）計算科学アプリからの検討・評価



評価対象アプリケーション

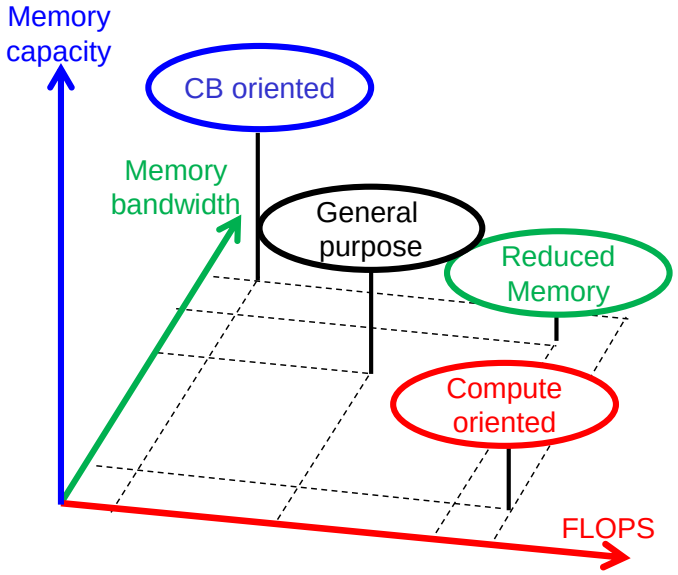
- 計算科学に対する**社会的・科学的課題の達成のために必要なアプリケーション**のうち、本調査研究で対象とするメモリ削減型(RM)および演算重視型(CO)で、ある程度の実行効率が期待できるものの洗い出しを進めている。
  - 生命科学、物性科学における分子動力学計算、生命科学、物性科学、ものづくり分野における第一原理計算、素粒子物理における格子QCD、原子核物理における様々な手法、宇宙物理における粒子シミュレーション、流体計算等(合同作業部会報告より)

（合同作業部会報告より抜粋）



- 本調査研究では科学技術計算における**典型的な計算の一つであるステンシル型の並列アプリケーション**について適用可能の検討を進めている
  - 地震シミュレーション　地震波計算コード　FDM
  - 気象シミュレーション NICAM

- 以下の5つのアプリのカーネルをターゲットとしてアーキテクチャとのco-designを進めている。
  - 格子QCD　（素粒子分野）
  - 重力多体計算　N-body（宇宙物理分野）
  - 磁気流体コード HMD　（宇宙物理分野）
  - 分子動力学　MD（生命科学）
  - 地震波計算コード（地球物理）



- 現在、検討中：NICAM（気象）、RS-DFT（物性）、FMO（化学）

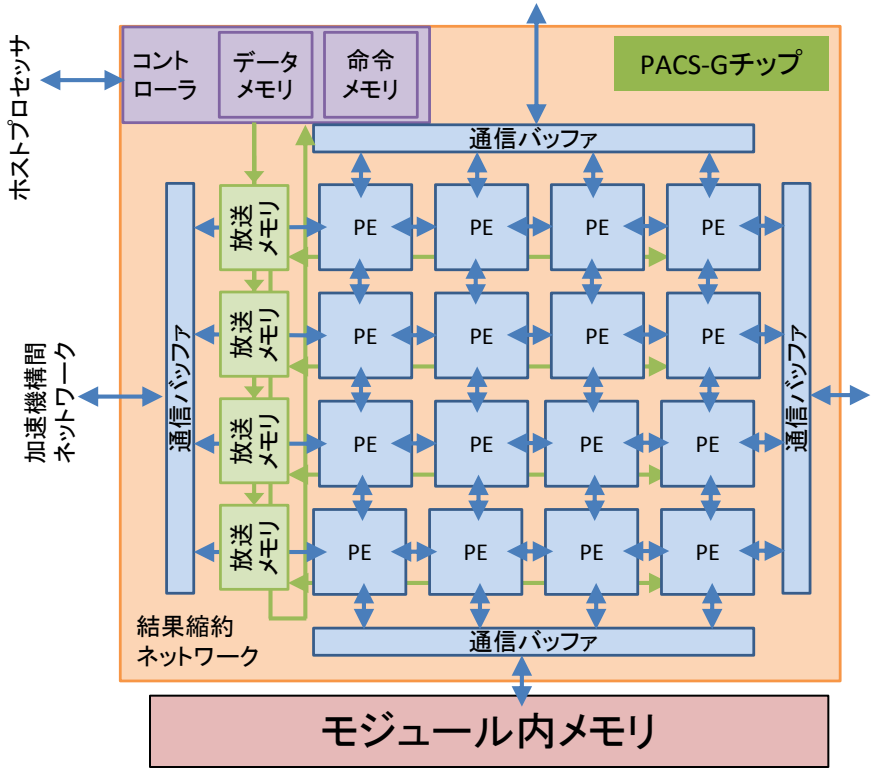
PACS-G アーキテクチャの概要： ノード（チップ）

- 以下のアーキテクチャを、**Straw man（たたき台）アーキテクチャ**として設定

- 演算集約型とメモリ削減型のステンシル計算を両立させるアーキテクチャ（プロセッサ、ネットワーク）をターゲットに設定
- 2018～2020年のLSIテクノロジーとして、14nmを想定。チップサイズを20mm^{2}として、メモリ(SRAM)換算で1GB/チップを想定
- チップの基本アーキテクチャは、SIMD
- チップ内は、2次元のメッシュ・ネットワークを（当面）想定 （＋ブロードキャスト・リダクションネットワークを検討）、コア間 16GB/s (双方向)

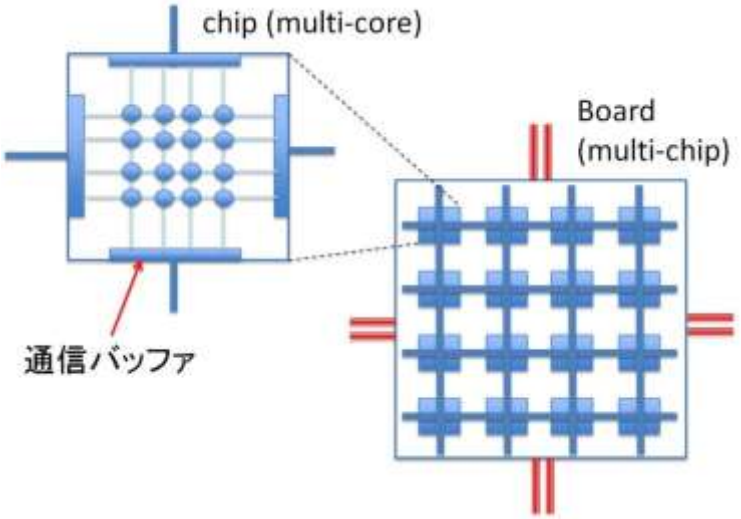
- コアとメモリの比を1:1 として、チップあたり4096 コア（PE）= 64 x 64
- チップ内メモリ 512MB/チップ, 128KB/コア
- コアの基本性能は2FMA@1GHz, したがって、4GFlopsx4096 = 16TFlops/チップ (64Kチップ/1EF)
- TSV 2.5次元実装によるモジュール内メモリを想定。HMC もしくはWide IO DRAM で、
  - バンド幅は1000-1500GB/s
  - サイズは、16-32GB/chip程度
- チップ外付けメモリ(DDR/DIM)は、想定しない

- 電力は250W/チップを目標 （16MW/1EF）
- 2048 チップ/group, Group内のチップは演算加速機構ネットワークで結合



PACS-G アーキテクチャの概要： ノード間、ホスト間

- 1024～2048チップ（グループ）ごとに演算加速機構間ネットワークで結合
- チップの2次元メッシュネットワークをボード上のチップ間ネットワークに展開する際、ボード上のチップ（例えば4x4=16個）を同様に2次元メッシュ結合すると、隣接チップ間（数cm～10cm程度）接続のバンド幅は、チップ内隣接ネットワークの20～40%程度で実現可能（電気配線）
- さらに、ボード間ネットワークまでも2次元（あるいはより高位の多次元）メッシュ展開とすると、インタフェースチップからの光コネクションが使えればチップ間バンド幅と同等（=チップ内ネットの40%程度）が実現可能。

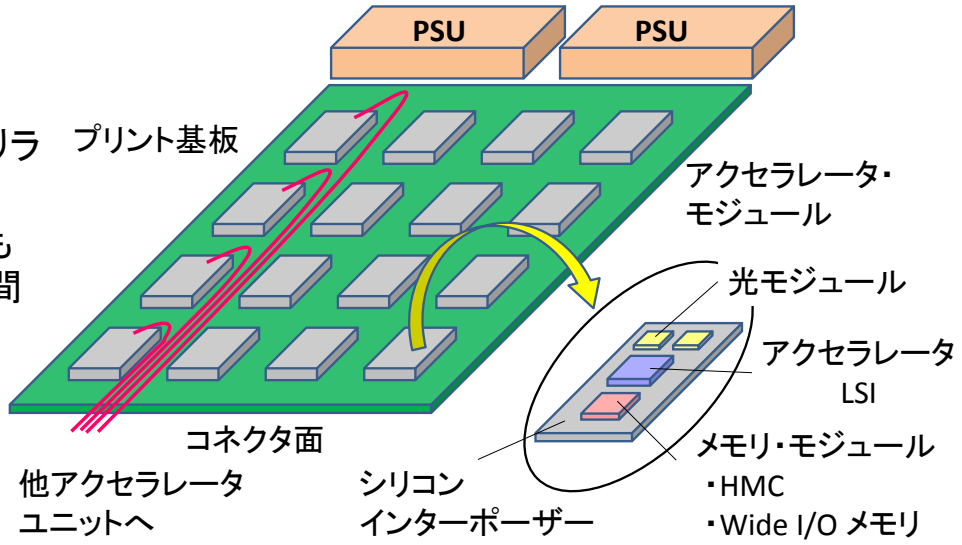


チップ間ネットワークの検討例

チップ内コア間：
16GB/s (= 1GHz x 8B x 2(双方向)@コア間)→1024GB/s (= 16GB/s x 64コア)
チップ間（ボード内、ボード外）：
200～400GB/s (= 32ch. x 25～50Gbps x 2(双方向))

- QCDのような隣接通信のアプリであれば演算のB/F値よりラックサイズのシステムまではメッシュのままで対応可能
- 当面、2次元メッシュで考えるが、もう少し高次元の実装も検討。また、メッシュをトーラスに変更することは、チップ間配線の実装で対応可

- 検討しているシステムは、汎用CPUを基本とした超並列システムにアタッチされることを想定
- ホストとのインタフェースは、PCI Express Gen4 x 16 相当の性能を期待



システムの実装イメージの検討例

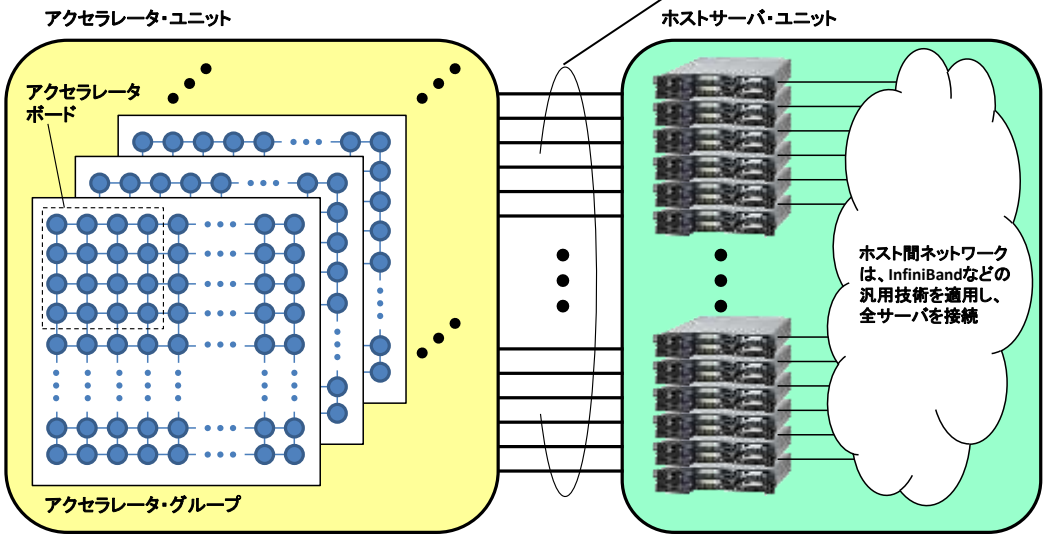
汎用システム（ホスト）との統合イメージ

- 電力・性能等を勘案して、次の2つの構成を想定 (目標全体性能 1EF)
  - ケース1： ホスト 100PF, 演算加速機構部 900PF
  - ケース2： ホスト 300PF, 演算加速機構部 700PF

- システム（論理）構成イメージ
  - 1,024～2,048のアクセラレータ・チップで、グループを構成（図は、2次元メッシュ）
  - 20～50セットのアクセラレータ・グループにより、アクセラレータ・ユニットを構成 （グループ数は、想定性能、グループあたりのチップ数による）
  - ホストサーバーは、Xeon Phi相当のサーバを想定

- ホストサーバとアクセラレータ・ボード間をPCI Express、もしくは、専用の高速伝送路で接続
- ホストサーバ 5～12台に対して、アクセラレータ・ボード 1台の割合で接続
- ホストサーバは、Xeon Phiサーバを想定時(2018年)、18,000～33,000台で構成
- 筐体の中でのアクセラレターボードとサーバの混在の案もあり。

- 演算加速機構の利用イメージ
  - 個々のホストの一部分のコードを演算加速機構にオフロードする。または、ライブラリとして呼び出す（現在のGPUと同様）
  - 並列実行する一部分のコードを演算加速機構にオフロードする。オフロードされ演算加速機構で実行される部分は、演算加速機構ネットワークで通信する。（XcalableMP +OpenACCで記述）
  - ある程度のプロセッサを演算加速機構に付加することも検討。これによりオフロードできる部分を増やす



システムの消費電力

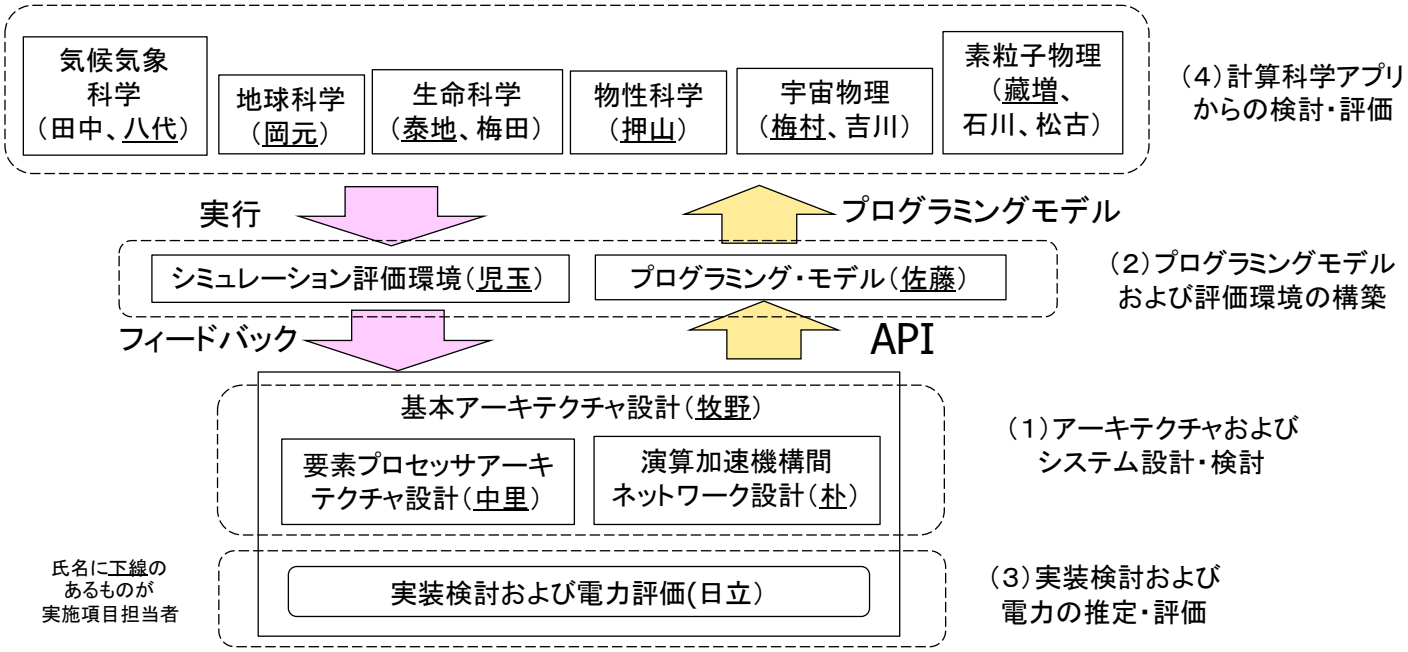
- 消費電力

項目	概要	ケース1	ケース2
構成	演算加速機構は汎用ホストに接続される。全体で1EF (peak)	演算加速機構部 900PFLOPSの場合	演算加速機構部 700PFLOPSの場合
チップ数	1EFでは、62,500チップ	56,250	43,750
チップ全体(MW)	チップあたりの電力は、320～200W(50～80GF/W)	18.0～11.3	14～8.8
メモリ全体(MW)	1チップあたりのモジュール内メモリは20～10W	1.1～0.56	0.8～0.4
周辺 (MW)	ボードあたりの電力100～50W	0.36～0.18	0.3～0.1
加速機構部分合計		19.5～12.0	15.1～9.3
ホスト消費電力	京、およびメニーコアの動向から推定	10MW （10GF/W（京の10倍）を想定）	15MW （20GF/W（京の20倍）を想定）
システム消費電力		29.5～22.0	30.1～24.3

プログラミングモデル、開発・評価環境

- PACS-Gソフト開発用シミュレータ
  - 命令セットの検討
  - プログラミングモデルの検討

- サイクルベース・シミュレータ
  - チップ内の命令実行をクロックサイクル精度で評価
  - PEアーキテクチャは、GRAPE-DRの構成をベースに拡張

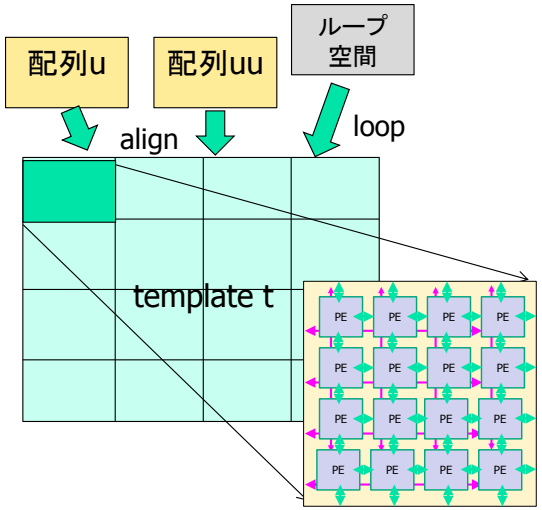


プログラミング・モデルの検討および評価環境の構築の状況

- 性能評価のためのクロックレベルのシミュレータと、ソフトウエア開発用命令レベルシミュレータを開発し、評価を進めている。
  - 命令セットの設計、ネットワーク構成などの設計・評価、モジュールメモリ内の利用、
  - 具体的コードによる定量的評価

- プログラミングモデルの検討
  - アセンブラレベルのSIMDプログラミングをするためのCのsubsetのような言語
  - ユーザに提供するための言語として、XcalableMPの拡張を検討
  - （C＊などのデータ並列言語）

- XcalableMP + OpenACCによるプログラミングモデル
  - ホストからチップにオフロードするために、OpenACCの指示文を用いる。
  - チップの中のプログラミングに、XcalableMPのtemplateを利用
    - Templateは、データやindex空間をマップするための仮想格子
  - ループのプロセッサと配列を整合させることで、PEあたりのコードを生成することができる。
  - データを積層メモリに置く場合は、仮想プロセッサという形でマッピング（？）



- Template directiveで、宣言
- (distribute directiveで、templateをpeにmapping)
- align directiveで、配列を整列
- loop directiveで、ループの実行プロセッサを割り当て

```
#pragma xmp template t(0:XSIZE+2, 0:YSIZE+2)

double u[XSIZE+2][YSIZE+2],uu[XSIZE+2][YSIZE+2];
#pragma xmp align u[i][j] with t(i,j)
#pragma xmp align uu[i][j] with t(i,j)

#pragma xmp loop on t(x,y)
  for(x = 1; x <= XSIZE; x++)
   for(y = 1; y <= YSIZE; y++)
      u[x][y] = (uu[x-1][y] + uu[x+1][y]
              + uu[x][y-1] + uu[x][y+1])/4.0;
```

評価アプリによる性能概算

- 計算の1ステップでの演算、メモリアクセス、通信のパターンと発生量を分析し、想定したプロセッサアーキテクチャ、ネットワークアーキテクチャから期待できる性能を推定。現時点では、グループ単位（～2048チップ,32PF）に限定。
  - 全部データがオンチップに乗る場合：演算・メモリ性能　4 B/F, メモリ1TB/group
  - データが積層メモリに乗る場合：演算・メモリ性能　0.05B/F, メモリ32TB/group

アプリケーション	想定問題サイズ	効率・性能	コメント・比較
格子QCD （素粒子物理）	物理体積$(12fm)^4$ ハドロン多体系 128^4格子	12%～53% 7.9～34.7PF　2048チップ(単精度peak 65.5 PF)	• 評価対象アルゴリズム：領域分割前処理単精度クォークソルバー（ウィルソンクォーク型、BiCGStab法） • オンチップのメモリのみを利用 • 通信レイテンシパラメータの範囲で性能に幅がでる。 • 京では、効率が26%，32768ノード、1.1 PF
磁気流体コード （宇宙物理）	セル数　1984^3	1.89 PF, 22.5% 512チップ(8PF)	• HLL近似リーマン解法、磁場をflux-CT法よる有限体積法。時間積分を2次精度のTVD Runge-Kutta法 • グローバルメモリを利用 • 210-220ms/step，Intel Core i7 4096コアで4.5s/step
重力多体計算 （宇宙物理）	814G interaction/sec/chip （単精度、無衝突系）		• 重力計算を演算加速機構で加速。粒子の軌道計算はホスト計算機で行う。オンチップメモリのみ使用 • Intel Xeon E5-2670 の 66.7倍
分子動力学 （MD）カーネル （生命科学）	1セル/コア、1セル$(5Å)^3$，カットオフ半径12Åを仮定 2580原子/コア	3.67PF、 最大15M原子/256チップ， 784.4us/ステップ	• 近距離相互作用の直接和計算を計算。遠距離相互作用計算、結合力計算は未評価 • セルインデックス法（空間座標分割）とハーフシェルスキームを仮定 • 通信ネックになっていないため小規模問題ではさらに高速化可能? • 京では、全ノードで500M原子、4.6PF, 114ms/ステップ
地震波 計算コード （地球科学）	格子サイズ 2048x2048x512	3.5 PF /1024チップ	• 3次元時間領域差分法（FDTD）、空間差分4次精度、時間差分2次精度、弾性体、速度と応力を変数とするスキーム • オンチップのみ。今後、グローバルメモリも検討。 • 格子間隔 50 m, 最小横波速度 300 m/s を想定した場合 (100 × 100 × 25 km, Δt～0.001s), 実時間の10倍程度の速さ • 現状のGPUクラスタでは実時間の1/20 程度の計算速度(1/200)

仕様の見直し

- アプリの性能および電力とのトレードオフで通信バンド幅の見直しを行った
  - 演算加速コプロセッサのチップ間のネットワークのバンド幅が226GB/sと過大なものになっていたが、内部ネットワークからの延長としてmaxで考えていたもの
  - 結論：28GbpsのSERDESを用いて、8レーン／ポート、ポートあたり22.3GB/s, 3次元トーラスのために6ポートで、十分な性能が得られる
    - 地震波シミュレーション（差分法）：strong-scaleを見込むオンチップメモリだけの場合は50GB/sの場合、効率23.3%, 15GB/sの場合効率14.3%。モジュールメモリを使う場合は、計算律速。
    - 宇宙磁気流体コードHMD：50GB/sの場合、効率21.5%, 15GB/sの場合効率18.5%
    - 分子動力学(MD)コード: バンド幅は20GB/sぐらいで十分。むしろ、レーテンシが重要
    - QCD:バンド幅は20GB/sぐらいで十分。むしろ、レーテンシが重要

まとめ・課題

- エクサスケール（エクサフロップス）システムを実現するには、演算に特化した演算加速機構が必要。（電力性能50GF/W以上、汎用では難しい）
- 次世代の計算科学を展開するには、strong scalingが必要。
- ポイントは、2018-2020で可能になる、14nm-10nmテクノロジーを用いて、オンチップの計算がどのくらいできるか（16TF/chipの場合のmemory wall問題、...）

- 次のステップはより詳細なプロセッサアーキテクチャの決定、命令レベルシミュレータの開発と、それによるアプリケーション（システム全体）の性能評価である
  - ネットワークトポロジーの検討（現在は、２次元メッシュ）
  - ある程度のプロセッサを演算加速機構に付加することも検討
  - ネットワークのルーティング機能の検討
  - プログラミングモデルとコンパイラの実装
  - 実際のコードを用いた定量的評価、詳細な電力評価 (アプリFS mini-app)
  - NICAM（気象）、RS-DFT（物性）、FMO（化学）

- 全体システムの検討
  - ホストとの接続形態の検討
  - システム全体としての耐故障機能
  - I/O等