RIKEN BDR Center for Biosystems Dynamics Research

分子動力学シミュレーション専用計算機の開発：MDGRAPE-4A and beyond

泰地　真弘人 (Makoto Taiji)　taiji@riken.jp

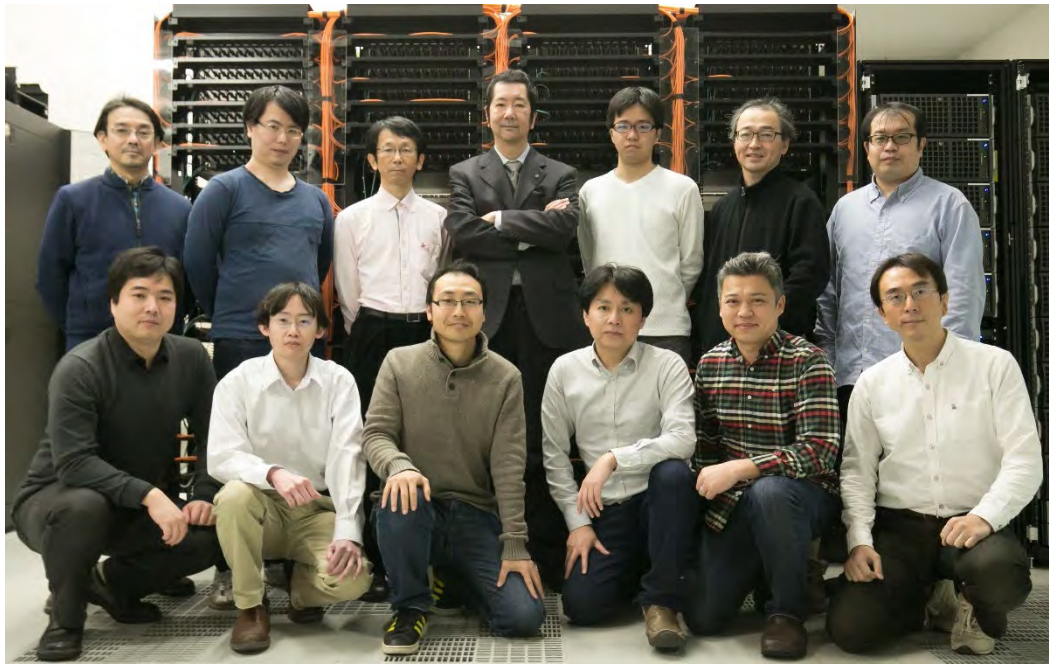
理化学研究所　生命機能科学研究センター
副センター長

Acknowledgements

- RIKEN

  Yousuke Ohno
  Gentaro Morimoto
  Teruhisa Komatsu
  Yohei Koyama
  Hao Zhang
  Itta Ohmura
  Keigo Nishida
  Hiroshi Koyama

Ohmura Ohno Nishida Komatsu



Koyama Y. Zhang Morimoto

- Taisho Pharmaceutical Co.
- Prof. Erik Lindahl and his colleagues, KTH
- JSPS / AMED / RIKEN Dynamic Structural Biology Project

- Companies

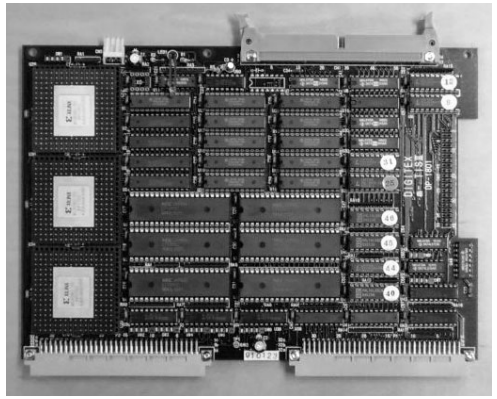
  D-Clue / Alchip / X-ability / Tokyo Electron Device / HOKS / Fujikura / Integran/
  Japan Cadence / Japan Synopsys

GRAPE(GRAvity PipE）：粒子系のための専用計算機

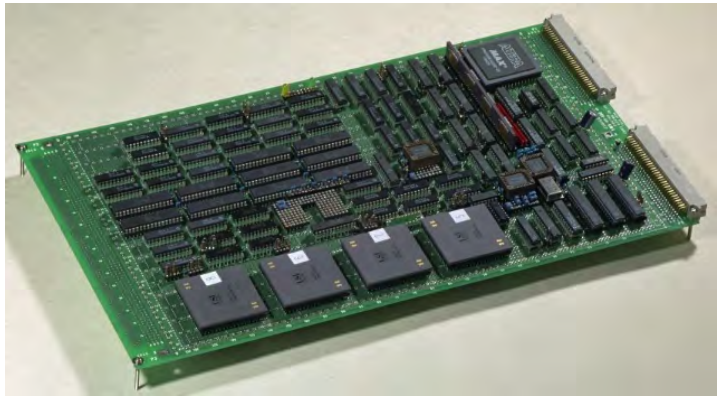
ヘテロジニアス計算機開発の先駆

m-TIS II (1991)

最初期のFPGAを用いた再構成可能計算機
（東大理）



MD-GRAPE (1996)

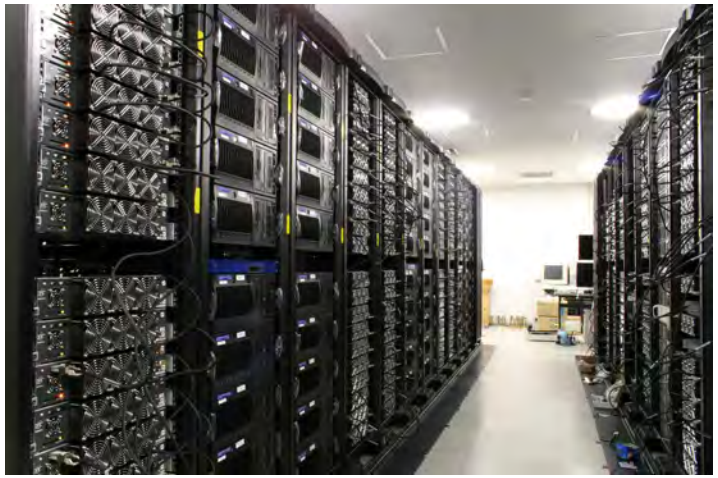


最初のASICベースMD加速器
（東大教養）

GRAPE-4 (1995)

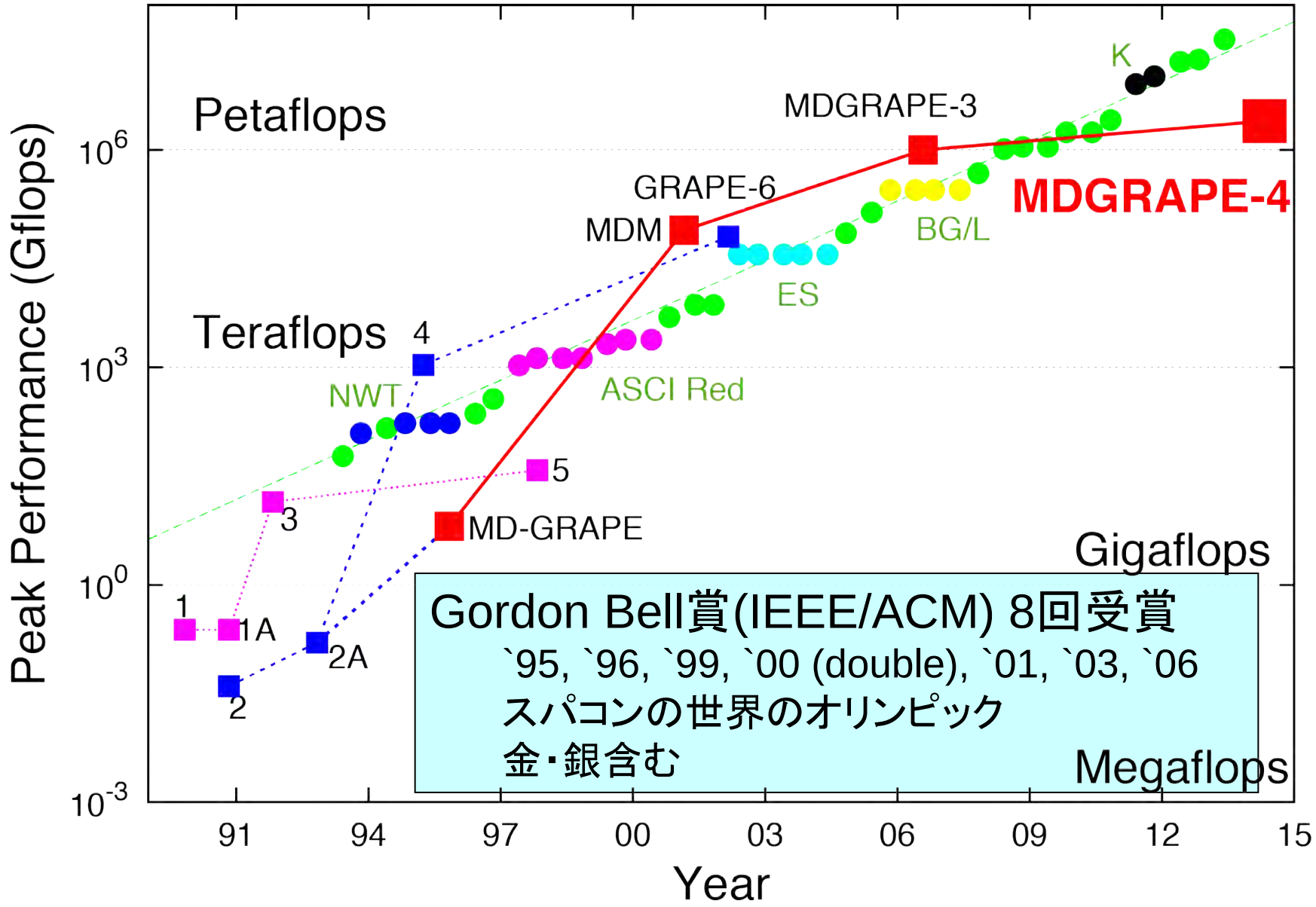
世界初のTFLOPS級計算機
(東大教養)

MDGRAPE-3 (2006)



世界初のPFLOPS級計算機
（理研）

GRAPEの歴史


Petaflops
Teraflops
Gigaflops
Megaflops
Peak Performance (Gflops)
Year
10^6
10^3
10^0
10^{-3}
91
94
97
00
03
06
09
12
15
K
MDGRAPE-3
MDGRAPE-4
GRAPE-6
MDM
BG/L
ES
NWT
ASCI Red
MD-GRAPE
1
1A
2
2A
3
4
5
Gordon Bell賞(IEEE/ACM) 8回受賞
`95, `96, `99, `00 (double), `01, `03, `06
スパコンの世界のオリンピック
金・銀含む

なぜ専用計算機を開発しているのか？

- 70年に亘る長期の発展を遂げたコンピュータであるが、近年壁が見えつつある

1. 並列度の上昇の問題
2. 消費電力の問題
3. ムーアの法則の終焉

専用化により、問題を大きく緩和可能

現在のコンピュータの課題：並列度の上昇

■CPUコアの速度が
頭打ちになり、並列化で
性能を上げる時代に

■最先端の並列計算機には、
100万以上のコアがある
▷100万人に指示を出して動かす
▷どうしても効率が悪くなり、実際の性能は100万倍にならない。

専用化により、並列度を上昇させることが可能

どうやって並列度を活かすか？

■問題の規模を十分大きくする

「人海戦術」が適用可能
基本的に同じ作業をさせることで性能向上

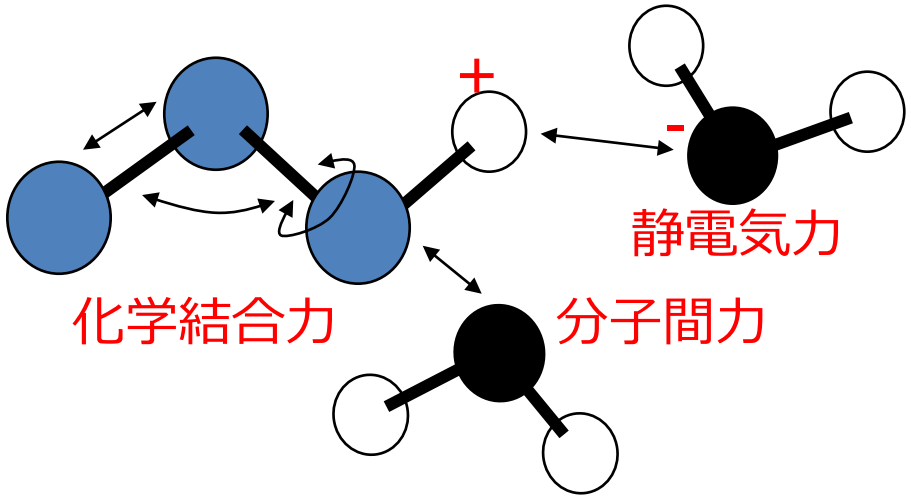
■問題を速く解く

高度なオーケストレーションが必要

⇨ 問題の性質に合わせた設計により使える並列度を100倍に

分子動力学シミュレーション

- 原子を質点であらわす
- 原子間の力を古典的な力で近似
- 運動方程式を積分して原子の運動を計算
- タンパク質の柔らかさを考慮できる
- 1ステップあたりの計算量はそこそこだが、
  時間を延ばすためには多くの計算コストがかかる



Bengt Nyman / CC-BY 2.0, trimmed

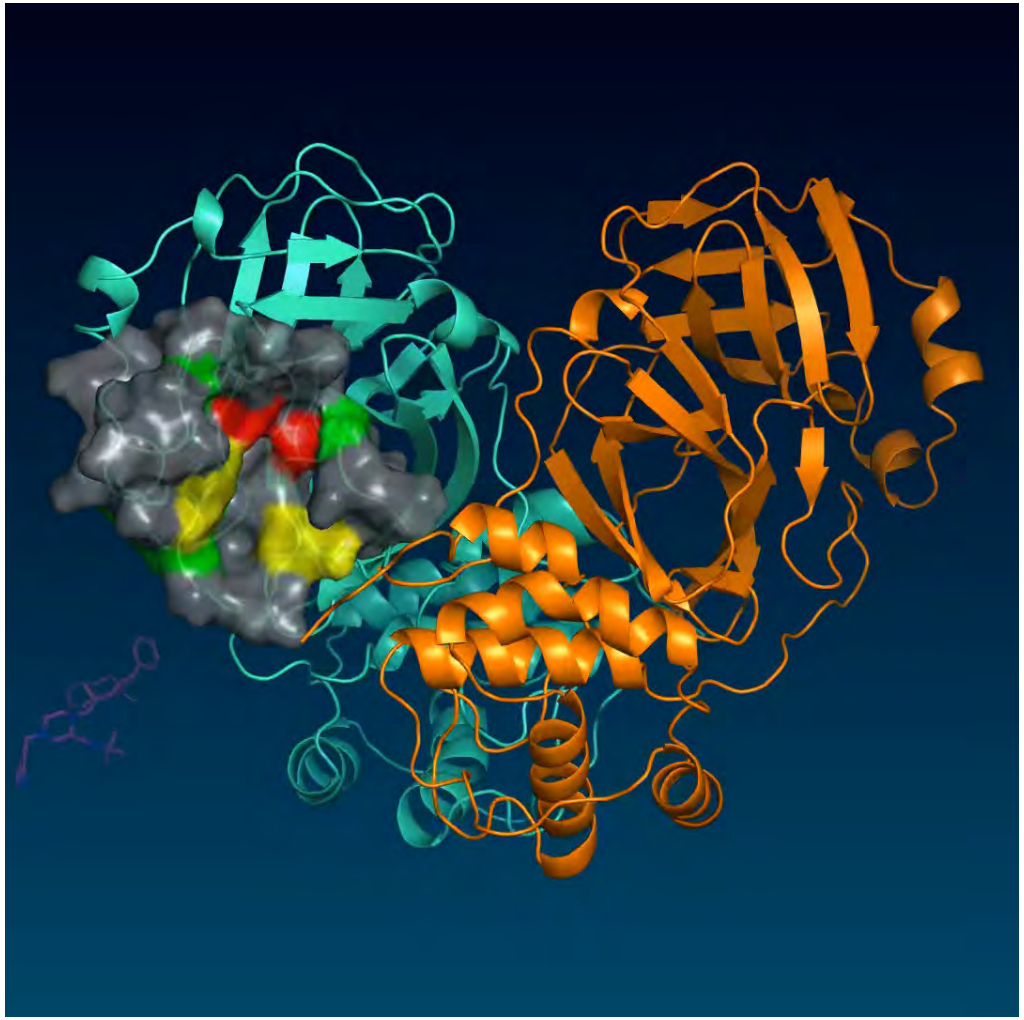
2013年ノーベル化学賞
Arieh Wershel
Michael Levitt
Martin Karplus

Example application of MDGRAPE-4A

- SARS-CoV-2 Main Protease (M^{pro}, 3CL protease)とligandの結合過程
  - ▷HIV-1 proteaseの阻害剤7種と SARS-CoV-2 M^{pro}.の結合過程をシミュレーション
  - ▷柔らかなポケット構造の解明

Dr. Komatsu

Dr. Okimoto

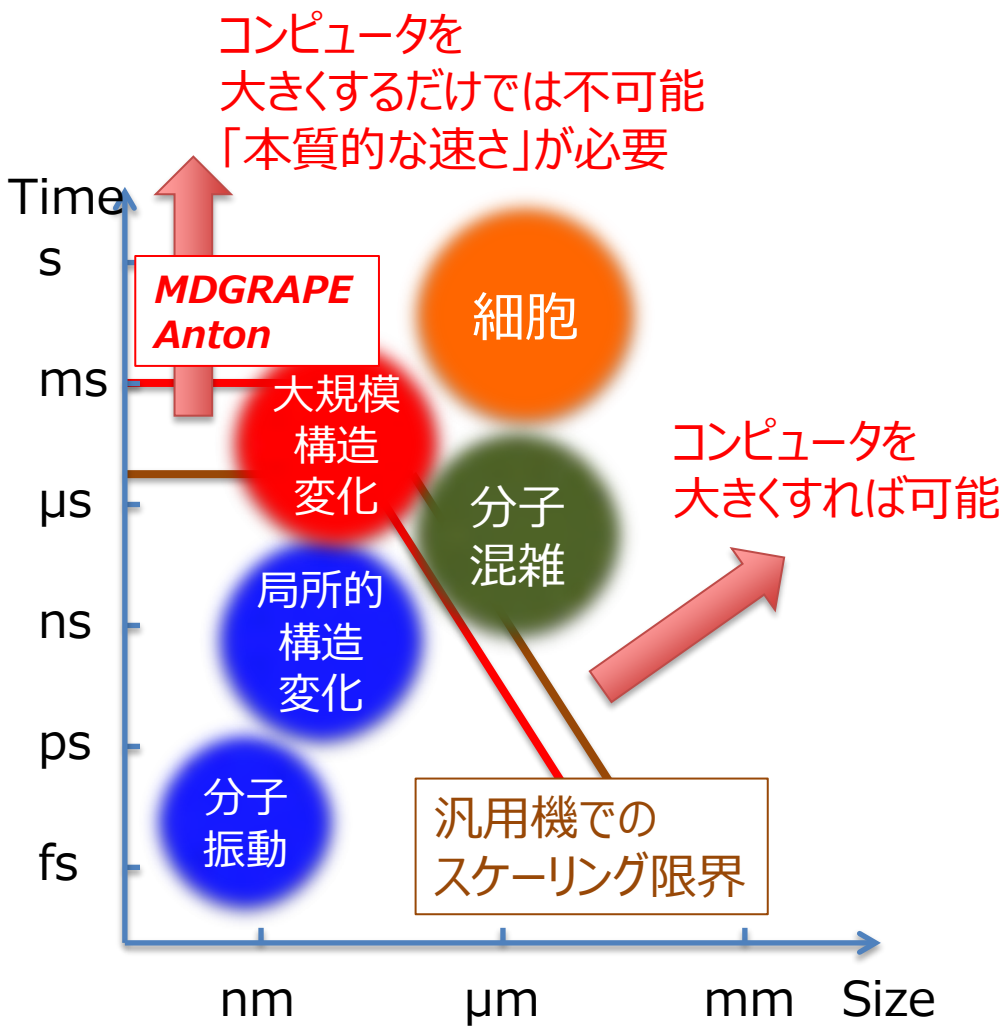


SARS-CoV-2 Main Protease (6LU7)
+ indinavir, 6 μs simulation

Komatsu T.S et al., *Sci. Rep.* **10**, 16986 (2020).
https://doi.org/10.1038/s41598-020-74099-5

Scaling Challenges in MD

MDは強スケーリング性能の問題に最初に到達した問題の一つ：
計算機のスケールだけでは問題を解決できない



- ～30,000 FLOP/particle/step
- System size : $N=10^5$
- 3 GFLOP/step
- 実効性能3TFLOPSの場合

  1msec/step = 200nsec/day

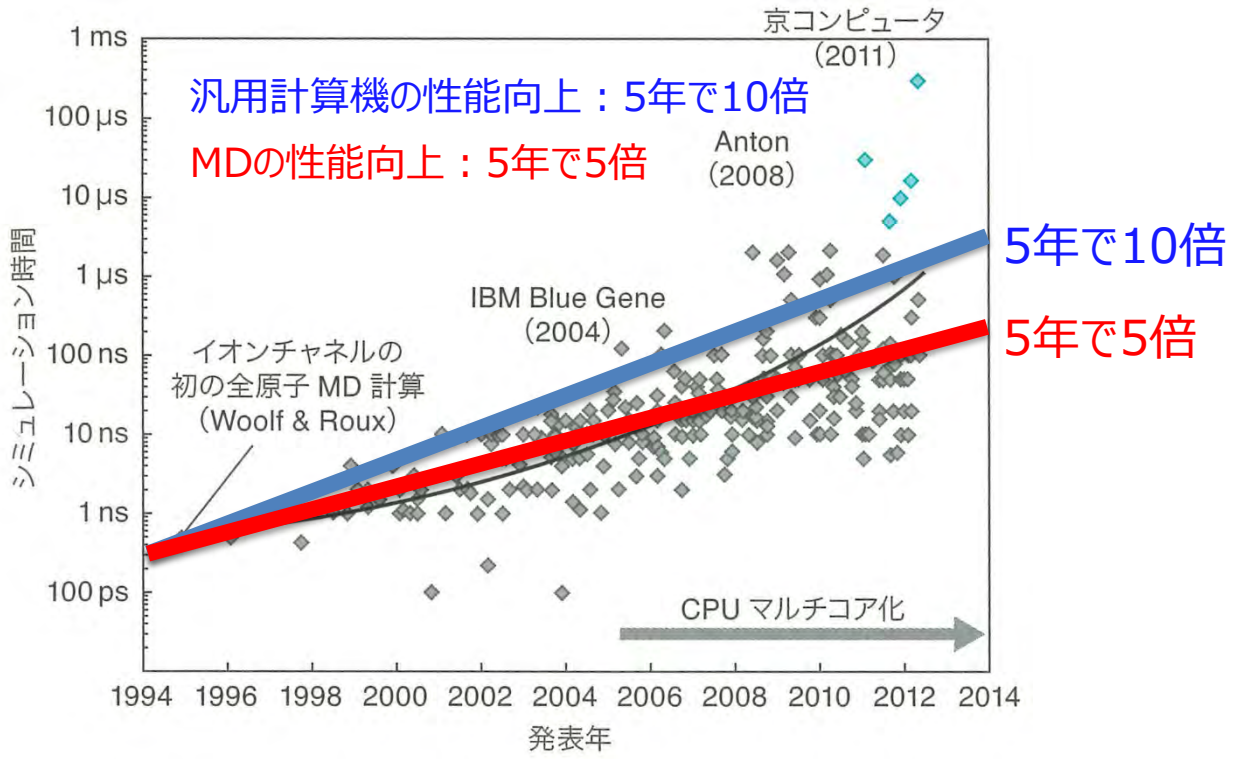
  Possible
- 実効性能3PFLOPSの場合

  1μsec/step = 200μsec/day???

  Difficult, but important

MDシミュレーションの場合

- 問題の規模は通常小さい
  典型的なタンパク質では水を含めて５万～20万原子
- 原子数＜コア数
- 通常の計算機（GPU含む）では、すぐに性能が頭打ちになってしまう
- 実効1PFlops（京の性能近く）の計算機があれば、一日に100μ秒近い計算ができるはずだが、実際には0.3μ秒程度が現在の限界



杉田有治他:スーパーコンピュータを用いた膜タンパク質の分子動力学シミュレーション、膜タンパク質構造研究、岩田想編、化学同人(2013).

Anton (D. E. Shaw 研究所)

- MD専用計算機
- 専用回路と汎用プロセッサ・ネットワークを統合
- Antonの基本要素とアイデアはMDGRAPEと同じ
- 統合システムの重要性を示した
- **Anton-1: ~20μs/step, ~10μs/日**
- **Anton-2: ~2μs/step, ~100μs/日**
- **Anton-3: ~200μs/日**
- **通常のスーパーコンピュータの1,000倍**

MDGRAPE-4Aの開発

- 2017～2020
- >マイクロ秒/日以上の専用計算機開発
  - ▷現在 ~1.1マイクロ秒/日　（10万原子系）：**通常のスパコンの約5倍**
  - ▷当初目標は8 μsec/dayでしたが、そこまで至らず…
- Anton類似のシステム化
  - ▷MDGRAPE-4: バグ等の問題あり
  - ▷改良版MDGRAPE-4Aを開発
    - – 多くの改良
    - – 新規アルゴリズム（改良版MSM）の開発
- やってみて明確になったこと
  - ▷**汎用機では、同等の性能を出すのはほぼ無理**
  - ▷1ステップ1ミリ秒台に突入したときの様々な並列化ボトルネックの解消：**専用化が必須**

MDGRAPE-4A system

MDGRAPE-4Aの新技術・アルゴリズム (1)

■ クーロン力長距離部分計算の新アルゴリズム

Tensor-structured Multilevel Ewald Summation Method
(TME, Morimoto, G. et al., Proc. Supercomputing 2021, accepted; Koyama, Y. et al., in preparation)

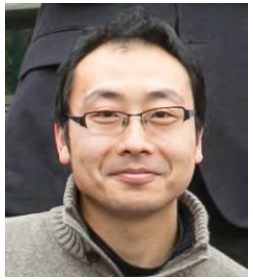
- ▷Multilevel Summation Method (MSM)の改良版
- ▷高次元コンボリューションが必要な多くのアルゴリズムに適用可能(CNN等?)

Dr. Koyama

■ 上記アルゴリズムのハードウェア化

- ▷「長距離力ユニット」の新開発：格子点への電荷割当・格子点からの力の計算専用ユニット
- ▷ネットワークへの一次元コンボリューションの埋め込み(特願2018-197597)
- ▷FPGAによる高速な16x16x16 FFTベースコンボリューションの実装

Dr. Morimoto



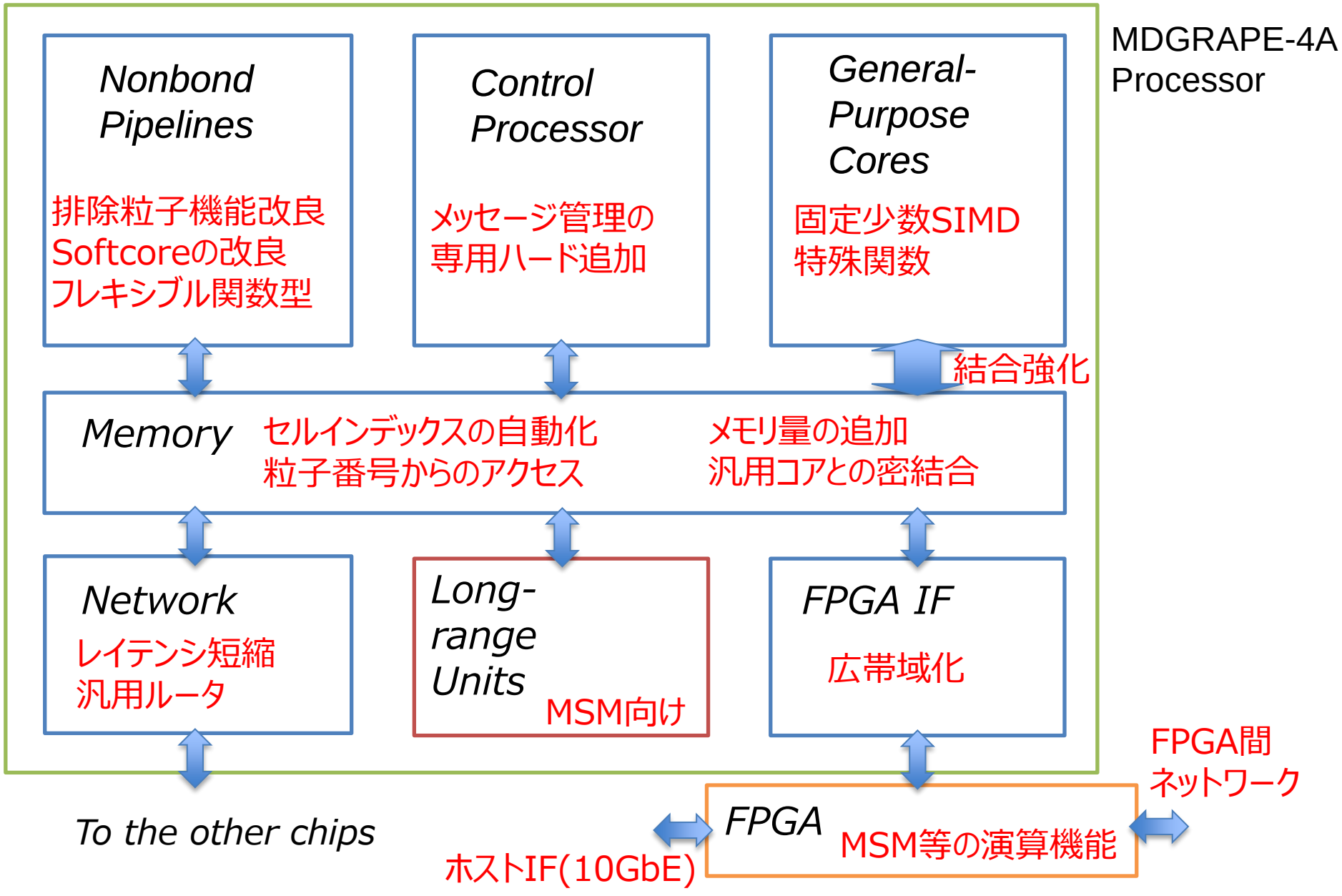
Dr. Zhang

MDGRAPE-4Aの新技術・アルゴリズム (2)

- クーロン・分子間力で結合ペアを排除するシンプルな方法
  - ▷タンパク質配列の特徴づけに基づく
    (特願2018-197597)
- 高速な３次元トーラスネットワーク
  - ▷1 hop 遅延 ~ 200 nsec
  - ▷MDGRAPE-4 (400 nsec) / 汎用ネットワーク(~1 μsec)に比べ高速
- RISC-Vベースの拡張汎用プロセッサ
  - ▷初のRISC-Vの大規模応用

Dr. Komatsu

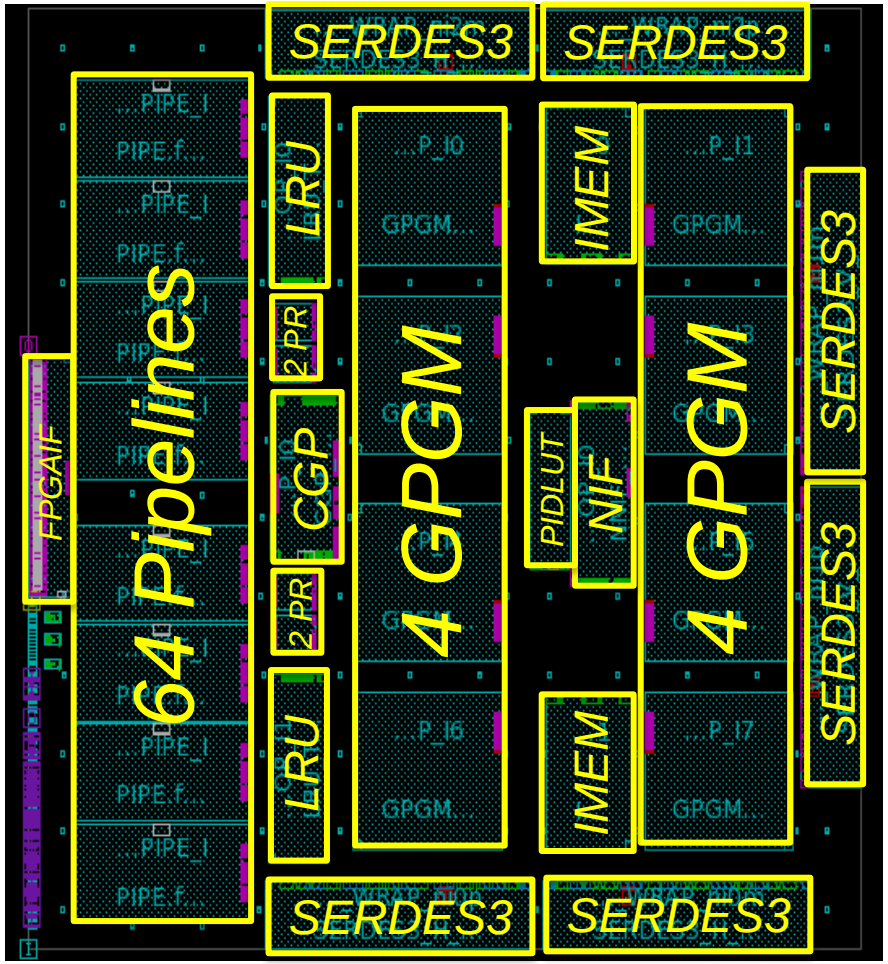
MDGRAPE-4A improvements



MDGRAPE-4A SoC

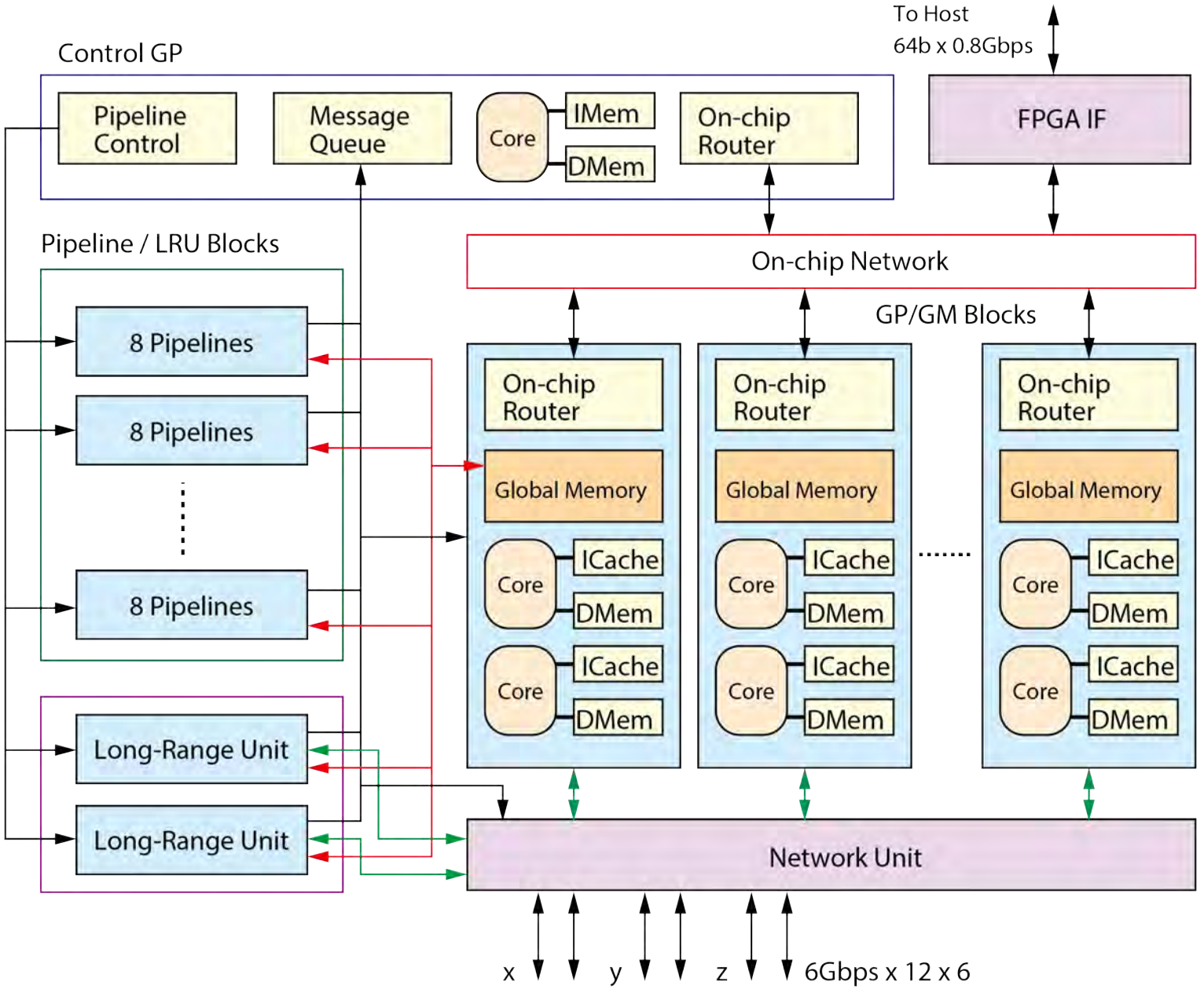
- MDGRAPE-4 SoCをほぼ1から設計しなおした
- Production：Alchip
- TSMC 40G (40 nm: 10年以上前のテクノロジ), 16.2mm X 16.2mm
- Logic ～ 50M Gates, SRAM ~ 72M bits
- 85W (worst case, 実際は 40～50W)
- 0.8 GHz (pipeline) / 0.6 GHz(その他)

- 大部分を理研で設計、一部外注



LSI Block Layout

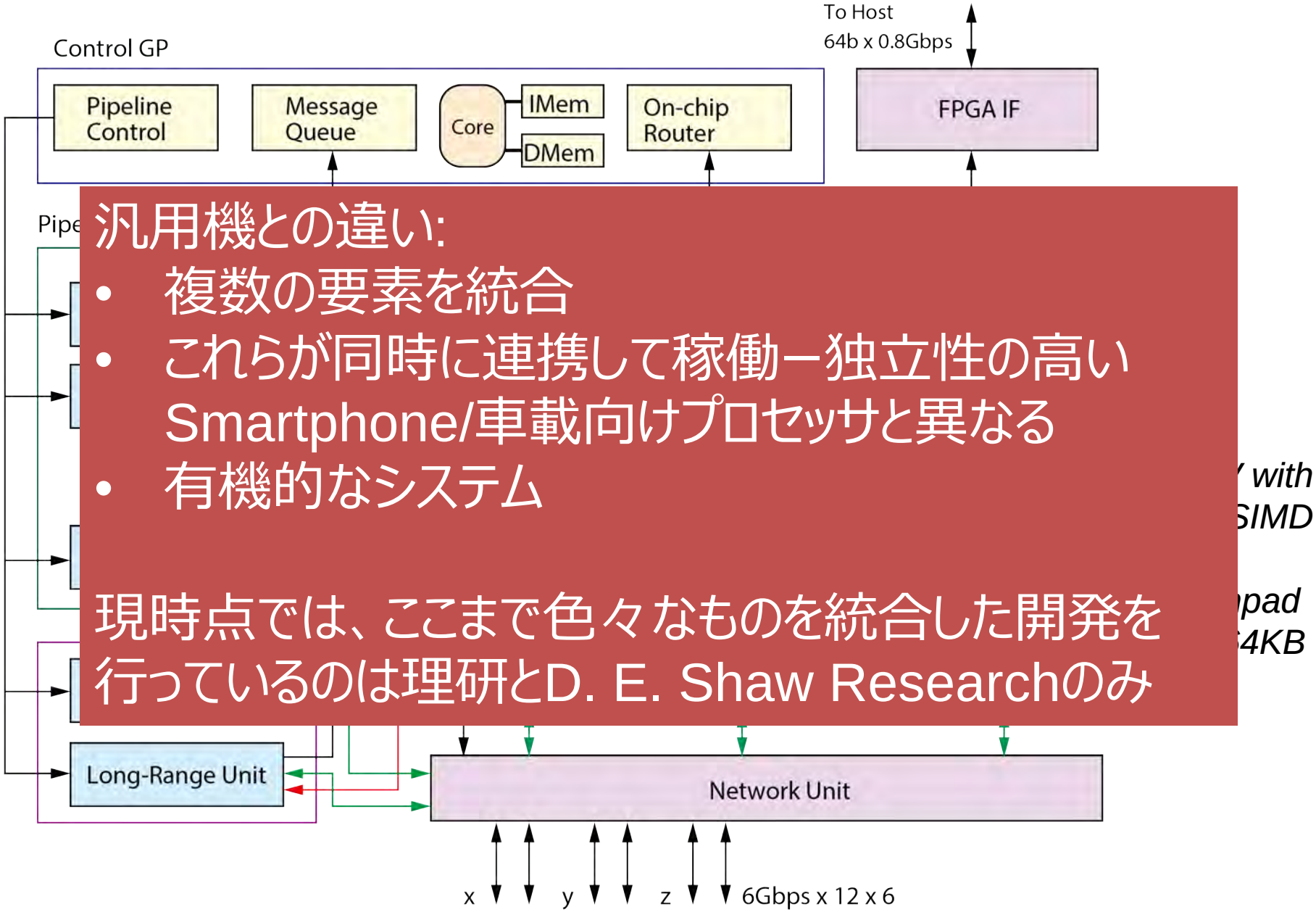
MDGRAPE-4A System-on-Chip



Core:
RISC-V with
4-way SIMD

Scratchpad
mem. 64KB

MDGRAPE-4A System-on-Chip



汎用機との違い:

- 複数の要素を統合
- これらが同時に連携して稼働ー独立性の高い Smartphone/車載向けプロセッサと異なる
- 有機的なシステム

現時点では、ここまで色々なものを統合した開発を行っているのは理研とD. E. Shaw Researchのみ

演算以外の専用化のポイント（１）

- メモリやネットワーク周辺の最適化が重要に
- **メモリでの計算**
  - ▷メモリでの固定小数点累算
- **メモリでの原子管理**
  - ▷セルへのatomicな原子の追加
  - ▷原子メモリのセルへの割当と解除
    - – 空き領域の管理と割り当て
  - ▷セル単位でのDMA read
    - – Pipeline, Network, GPのローカルメモリ
  - ▷２「ページ」の粒子管理テーブル: 粒子の移動時に便利

Lock動作が不要な並列化の実現

演算以外の専用化のポイント（２）

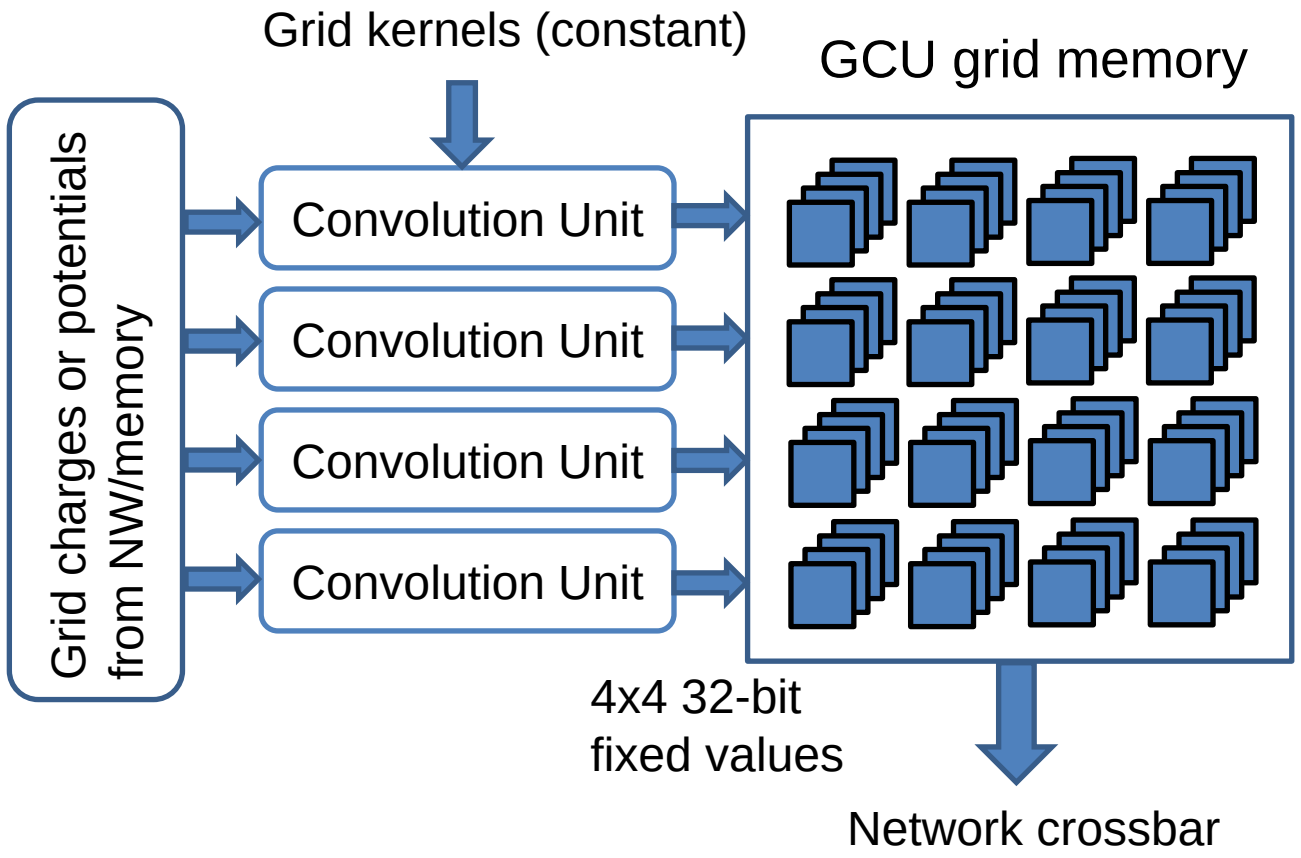
- Grid Convolution Unit の例
- 長距離力計算で、Convolution計算を行うユニット
- x,y,z軸を入れ替えつつ計算
- 行列の転置を行わず、

  並列に読み書きする面を

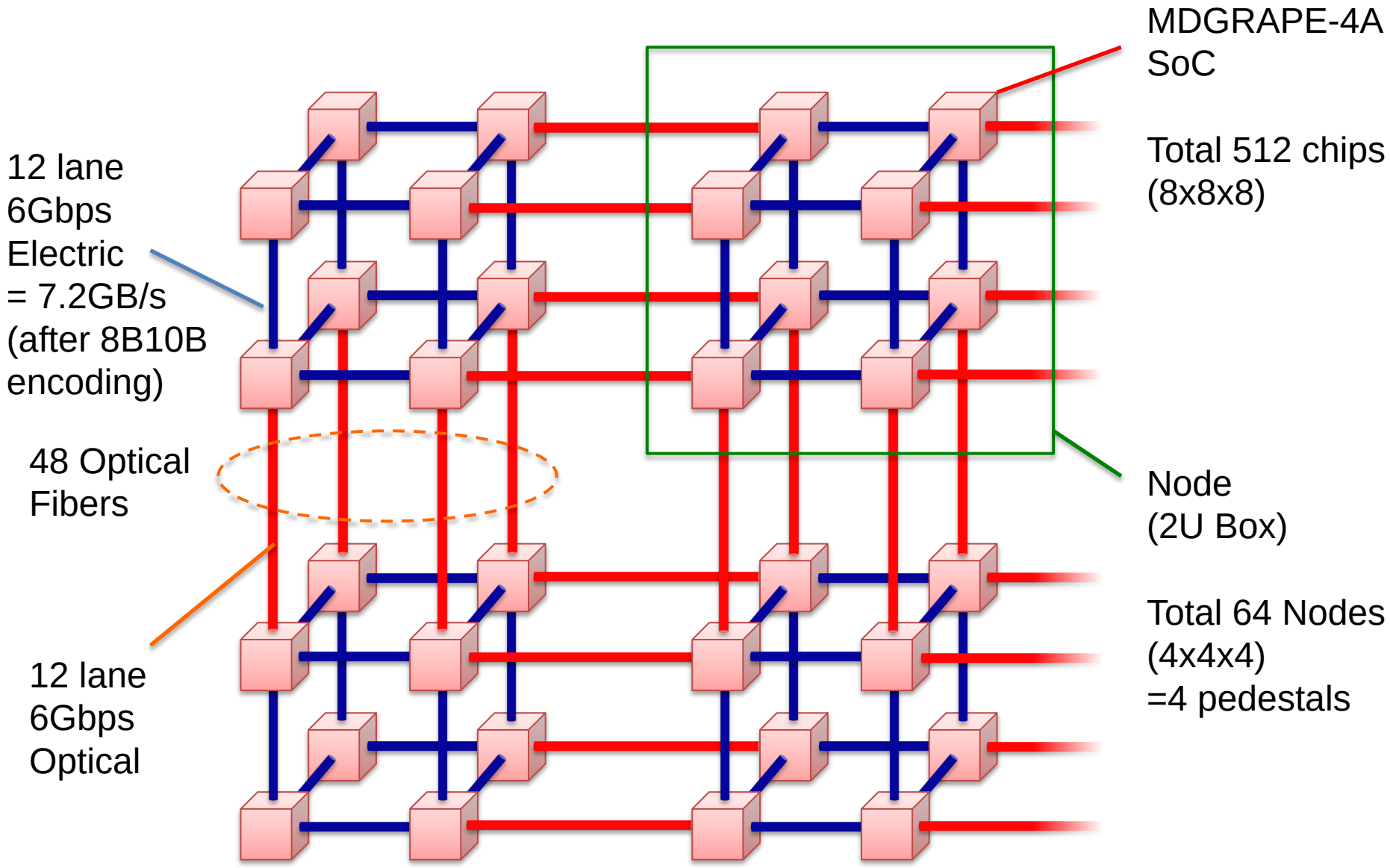
  xy, yz, zxで変更可能な

  専用メモリを設計

- 類似の専用化を各所で
  行っている

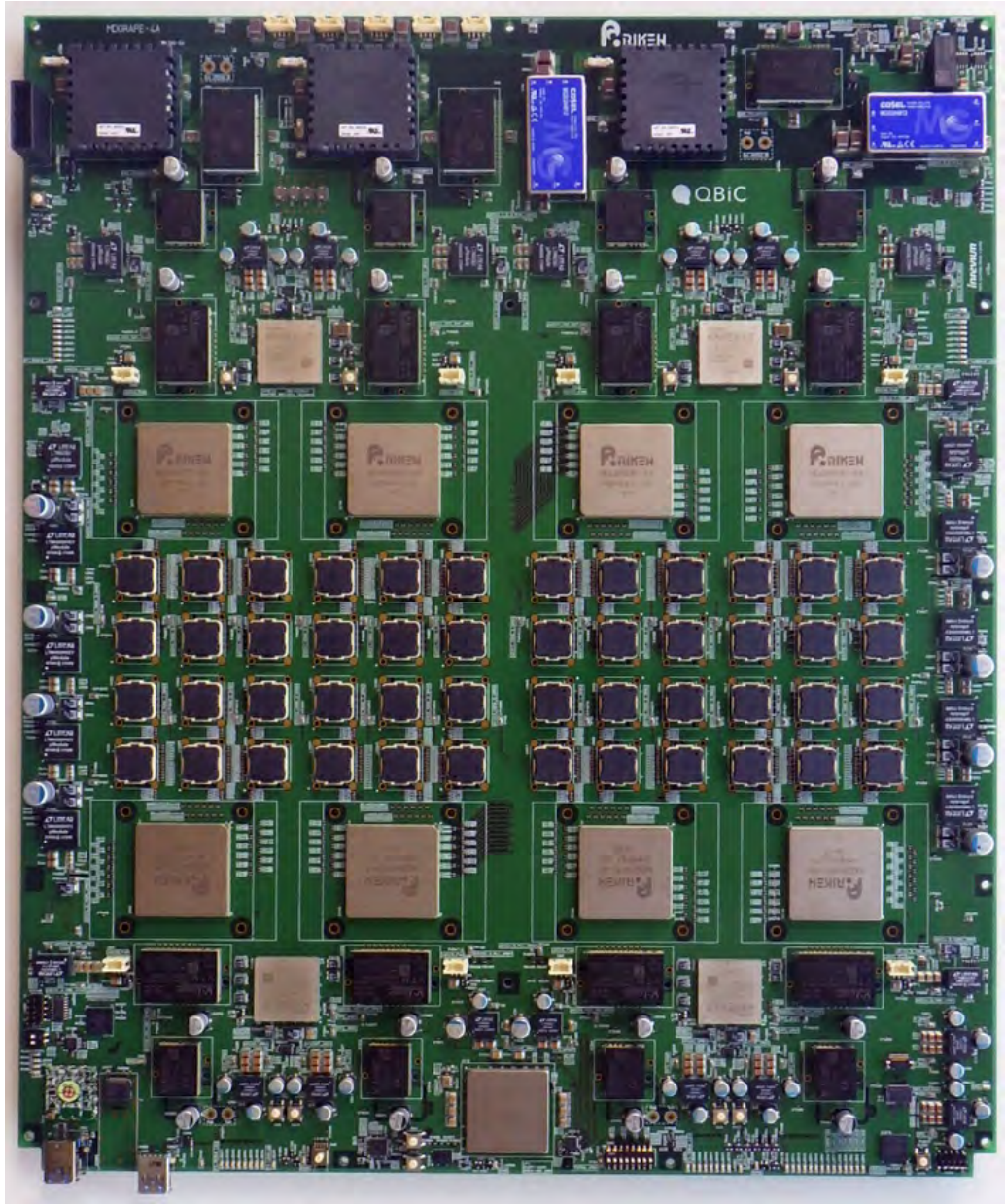


MDGRAPE-4A System



+ FPGA based oct-tree network (5GB/s duplex)

システム基板



- SoC 8個
- 28層
- 光モジュール
  TX/RX 各24個
- 通信・制御用
  FPGA 5個
- 電源

MDGRAPE-4A system

- 512 MDGRAPE-4A SoC
- ~1.3 PFLOPS equivalent
- 7.2GB/s 双方向トーラスネットワーク
- FPGAベースの８分木ネットワーク
  - ▷ 長距離力計算用, 5GB/s
- 10万原子系で1.1μs/day
  - ▷ 汎用スパコンの５倍程度
  - ▷ 少しバグが残り、性能が落ちてしまいました
- 40kW (実測値)
  - ▷ MN-3（Green500 Top1 as of 2020 May）: 21.1GFLOPS/W (実効値)
  - ▷ 当方実効性能を0.2PFLOPSとすると、 5GFLOPS/W
  - ▷ 10年前の技術とすればまずまず
  - ▷ 先端プロセスで作って、さらに実効性能を上げれば1TFLOPS/W程度の性能を実現可能

MDGRAPE-5の開発

- 次世代機MDGRAPE-5の予備的な開発を開始
  - ▷FPGAベースで2022年度中に完了を目指す
  - ▷FPGAでも汎用機を超えるスピードを達成したい
- 目標性能：
  - ▷>100 TFLOPS 等価演算性能/chip (System-on-Chipの場合)
  - ▷2～300 μsec/day by 64 chip system
  - ▷Single module with 8 chip
  - ▷SoCまで作る資金のあてはまだないですが…これまでの反省からも設計は先行させておきたい
- We plan to deliver an FPGA implementation also.
- 粗視化モデル(coarse-grained, CG models)への対応に挑戦

MDGRAPE-5での課題

一般的な課題として

1. 設計複雑度への挑戦：回路のパラメタ化
2. 細粒度並列化のハードウェア・ソフトウェアによる支援

これらは、今後の高性能計算において
専用ハードウェアのみならず
汎用機・準汎用機においても重要な課題

MD specificな課題として

3. Bond Engineによる結合力計算・積分の加速
4. 粗視化モデルへの対応
5. Lennard-Jones Ewald (分子間力の周期境界計算)への対応
6. その他CMAPの実装等

MDGRAPE-5: 設計複雑度への挑戦

- More specialization for more speed:
  - ▷専用部の増加に伴い、設計複雑度が上昇
    - MDGRAPE-4: MDGRAPE-3の~3倍
    - MDGRAPE-4A: MDGRAPE-4の~1.3倍
  - ▷汎用プロセッサよりも複雑な面も
- 設計複雑度への挑戦
  - ▷Anton-3では、同じもののタイル化でsimplify
  - ▷別のアプローチも可能か？
    - **設計のパラメタ化と最適化**

RTL line counts

PIPE	14,210	
LRU	10,085	
PIPERELAY	1,447	
FPGAIF	7,212	
GPGM	17,438	w/o GP core
CGP	9,675	w/o GP core/GMNW
NNIF	12,075	
SERDES3	6,400	
PIDLUT	1,435	
IMEM	803	
TOP	3,635	w/o slicer
Total	84,415	

Since MDGRAPE-3

Since MDGRAPE-4

Since MDGRAPE-4A

MDGRAPE-5における設計手法の革新

新しいハードウェア記述言語ChiselによるFlexibilityの実現

- ▷Chisel: Scalaをベースにしたhdl
- ▷現在、全てChiselで書き直し中
- ▷RISC-V Rocket / Boomの利用
- ▷Chisel言語の場合、パラメタによる回路ジェネレータを書きやすい。
  原理的には、Verilog/SystemVerilogのgenerate文でもできるが…

回路記述	回路記述の制御
Verilog 主	generate
Chisel	Scala 主

- ▷パラメタ化により、設計空間の自動探索が可能になる。ベイズ最適化などの手法によるハードウェア自動最適化を試したい。

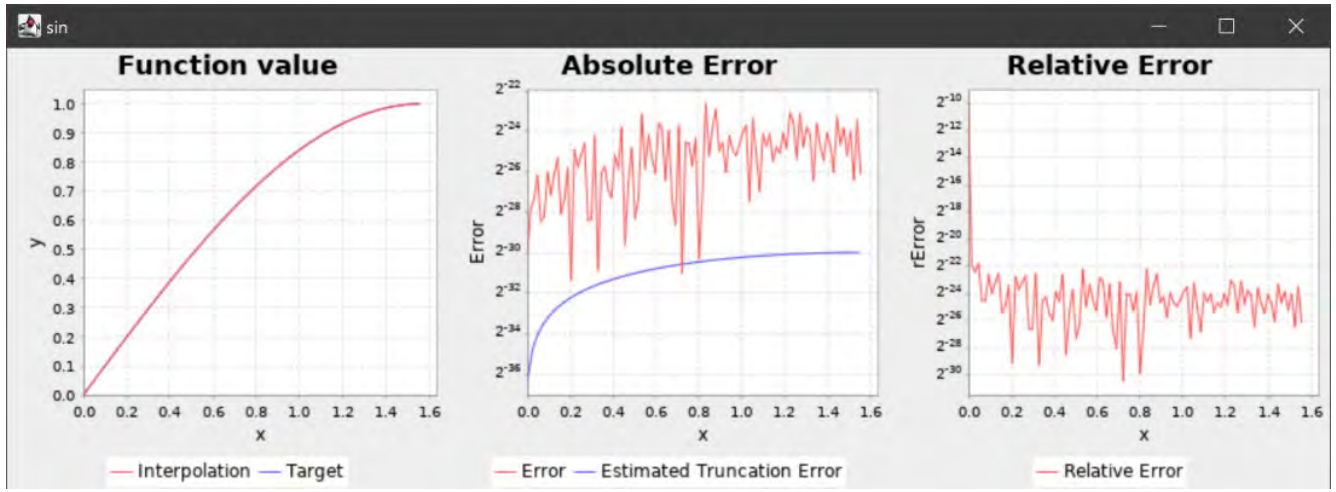
Chiselによる例：関数評価器

- 区分多項式による関数の近似値評価回路を自動生成
  - ▷ Chisel3の機能を活用
- 任意の（なめらかな）関数に適用可能
- 関数と精度、分割区間を与えるのみで自動生成、誤差評価

```
val xMin = 0.0
val xMax = 0.5*Pi
val nDiv = 64
val bp   = 24
val cbit = Array(bp+2)

val t = new FuncTableDouble( sin, 3 )
t.addRange(xMin, xMax, nDiv )
val tf = new FuncTableFloat( t, true, cbit )
```

これだけでsin関数の3次式近似
0~π/2 64区間のテーブル作成が
可能



結果の確認と誤差の表示

細粒度並列化のためのハードウェア

- まずMDの計算を、マイクロタスクに分割
  - ▷セル間の非結合力計算
  - ▷結合力計算
  - ▷セル単位でのCharge assignments など
- これらをデータフロー的に処理
  - ▷ネットワーク・演算器からのメモリ書き込みイベントを基にハードウェアによりイベント駆動
  - ▷Cache injection等のpush型処理も検討

- 伝統的なGRAPEは粗粒度の力計算のハードウェア加速が中心だったが、**現在のMD向け開発は細粒度並列化の支援が中心的課題**。演算量の多くは未だに専用部に依るが、**単純な演算の高密度化のみではもはや加速が難しい**。

細粒度並列化のためのソフトウェア

- 汎用機、専用機双方で動くソフトウェアを維持しやすい環境を整備
  - ▷専用機のモデルに近い、非同期プログラミングモデルを導入
- **メタプログラミング**アプローチを検討中
  - ▷メタプログラミングにより、
    - –専用機でのタスク割り当てに必要な計算グラフ・通信パターン
    - –専用機汎用部でのコード生成
    - –汎用機上でのコード生成

    を行う
  - ▷プログラミングが容易に（デバッグに課題あり）
  - ▷専用機だけでなく、他のヘテロジニアス環境でも有効かも
  - ▷将来的には、ハードウェア生成も

Evolution of MD processors

	MD-GRAPE	MDG2	MDG3	Anton-1	MDG4	Anton-2	MDG4A	MDG5
Year※	1995	2000	2004	2009	2013	2013	2018	2022?
Performance	0.8GF	16GF	200GF	2.73TF	2.5TF	12.7TF	2.5TF	50TF
	50MHz	0.1GHz	0.25Hz	0.97GHz	0.8GHz	1.65GHz	0.8GHz	1.28GHz
Pipelines	1	4	20	66	64	152	64	800
Long-range	Ewald	Ewald	PME	GSE	GSE	u-series	MSM	MSM
Memory			○	○	○	○	○	Enlarge
Network				○	○	○	○	NoC integration
GP processor				○	○	○	○	+Bond engine
Soft-core					△	?	○	○
in-memory op				?	○	○	○	○
Exclusion				△	△	○	○	○
Event management				?	Queue	○	△	Enhanced
Particle management				?		?	○	○

※For Anton, year of system release; For MDGRAPE, year of chip release

今後の開発に向けた課題

- Antonとの差異化
  - ▷両方とも普通の技術のシステム化で作っているため、難しい
    - 例えばAnton-3では結合力の一部専用化を行っている。MDGRAPE-5で準備していたものと同じ方向。設計複雑度の話もAnton-3では強調されていた。
  - ▷予算規模の差：オープン化も可能だが、逆に投資を難しくする可能性もあって踏み切れないでいる。
  - ▷作ってみて予想以上に大変だったが、見通しが立った段階で捨ててしまうのももったいないと感じているところ。Antonの論文には書いてないことも多く対応する必要があり、積み残しもあるが技術的な蓄積はできた。
  - ▷粗視化モデルへの対応などでの差異化
- 準汎用機への展開
  - ▷細粒度並列化のHardware・Software支援は、一般的に有効
  - ▷Chipletなどもありうる
- その他のアプリケーションの加速
  - ▷量子化学・データ科学関係など

古典分子シミュレーションのゴール

~2028（既にAnton-3では近いですが…）

- ~msec simulation of 100K atom / day

  ~0.5 μsec/step, 5fs/step

- 1msec simulations X 1,000

  ▷累積 1秒のシミュレーション

  ▷複合体形成や化合物結合後の構造変化など、
  多くの遅いダイナミクスを解明

- 100万原子系の10 msシミュレーション

  ▷細胞のレオロジー

- 細胞システムの粗視化シミュレーション

- MDは細胞構造生物学のための基本的なツールになりうる

Summary

- MDシミュレーション専用計算機MDGRAPE-4Aを開発した
  - ▷現時点での性能： ~ 1μs / 日（10万原子系）
  - ▷多くの新規ハードウェア要素を開発し、曲がりなりにも全体を稼働させることができた。
- 次世代機MDGRAPE-5の開発に着手
  - ▷Anton-3をターゲットとした開発
  - ▷FPGAでの実装から開始
  - ▷回路パラメタ化による設計複雑度への挑戦
  - ▷細粒度並列化のハードウェア・ソフトウェアによる支援
  - ▷粗視化モデル等の新機能の拡充
- 技術移転の促進
  - ▷若手の育成、国内外で技術の受け取り先の探索