資料1

NGACI：次世代先端的計算基盤の開発に向けたコミュニティ活動の紹介

慶應義塾大学理工学部
理化学研究所計算科学研究センター（兼務）

近藤 正章

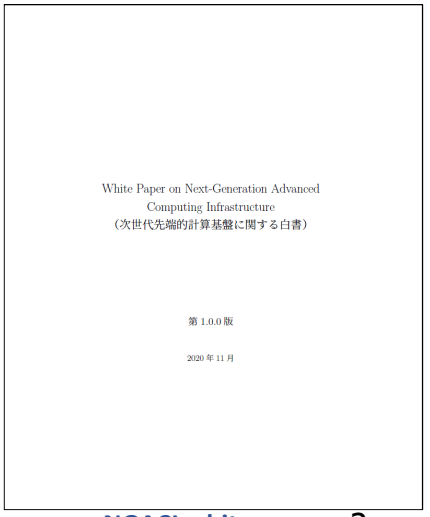
NGACI活動の紹介

- NGACI: Next-Generation Advanced Computing Infrastructure
  - 概要と活動目的

今後の高性能計算機の持続的な発展を考えるにあたり、AIやビッグデータ技術とのさらなる融合、Society5.0といった新しい応用分野への展開など、さらなる発展も期待されますが、ムーアの法則の終焉など多くの技術的課題が待ち受けていることも事実です。本活動(NGACI)は、将来の高性能計算環境として、また共用計算機資源としてどのような技術的課題があり、どのような研究開発が必要なのか、コミュニティーとしてどのような活動をしていくべきなのかなどに関して、オープンに意見交換をしつつそれをWhite Paperとしてまとめることで本分野の発展に寄与することを目的としています。

https://sites.google.com/view/ngaci/home

  - これまでの実績
    - 本活動に登録して頂いているコミュニティのメンバー数：104人
    - 7回の全体ミーティングと3回のセミナーを実施
    - 4つのWGにより将来のシステム像や課題を集中的に議論
      - アーキテクチャWG、システムソフトWG、アプリ/ライブラリWG、システム運用WG
  - White Paperについて
    - 1.0.0版（164ページ）を公開中（https://sites.google.com/view/ngaci/home）


White Paper on Next-Generation Advanced Computing Infrastructure
(次世代先端的計算基盤に関する白書)
第 1.0.0 版
2020 年 11 月

NGACI white paper

White Paperの執筆協力者

所属等は2020年11月時点

- 取りまとめ: 近藤（東大・理研）
- アーキテクチャWG
  - WGリーダ: 三輪（電通大），佐野（理研），谷本（九大）
  - WGメンバ: 安島（富士通），井口（北陸先端大），井上（九大），江川（電機大），岡本（Spin Memory） 小野（九大），鯉渕（NII），児玉（理研），小林（筑波大），小松（東北大），佐藤（東北大），塩見（京大），田邊（東大），中里（会津大），吉川（富士通研），福本（富士通研），星（NEC），三好（わさらぼ），宮島（理研）
- システムソフトWG
  - WGリーダ: 佐藤（理研），佐藤（豊橋技科大）
  - WGメンバ: 合田（NII），遠藤（東工大），小柴（理研），小松（東北大），坂本（東大），高野（産総研），滝沢（東北大），辻（理研），ゲローフイ（理研），中島（富士通研），深井（理研），山本（理研），和田（明星大）
- アプリケーション・ライブラリ・アルゴリズムWG
  - WGリーダ: 深沢（京大），今村（理研），中島（東大・理研）
  - WGメンバ: 岩下（北大），小野（九大），笠置（富士通研），片桐（名大），白幡（富士通研），住元（富士通研），高橋（筑波大），寺尾（理研），長坂（富士通研），椋木（理研），村上（都立大）
- システム運用WG
  - WGリーダ: 塙（東大），野村（東工大）
  - WGメンバ: 大島（名大），實本（理研），庄司（理研），滝澤（産総研），竹房（NII），藤原（NII），三浦（理研）

White Paperの章構成

2028年頃に実現可能な次世代システムの予測

- 次世代システムの構成としていくつかのアーキテクチャタイプを検討
  - 汎用システム型
    - メニーコアCPU型：富岳の構成の延長として考えられるシステム
    - メニーコアCPU＆GPU混載型：GPUとホストCPUで構成（現在多くのシステムでも採用）
    - ベクトルプロセッサ混載型：ベクトルプロセッサとホストCPUで構成（例： SX-Aurora TSUBASA ）
  - 専用システム混載型（ムーアの法則減速により重要な検討事項に）
    - CPU拡張型
      - ISA（SIMD）の専用的な命令をCPUに拡張機能として搭載（例：Intel AMXやARM SVEのFMMLAなど）
      - BFloat16やINT8、INT4などの応用に特化したデータ型の導入
    - アクセラレータ主体型/ヘテロジニアス型
      - システム搭載方式：チップ内拡張(SoCやMCM)、ノード内拡張、ラック間疎結合
      - アクセラレータ構成方式：専用、準専用、汎用
    - Processing-In-memory主体型
      - 演算器とメモリの密接実装によるメモリアクセスの高バンド幅化と低遅延化
    - 新計算原理の混載

汎用システム型の性能予測方法

- システムコンポーネント毎に以下の文献データから予測
  - プロセッサ：IRDS Roadmap - Systems and Architectures (2017 and 2020 edition)
    - ソケットあたりコア数: 70コア, SIMDビット長：2048-bit x 2, クロック周波数：3.9GHz
    - CPUソケットのTDP: 351W
  - GPU
    - 保守的な予測：NVIDIA社の過去のハイエンドGPUの性能をもとに線形で外挿
    - 積極的な予測：将来のCPUの性能予測値に現行のGPU/CPUの性能比を乗じることで予測
  - ネットワーク：“Ethernet Alliance Roadmap 2018”
    - リンクあたり1.6 Tbyte（100Gbps x 16レーン）
    - ノードあたり1リンク（リンク数増加によってアプリのカバー範囲が変わらないため）
  - ストレージ：“Lustre: The Next 20 Years”, HPC-IODC Workshop, 2019.
    - LustreでI/O性能が1.36x/年、容量が1.38x/年で向上するとの予想を利用
- 制約：システム全体の電力
  - 3種類のシステム電力バジェット：30, 40, 50MW（cf. 富岳では28.3MW）および PUE=1.1
  - 3種類のCPU（あるいはGPU）の電力バジェットの比率：60, 70, 80%

2028年のメニーコア型システムの予測性能

- 最も積極的な予測でも最大1.8 EFLOPS (富岳の性能の3.37倍)

	30MW			40MW			50MW		
	60%	70%	80%	60%	70%	80%	60%	70%	80%
ソケット数	46620	54390	62160	62160	72520	82880	77700	90650	103600
総コア数	3.3×10^6	3.8×10^6	4.4×10^6	4.4×10^6	5.1×10^6	5.8×10^6	5.4×10^6	6.3×10^6	7.3×10^6
PFLOPS	815	950	1086	1086	1267	1448	1358	1584	1810
DDR 総 BW (PB/s)	102	120	137	137	160	182	171	200	228
HBM 総 BW (PB/s)	307	358	410	410	478	547	512	598	683
DDR 総容量 (PB)	17	20	23	23	27	31	29	34	39
HBM 総容量 (PB)	4	5	5	5	6	7	7	8	9
インジェクション BW (Tb/s)	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6
総 I/O 性能 (TB/s)	34	34	34	34	34	34	34	34	34
総ストレージ容量 (EB)	3.45	3.45	3.45	3.45	3.45	3.45	3.45	3.45	3.45

← システムの電力制約の仮定
← CPUの電力バジェット

参考）富岳の諸元

158,976
8.3×10^6
537
—
163
—
4.85
0.33

2028年のGPU混載型システムの予測性能

- 最も積極的な予測で最大18.0 EFLOPS (富岳の性能の33.5倍)

保守的な予測の場合（NVIDIA GPUの性能トレンドから外挿）

	30MW			40MW			50MW		
	60%	70%	80%	60%	70%	80%	60%	70%	80%
GPU 数	50661	59104	67548	67548	78806	90064	84435	98508	112580
総コア数	5.3×10^8	6.2×10^8	7.1×10^8	7.1×10^8	8.3×10^8	9.4×10^8	8.8×10^8	1.0×10^9	1.2×10^9
PFLOPS	1279	1492	1706	1706	1940	2474	2132	2487	2843
HBM 総 BW (PB/s)	91	107	122	122	143	163	153	178	204
HBM 総容量 (PB)	1	1	2	2	2	2	2	3	3

積極的な予測の場合（CPUとの性能比のトレンドから外挿）

	30MW			40MW			50MW		
	60%	70%	80%	60%	70%	80%	60%	70%	80%
GPU 数	50661	59104	67548	67548	78806	90064	84435	98508	112580
総コア数	3.4×10^9	3.9×10^9	4.5×10^9	4.5×10^9	5.2×10^9	6.0×10^9	5.6×10^9	6.5×10^9	7.5×10^9
PFLOPS	8083	9431	10778	10778	12574	14371	13472	15718	17963
HBM 総 BW (PB/s)	334	390	445	445	520	594	557	650	743
HBM 総容量 (PB)	4	5	6	6	7	8	8	9	10

アクセラレータの構成法式

- 専用アクセラレータ
  - 特定の計算問題または計算ドメインのみを高速に処理可能な構成方式
  - 最も高性能・高電力効率であるが、柔軟性やプログラマビリティが低い
  - 例）ディープニューラルネットワークアクセラレータ、FFTアクセラレータ
- 準専用アクセラレータ
  - 複数の計算モデル・計算問題を高速に処理可能＋ある程度のプログラマビリティを持つ構成
  - 専用アクセラレータに比べ若干性能や電力効率で劣る
  - 例）データフロープロセッサ、コンボリューションプロセッサ
- 汎用アクセラレータ
  - 特定のアクセラレータを再構成可能デバイス上に構成する方式
  - 例）FPGA, Coarse-Grain Reconfigurable Arrays (CGRA)
- 特定計算ドメインの候補
  - 疎行列計算、N体問題、FFT、ソート、グラフ処理、脳型計算、量子計算シミュレーション、エージェントシミュレーション

汎用 vs. 演算加速機構（アクセラレータ）

- ターゲットアプリケーションが複数＋各アルゴリズムも日々改良

  →特定ドメインで優位性を発揮できる一方で、様々な処理を実行可能な「広義」のアクセラレータ を前提とすべき

  – 性能優位性とコストのトレードオフを考慮が必要
  – プログラミング生産性が重要
  – 特定ドメイン向けの専用システム開発とは異なるアプローチが必要
- 技術的な課題
  – ホストCPUとの役割分担
  – アクセラレータのメモリ階層の設計
  – 共有メモリ空間の見せ方とメモリー貫性維持方式
  – ホスト・アクセラレータ間やアクセラレータ間のネットワークトポロジ
  – アクセラレータのプログラミングモデル、デバッガやプロファイラの実装
  – :

アプリケーションの要求性能分析

- 計算科学ロードマップやアンケートに基づき37個のアプリの要求性能を解析
  - そのうちノード数の要求について記載のないものは除く
  - より多くのアプリ（重点課題やビッグデータアプリ）の分析は今後の課題
- 分析の目的
  - 性能要求の分析により必要なシステムのタイプを分類
  - 汎用的なシステム構成によりどの程度のアプリケーションがカバーできるかの調査
- 分析の際に仮定するシステム構成（メニーコア型・GPU混載型の積極的な予測の場合）

	Manycore (50MW, CPU80%)	GPU (50MW, CPU80%)
# of CPU Sockets or GPUs	103,600	112,580
# of total cores	7,252,000	1.2×10^9
PFLOPS (double)	1,810	17,963
DDR total BW (PB/s)	228	—
HBM total BW (PB/s)	683	743
Total Size of DDR (PB)	39	—
Total Size of HBM (PB)	9	10

各アプリケーションの要求性能との比較

- メニーコア型システム（〇および×は各コンポーネントが要求を満たしているかどうかを表す）

App. Area	Name of Application	Node	CPU	Memory	Interconnect	Storage	Total
素粒子・原子核	(unknown)	〇	〇	〇	〇	〇	〇
	(unknown)	〇	〇	〇	〇	〇	〇
	(unknown)	〇	〇	〇	〇	〇	〇
	rmcsm	〇	〇	〇	〇	〇	〇
	(unknown)	〇	〇	×	×	〇	×
物質科学	HPhi	×	〇	〇	〇	〇	×
エネルギー・資源	NTChem	〇	〇	×	〇	〇	×
	SMASH	〇	〇	×	〇	〇	×
	paraDMRG	〇	〇	〇	〇	〇	〇
	GELLAN	〇	〇	〇	〇	〇	〇
	MODYLAS	〇	〇	〇	〇	〇	〇
脳科学・AI	WHC	〇	〇	〇	〇	〇	〇
	Realtime cerebellum	〇	〇	〇	〇	〇	〇
	NEURON K+ Stochastic	〇	〇	〇	〇	〇	〇
	CNN (Forward & Back-prop)	〇	×	×	〇	〇	×
	CNN (Forward)	〇	×	×	〇	×	×
地震・津波	GAMERA	×	〇	〇	〇	〇	×
気象気候	NICAM	×	〇	〇	〇	〇	×
	SCALE-RM	〇	〇	〇	〇	〇	〇
	CHASER-LETKF	O	〇	×	×	×	×
	NICAM-LETKF	×	〇	〇	〇	×	×
宇宙・天文	GreeM	〇	〇	〇	×	〇	×
	(unknown)	〇	×	〇	〇	〇	×
	EM-PIC	〇	〇	×	〇	×	×
	P3T	〇	〇	〇	〇	〇	〇
	AmaTeRAS	×	〇	〇	×	〇	×

各アプリケーションの要求性能との比較

- GPU混載型システム（○および×は各コンポーネントが要求を満たしているかどうかを表す）

App. Area	Name of Application	Node	CPU	Memory	Interconnect	Storage	Total
素粒子・原子核	(unknown)	○	○	○	○	○	○
	(unknown)	○	○	○	○	○	○
	(unknown)	○	○	○	○	○	○
	rmcsm	○	○	○	○	○	○
	(unknown)	○	○	×	×	○	×
物質科学	HPhi	×	○	×	○	○	×
エネルギー・資源	NTChem	○	○	×	○	○	×
	SMASH	○	○	×	○	○	×
	paraDMRG	○	○	○	○	○	○
	GELLAN	○	○	×	○	○	×
	MODYLAS	○	○	○	○	○	○
脳科学・AI	WHC	○	○	×	○	○	×
	Realtime cerebellum	○	○	○	○	○	○
	NEURON K+ Stochastic	○	○	○	○	○	○
	CNN (Forward & Back-prop)	○	×	×	○	○	×
	CNN (Forward)	○	×	×	○	×	×
地震・津波	GAMERA	×	○	×	○	○	×
気象気候	NICAM	×	○	×	○	○	×
	SCALE-RM	○	○	○	○	○	○
	CHASER-LETKF	○	○	×	×	×	×
	NICAM-LETKF	×	○	○	○	×	×
宇宙・天文	GreeM	○	○	×	×	○	×
	(unknown)	○	×	×	○	○	×
	EM-PIC	○	○	×	○	×	×
	P3T	○	○	○	○	○	○
	AmaTeRAS	×	○	○	×	○	×

次世代型運用への要求

- 新利用形態への対応
  - 観測データやセンサデータ、外部データベースを直接取り込みながらリアルタイム処理
    - 例）地震観測データによるデータ同化、ゲリラ豪雨予測、ビッグデータによる異常検知
  - 高信頼性・高セキュリティへの要求
- データアーカイブ・流通
  - 重要データの保護、ディザスタリカバリ → 複数拠点でのデータ連携
- 設備・管理
  - 省エネ運用、電力変動対応、冷却設備の負荷変動大への対応
  - 外気導入や湖水・海水の利用を含めた次世代冷却技術
- ユーザ利用・課金モデル
  - Service Level Agreement (SLA)の定義
  - 省エネ実行に対するユーザのインセンティブの明確化と課金モデルへの反映

技術課題と研究開発ロードマップ（抜粋）

- デバイス・アーキテクチャ
  - 電力効率の改善（350Wでノード当たり35TFLOPSの達成は簡単ではない）
  - アクセラレータアーキテクチャの検討
  - 積層メモリの広帯域化と大容量化、ロジックとメモリの3次元積層化
- システムソフトウェア
  - コンテナ化されたアプリケーション／ユニカーネルの効率的な実行
  - データフレームワークの導入
  - ディスアグリゲーション、外部資源（クラウドやIoT）との連携
- ライブラリ・アルゴリズム
  - 新アーキテクチャや非均質なプロセッサへの対応
  - 混合精度演算の考慮
  - 従来型計算機と量子計算のハイブリッド利用

次世代の先端的計算基盤へ向けて

- 次世代計算基盤の創出が果たすべき役割
  - 計算とデータによる科学の発展・進化・新興と社会貢献
  - 新時代のコンピューティング開拓とそれに向けた人材育成
- コデザイン強化＋新応用分野開拓／オープンイノベーションプラットフォームの構築を当初から意識した開発

新応用分野の開拓（一例）

- デジタルツインによるSociety5.0推進
  - 人の行動心理や感情も含めた社会シミュレーション

  → ソフト/AI/データの一体フレームワーク
- 量子・古典ハイブリッド計算環境構築

グランドチャレンジ自体の創出

オープンイノベーションPF

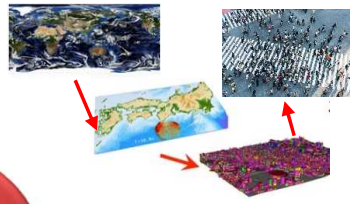
- 開発SW/HWの幅広い展開
- 戦略的な各種連携体制強化が重要
  - ベンダー間、ユーザ間連携
  - ベンダー・ユーザ・開発者間連携
  - 国際的な連携

エコシステムの構築と長期的な人材育成

次世代先端的計算基盤の可能性

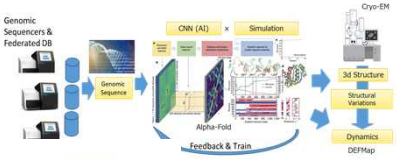
- 様々な科学（物理・工学・化学・生物・計算科学・情報科学など）の共創の場の創生へ

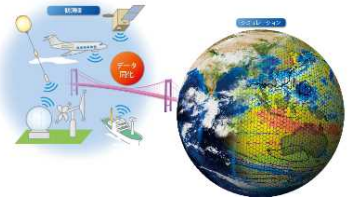
Science×Science
（多様な科学の融合）



Science
×
Computing

Science
×
Data





Next-Generation
Advanced
Computing
Infrastructure

Computing

Data

Computing
×
Computing
（将来の計算の探求）

Computing×Data

Data
×
Data
（データ融合による価値創造）

おわりに

- これまでに頂いているご意見
  - メニーコア型の性能見積りのベースとなるデータが保守的過ぎる
  - より幅広いアプリケーションで性能要求を調査することが必要
  - 将来の本分野の発展に繋がる技術的・アプリケーション的な方向性を示すべき
  - 新計算原理と古典コンピューティングのハイブリッド構成も検討すべき

- NGACIの今後の活動予定
  - 次世代システムの性能予測の精緻化と幅広いアプリでの要求性能調査
  - アプリケーション側の研究・開発コミュニティとの連携

- **アプリケーション・システムの協調設計がこれまで以上に重要**
  - コデザインによるHW/SWアーキテクチャ検討が重要、実装はオープンに検討を