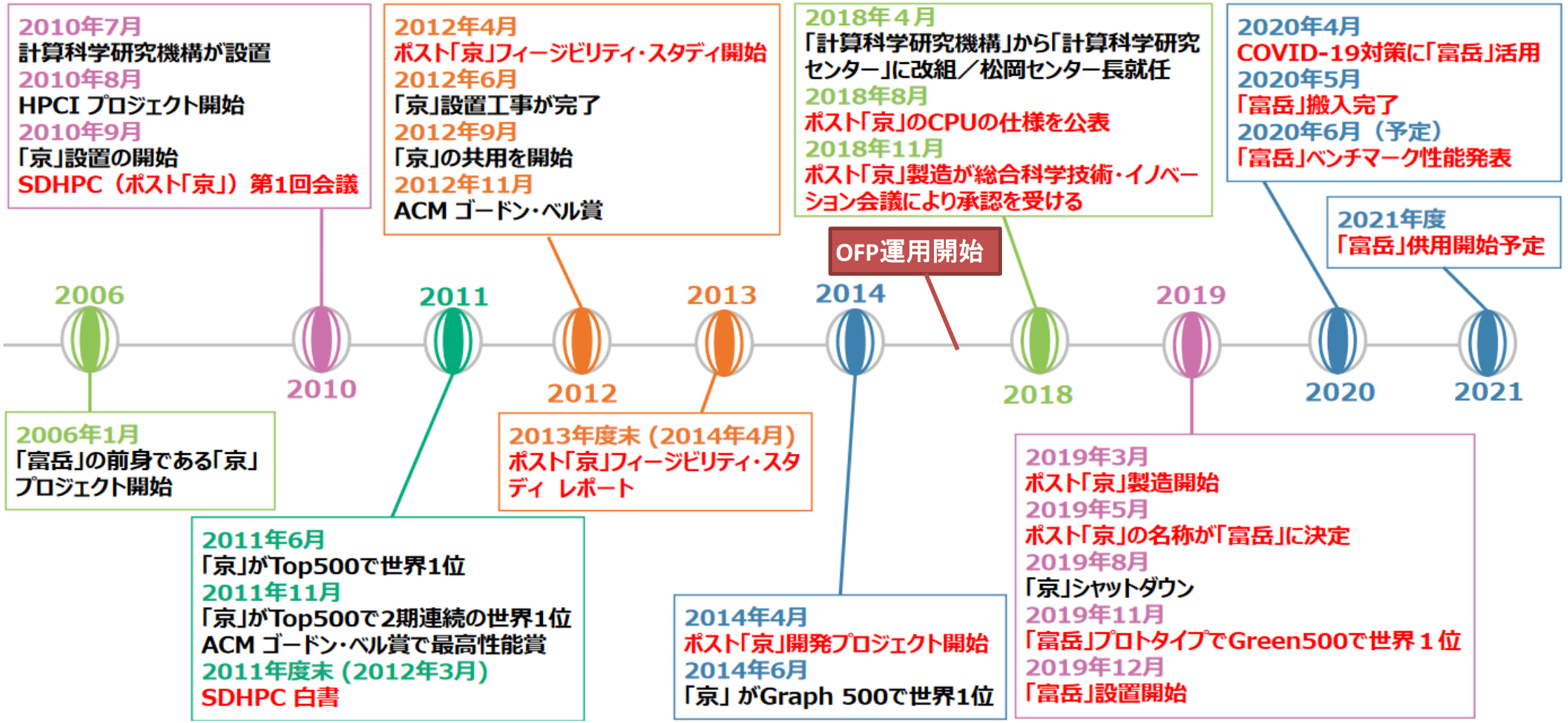
次世代先端的計算基盤の開発に向けた NGACIでの取り組み

慶應義塾大学理工学部情報工学科
理化学研究所計算科学研究センター

近藤 正章

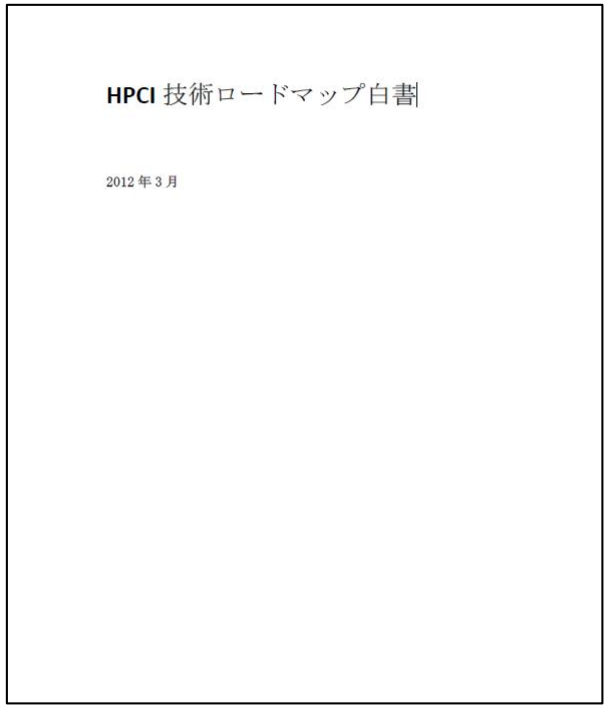
富岳開発への道のり

- SDHPCの活動がポスト京コンピュータ（富岳）開発の源泉に



SDHPC:戦略的高性能計算システム開発

- 2010年より高性能計算システムにおいて今後必要な技術などを議論
- 「HPCI 技術ロードマップ白書」を2012年3月にまとめる
  - この白書をベースに富岳開発プロジェクトの前身であるフィジビリティスタディが始まる


HPCI 技術ロードマップ白書

2012 年 3 月

SDHPCでまとめた白書

表1. プロセッサ・メモリ： 最大(20MW)システム性能の予想値

	総演算性能 PetaFLOPS	総メモリ帯域 PetaByte/s	総メモリ容量 PetaByte
汎用(従来型)	200~400	20~40	20~40
容量・帯域重視	50~100	50~100	50~100
メモリ容量削減	500~1000	250~500	0.1~0.2
演算重視	1000~2000	5~10	5~10

表2. ネットワークのレイテンシと帯域の性能の予想値

	Injection	P-to-P	Bisection	Min 遅延	Max 遅延
High-radix (Dragonfly)	32 GB/s	32 GB/s	2.0 PB/s	200 ns	1000 ns
Low-radix (4D Torus)	128 GB/s	16 GB/s	0.13 PB/s	100 ns	5000 ns

アプリに適するシステム構成の検討結果

次世代計算基盤の開発に向けたコミュニティ活動

- 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
    - 9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）
    - 更新版である1.1.0版の公開準備中


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

次世代計算基盤に関するFSプロジェクトも公募開始

次世代計算基盤に係る調査研究

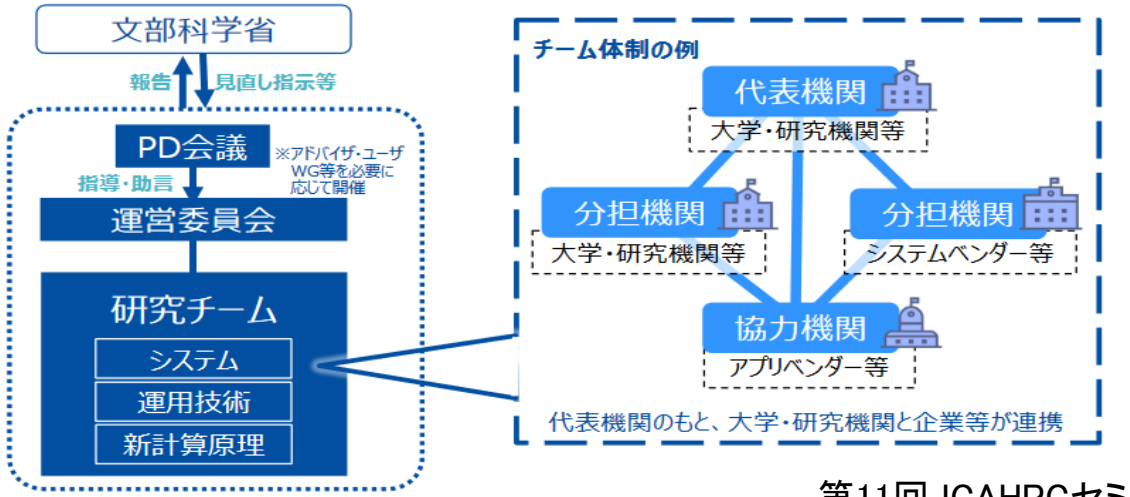
令和4年度予算額　4.3億円（新規）

背景

- 近年、大量かつ多様なデータの収集や活用が進展し、**データ駆動型科学が重要視される中で、シミュレーションやAI 等が連携した研究の重要性がより一層高まっている。**さらに、新型コロナウイルス感染症の拡大を契機として、研究のリモート化やスマート化、研究設備・機器への遠隔からの接続、データ駆動型研究の拡大など、**世界的にも研究活動のデジタルトランスフォーメーション（研究 DX）の必要性が高まっている。**
- 社会のデジタル化を進め、サイバー空間とフィジカル空間の融合によって新たな価値を創出していくSociety 5.0 を実現するため、スーパーコンピュータのみならず、データセンターからエッジコンピューティング、それらを繋ぐネットワーク等様々な形態の社会情報基盤がますます重要となっている。また、これらの**基幹技術を自国で保有することは経済安全保障の観点からも重要である。**
- これらの情勢を踏まえると、**ポスト「富岳」時代の次世代計算基盤を、国として戦略的に整備することは必要不可欠である。**

事業内容・目的

- ポスト「富岳」時代の次世代計算基盤の開発にあたり、我が国として独自に開発・維持するべき技術を特定しつつ、要素技術の研究開発等を実施し、具体的な性能・機能等について検討を行う。
- システム（アーキテクチャ、システムソフトウェア・ライブラリ、アプリケーション）、新計算原理、運用技術を対象に調査研究を実施。**サイエンス・産業・社会のニーズを明確化し、それを実現可能なシステムの選択肢を提案する。**



＜研究期間＞

令和4年度～令和5年度

※令和6年度以降の取組は、調査研究の進捗を踏まえ検討

＜スケジュール＞

公募開始	令和４年５月18日（水）
公募説明会	令和４年５月24日（火）
申請締切	令和４年６月21日（火）
審査等	令和４年６～７月頃（予定）
選定結果通知	令和４年７月頃～（予定）
委託契約等	令和４年７月頃～（予定）
事業開始	令和４年７～８月頃～（予定）

出典：次世代計算基盤に係る調査研究事業公募説明会資料

次世代情報基盤に向けたスケジュール予測

次世代情報基盤（富岳NEXT）の開発スケジュール予測

2020 R2	2021 R3	2022 R4	2023 R5	2024 R6	2025 R7	2026 R8	2027 R9	2028 R10	2029 R11	2030 H12	2031 H12
NGACI WP1.0	WP1.1										
FS1.5年版	FS公募	調査研究		開発PJ (基本設計)		開発PJ (詳細設計)			開発PJ (製造・設置・調整)		
FSを2.5年にした場合	FS公募	調査研究			開発PJ (基本設計)		開発PJ (詳細設計)			開発PJ (製造・設置・調整)	
TSMC N5		TSMC N4	TSMC N3			TSMC N2			TSMC N1.4?		
Intel 10nm		Intel7 (10nm+)		Intel4	Intel3	Intel20A		Intel18A		Intel15A?	

（参考）富岳開発スケジュール実績

▽ Top500#1

2010 H22	2011 H23	2012 H24	2013 H25	2014 H26	2015 H27	2016 H28	2017 H29	2018 H30	2019 H31/R1	2020 R2
ロードマップ策定 2010.8~ 2011.7~2012.3		調査研究 2012.7~2014.3		基本設計		詳細設計			製造・設置・調整	

- 理研ポスト「京」基本構成および性能目標等提案
- 構成案変更
- 夏 製造技術10nmから7nmに変更。当初計画から正式運用開始予定時期
- 夏 試作機完成
- 12月 ハードウェア搬入開始

https://www.ssken.gr.jp/MAINSITE/event/2020/20210121-sci/lecture-01/20210121_sci_ishikawa.pdf

White Paperの章構成（WP1.0.0より）

White Paperの執筆協力者（WP1.0.0より）

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）

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(WP1.0.0より)

- 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（WP1.0.0より）

- 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

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
  – ：

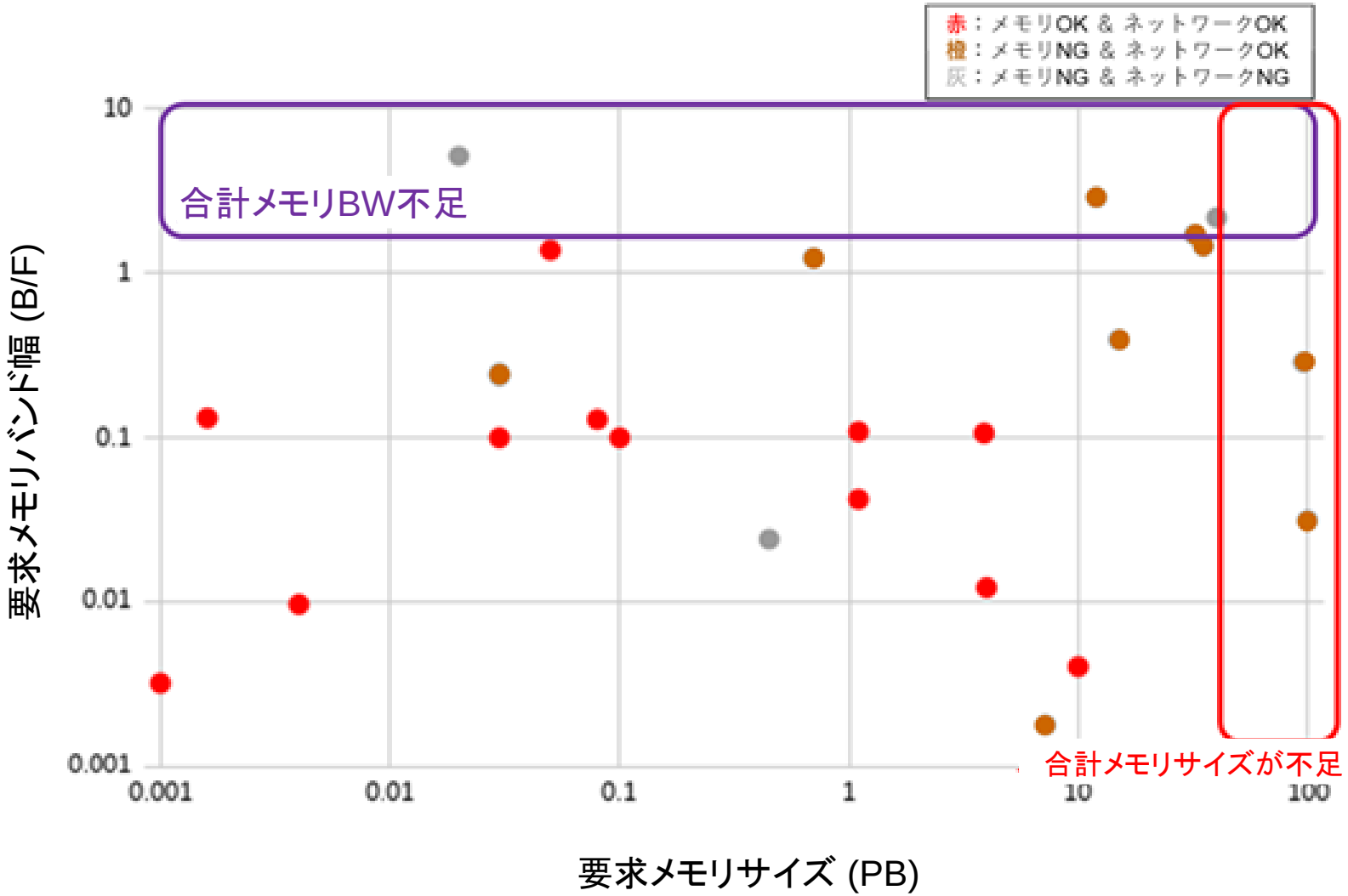
アプリケーションの要求性能分析

- 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

アプリケーションの要求性能との比較（WP1.0.0より）

- メニーコア型システム構成（システム電力50MW, CPU80%）の場合



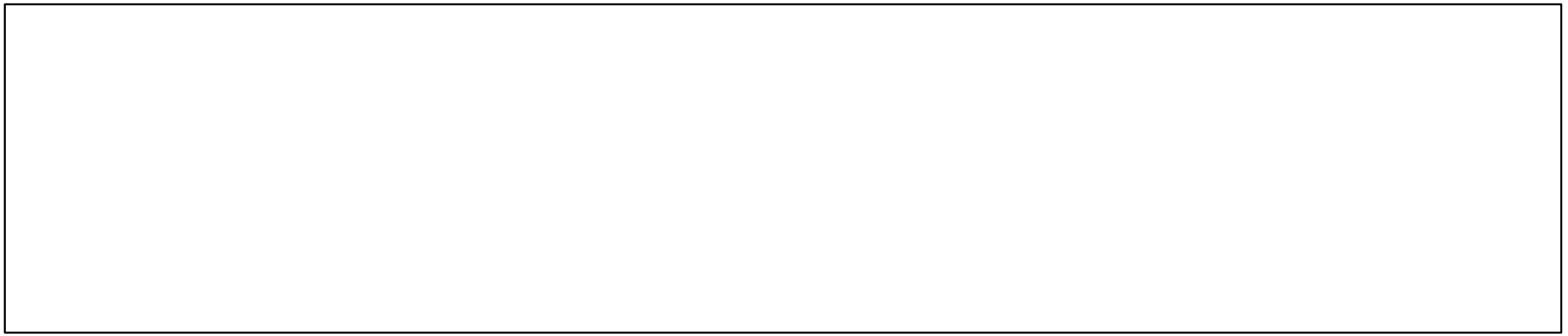
2028年のシステムの予測性能（更新版）

- 現在NGACIで議論中のシステム性能予測データの一部を紹介
  - 三輪先生@電通大、安島さん@富士通、塩見先生@阪大に感謝いたします

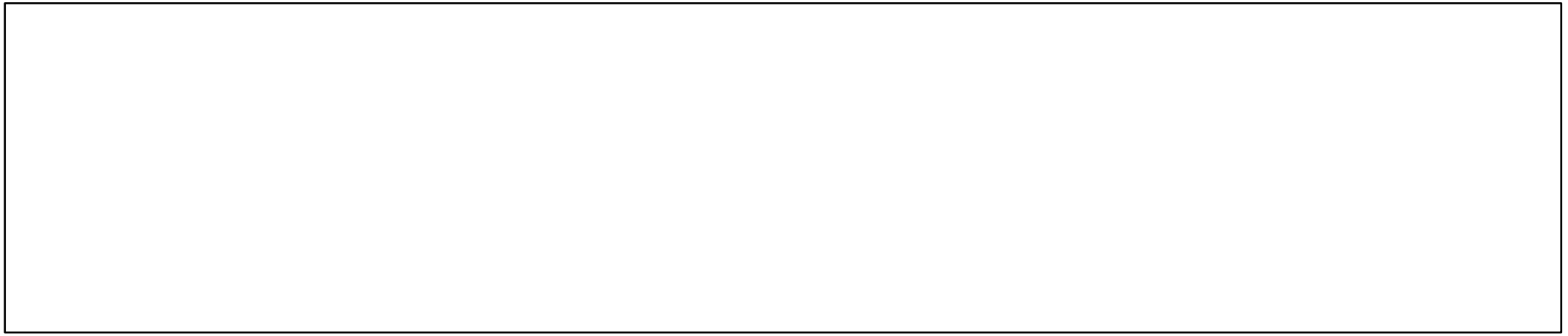
- 仮定するプロセッサのタイプ
  - Latency Sensitive（LS）
    - データベースなどランダムアクセス処理にも対応可能なアーキテクチャ
    - 大容量メモリ、大容量LLC、深いキャッシュ階層
    - 例）Xeon, EPYC, Amazon Graviton, etc.
  - Bandwidth Centric（BC）
    - グラフィックス、HPCなどストリーミング処理
    - 高帯域メモリ、帯域×遅延に必要十分なローカルメモリ／キャッシュ
    - 例）Tesla, Radeon Instinct, Intel Xe, A64FX, SX-Aurora Tsubasa, etc.
  - Compute Centric（CC）
    - 学習、推論などAI処理
    - 高密度な演算器、必要十分な入出力帯域
    - 例） Google TPU, Cerebras WSE, Tesla D1, Esperanto ET, Graphcore Colossus, SambaNova, etc.

2028年のシステムの予測性能(更新版)

- ソケットあたりの性能比較



- 50MW、80%におけるシステム性能の見積り



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＆データによる科学の発展・進化と社会貢献に向けたプラットフォーム化
  - 新時代のコンピューティングの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PF構築＋ポストムーア時代へのアプリ進化を当初から意識したアーキテクチャ選定・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

おわりに

- これから富岳Nextに向けた動きも加速すると予想
  - 調査研究（特に運用技術チーム）では第二階層との連携にも言及がなされている
- フラグシップだけでなく第二階層のシステムも含めた次世代計算基盤の議論が今後重要に
- **アプリケーション・システムのコデザインもこれまで以上に重要**
- **今後も皆様と密に協力・連携させて頂ければ幸いです**