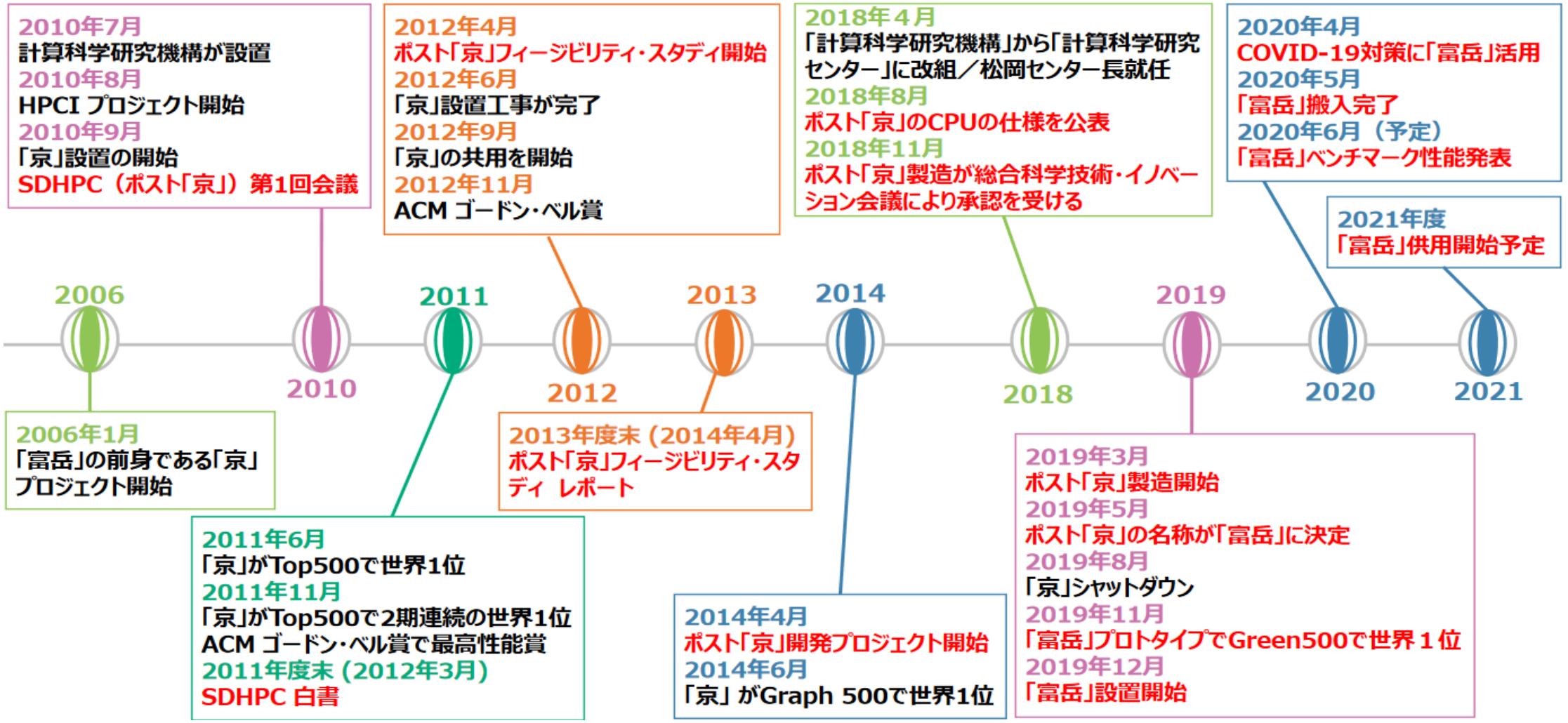
次世代情報基盤に関するコミュニティ活動と調査研究事業について

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

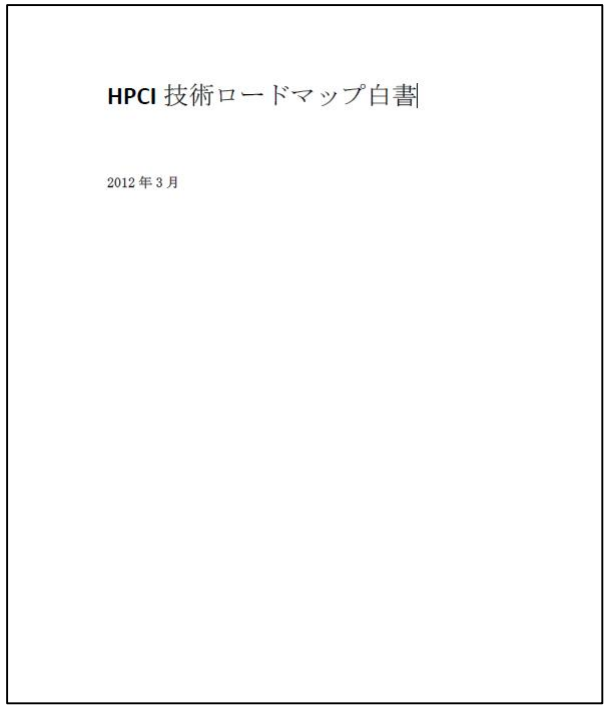
富岳開発への道のり

- 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

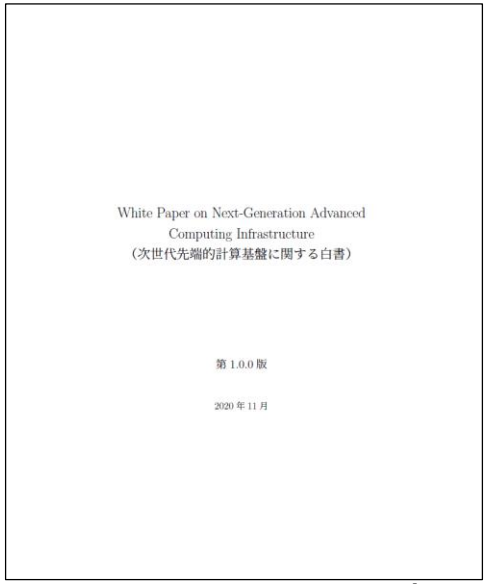
次世代計算基盤の開発に向けたコミュニティ活動

- 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

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%

参考：IRDS SA 2020 Editionでの予測

Table SA-1 Difficult Challenges

	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034
# cores per socket	38	42	46	50	54	58	62	66	70	70	70	70	70	70	70	70
Processor base frequency (for multiple cores together)	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90	4.00	4.10	4.2	4.3	4.4	4.5
Core vector length	512	512	1024	1024	1024	1024	2048	2048	2048	2048	2048	2048	2048	2048	2048	2048
L1 data cache size (in KB)	36	38	38	40	40	42	42	44	44	44	44	44	44	44	44	44
L1 instruction cache size (in KB)	48	64	64	96	96	128	128	160	160	160	160	160	160	160	160	160
L2 cache size (in MB)	1	1.5	1.5	1.5	2	2	2	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5
LLC cache size (in MB)	67	73	81	89	97	107	118	130	143	157	173	190	200	200	200	200
# of DDR channels	6	8	8	10	10	12	12	12	12	12	16	16	16	16	16	16
HBM ports	4	4	6	6	6	6	6	6	6	6	6	6	6	6	6	6
HBM bandwidth (TB/s)	2.4	2.4	6	6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6
Fabric lanes	64	72	80	88	96	104	112	120	128	136	144	152	152	152	152	152
Per lane (GT/s)	56	56	56	56	56	56	56	56	56	56	56	100	100	100	100	100
Socket TDP (Watts)	226	237	249	262	275	288	303	318	334	351	368	387	387	387	425	425

L1 = level 1 cache; LLC = last-level cache; Fabric = PCIe or new accelerator fabric (CXL/Gen-Z/openCAPI/CCIX); TDP = total power dissipation.

Source:
INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS 2020 EDITION SYSTEMS AND ARCHITECTURES

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）

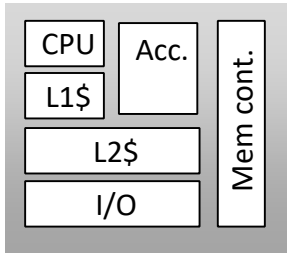
	30MW			40MW			50MW		
	60%	70%	80%	60%	70%	80%	60%	70%	80%
GPU 数	50661	59104	67548	67548	78806	90064	84435	98508	112580
総コア数	5.3×10^8	6.2×10^8	7.1×10^8	7.1×10^8	8.3×10^8	9.4×10^8	8.8×10^8	1.0×10^9	1.2×10^9
PFLOPS	1279	1492	1706	1706	1940	2474	2132	2487	2843
HBM 総 BW (PB/s)	91	107	122	122	143	163	153	178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

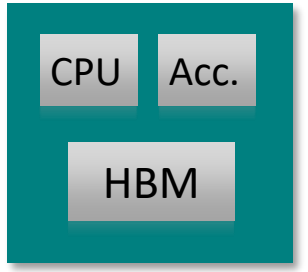
積極的な予測の場合（CPUとの性能比のトレンドから外挿）

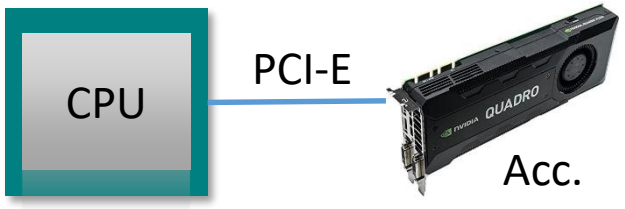
	30MW			40MW			50MW		
	60%	70%	80%	60%	70%	80%	60%	70%	80%
GPU 数	50661	59104	67548	67548	78806	90064	84435	98508	112580
総コア数	3.4×10^9	3.9×10^9	4.5×10^9	4.5×10^9	5.2×10^9	6.0×10^9	5.6×10^9	6.5×10^9	7.5×10^9
PFLOPS	8083	9431	10778	10778	12574	14371	13472	15718	17963
HBM 総 BW (PB/s)	334	390	445	445	520	594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

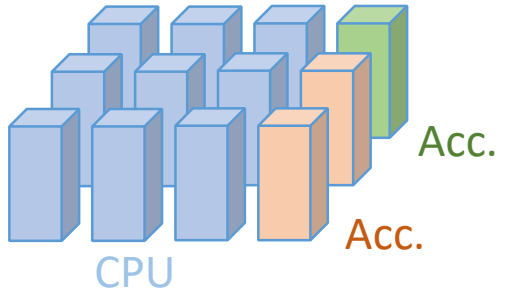
アクセラレータのシステム搭載方式

- チップ内拡張型（SoC）
  - CPUとアクセラレータが同一CPUダイ上で結合
  - オンチップキャッシュなどのメモリ階層の一部を共有
- チップ内拡張型（マルチチップパッケージ）
  - チップレットをインターポーザで結合
  - 主記憶を共有
- ノード内拡張型
  - アクセラレータがPCI-Express、CXL、CAPI等のインターコネクトでCPUと結合
  - メモリ間のデータ転送やアクセス制御の検討が必要
- 問題特化型ノードやラックによる疎結合型
  - 独立したアクセラレータ専用ノードやラックをネットワークで結合









→ 一方でアクセラレータは手段であり目的でないことに留意
→ 詳細なワークロードの分析により決定することが重要

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

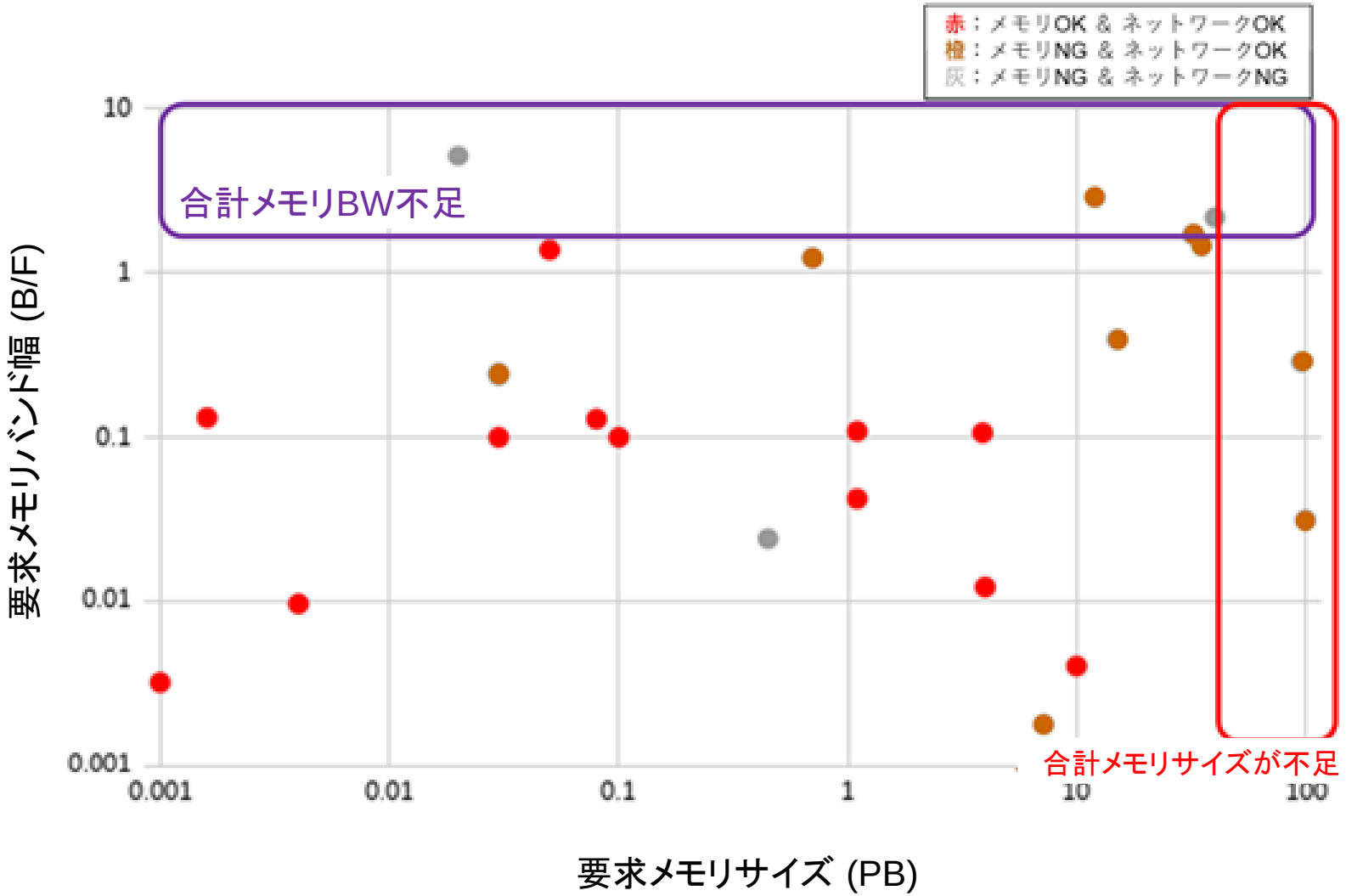
アプリケーションの要求性能

- サイエンスロードマップ(2017年版)をベースにNGACIで作成

App. Area	Name of Application	# of Required Node	Comp. Perf. (TFlops/node)	Memory BW (TB/s/node)	Memory (GB/node)	Interconnect BW (Tb/s/node)	File I/O BW (GB/s/node)	Storage Size (PB)
Elementally particle, Atomic nucleus	(unknown)	40,000	7.75	1	2	1.25	0.045	60
	(unknown)	4,100	7.56	0.073	0.98	0.088	0.039	5
	(unknown)	1,000	5.6	0	1	0	0.1	0.001
	rmcsm	6,000	5	0.5	5	0	0.017	0
	(unknown)	4,100	4.15	21.22	4.88	53.75	0.006	120
Material science	HPhi	1,000,000	0.03	0.001	100	1.025	0.009	340
Energy, Resources	NTChem	18,000	11.11	16.111	1,944.44	0.001	0.002	0.1
	SMASH	100,000	10	17	320	0.014	0.003	0.32
	paraDMRG	100,000	31	1.3	11	0.05	0.01	0
	GELLAN	100,000	2.7	0.033	40	0.018	0	0.02
	MODYLAS	100,000	5	0.5	1	0.003	0.001	500
Brain science, AI	WHC	50,000	30	3.2	78	0	0	1
	Realtime cerebellum	50,000	11.2	1.22	22	0	0	5
	NEURON K+ Stochastic	100,000	7.6	1	0.02	0	0.13	160
	CNN (Forward & Back-prop)	16,000	100	23.75	1.88	0.25	0.006	175
	CNN (Forward)	6,900	100	24.638	4.35	0.425	0.493	290
Earthquake, Tsunami	GAMERA	200,000	2.8	1.1	75	0.025	0.005	1
Weather, Climate	NICAM	400,000	1.23	3.5	30	0	0	230
	SCALE-RM	40,000	1.15	1.575	1.25	0	0	33
	CHASER-LETKF	32,000	0.2	0.438	1250	750	2.5	50
	NICAM-LETKF	160,000	1.38	1.688	4.38	0	0.5	10
Space, Astronomy	GreeM	100,000	6.9	0.028	100	25	0.3	300
	(unknown)	100,000	200	0.36	72	0.125	0.06	2,600.00
	EM-PIC	100,000	1.6	0.46	960	0.125	10	2,000.00
	P3T	10,000	3.1	0.01	0.1	0.001	0.001	400
	AmaTeRAS	1,000,000	1	0.024	0.45	41.25	0	20

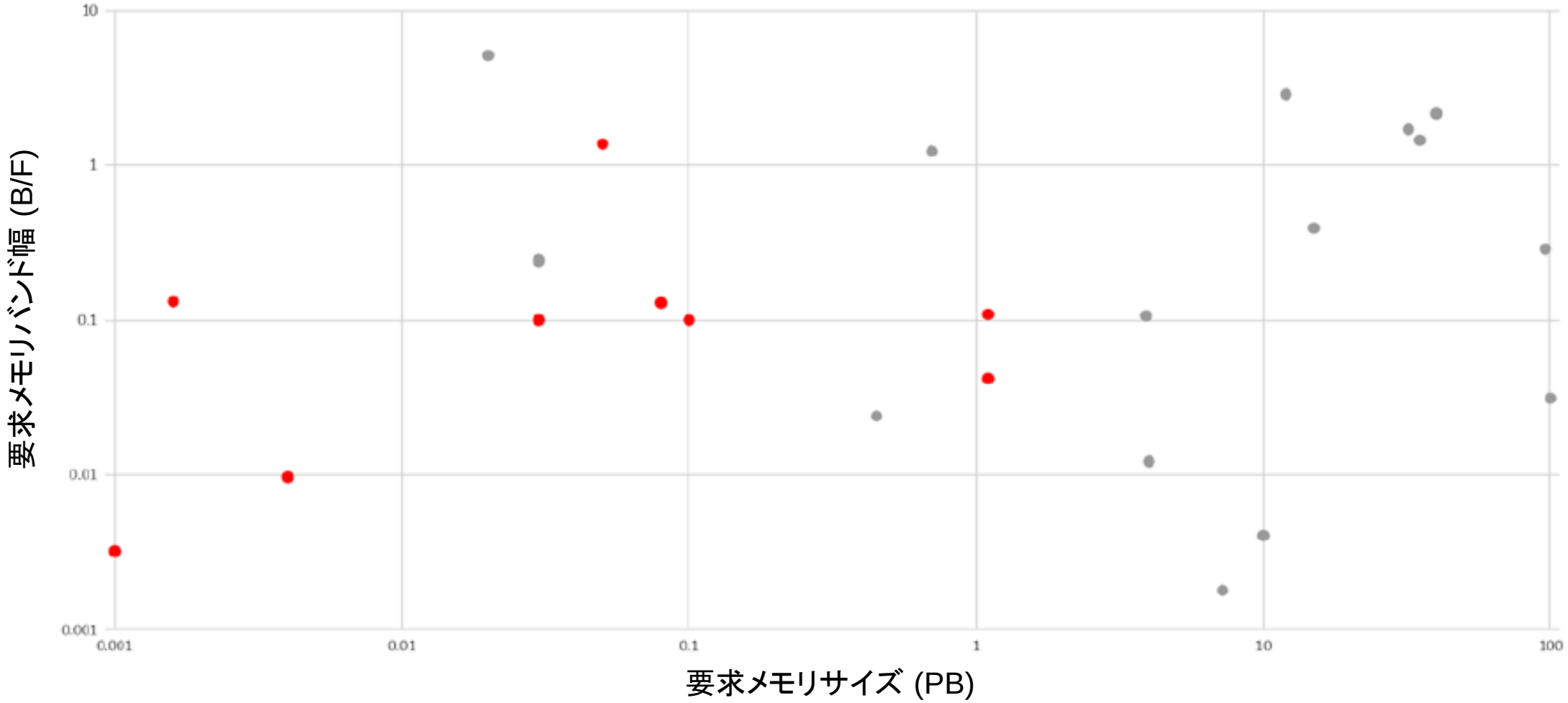
アプリケーションの要求性能との比較（WP1.0.0より）

- メニーコア型システム構成（システム電力50MW, CPU80%）の場合



アプリケーションの要求性能との比較

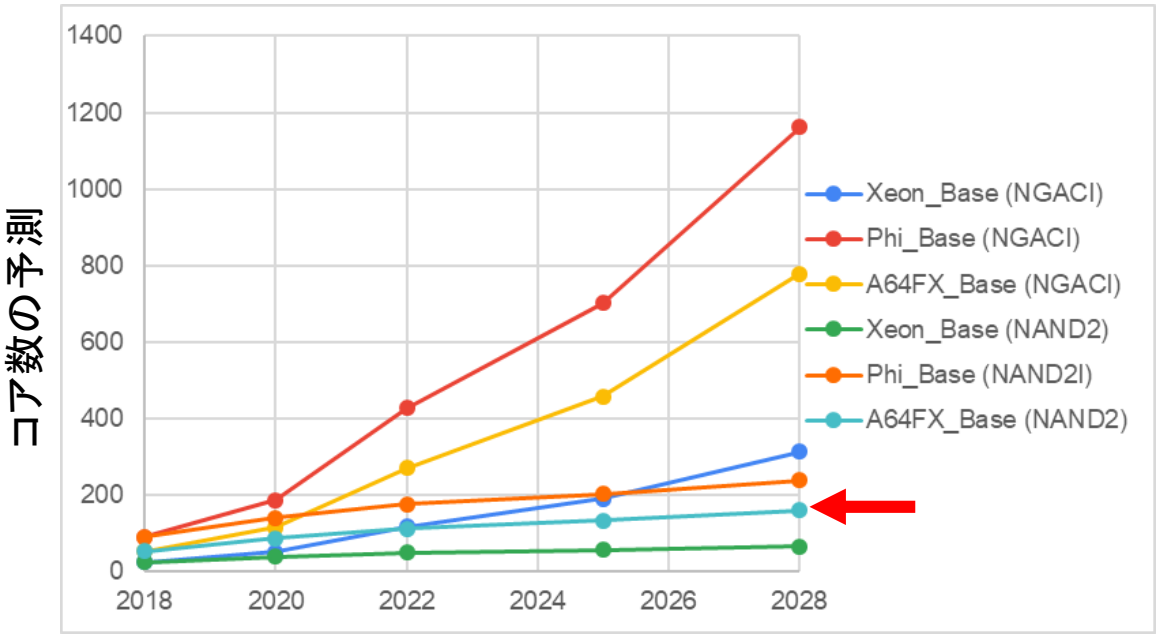
- GPU混載型システム構成（システム電力50MW, GPU80%）の場合
  - 赤: 要求性能を満たす, 灰色: 要求性能を満たしていない



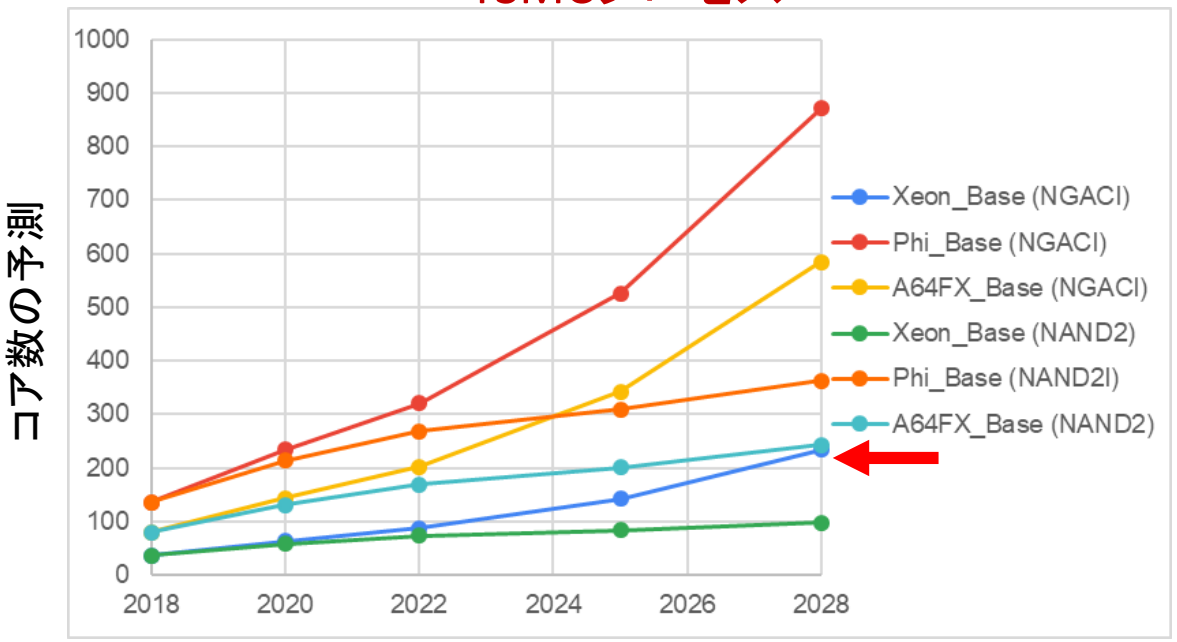
より高性能なシステム構成へ向けて

- HPC向けのプロセッサを仮定したメニーコアの（個人的な）見積り
  - Intelプロセスでの仮定：チップサイズ600mm^2、コア部割合=33%
  - TSMCプロセスでの仮定：チップサイズ450mm^2、コア部割合=33%

Intelプロセス

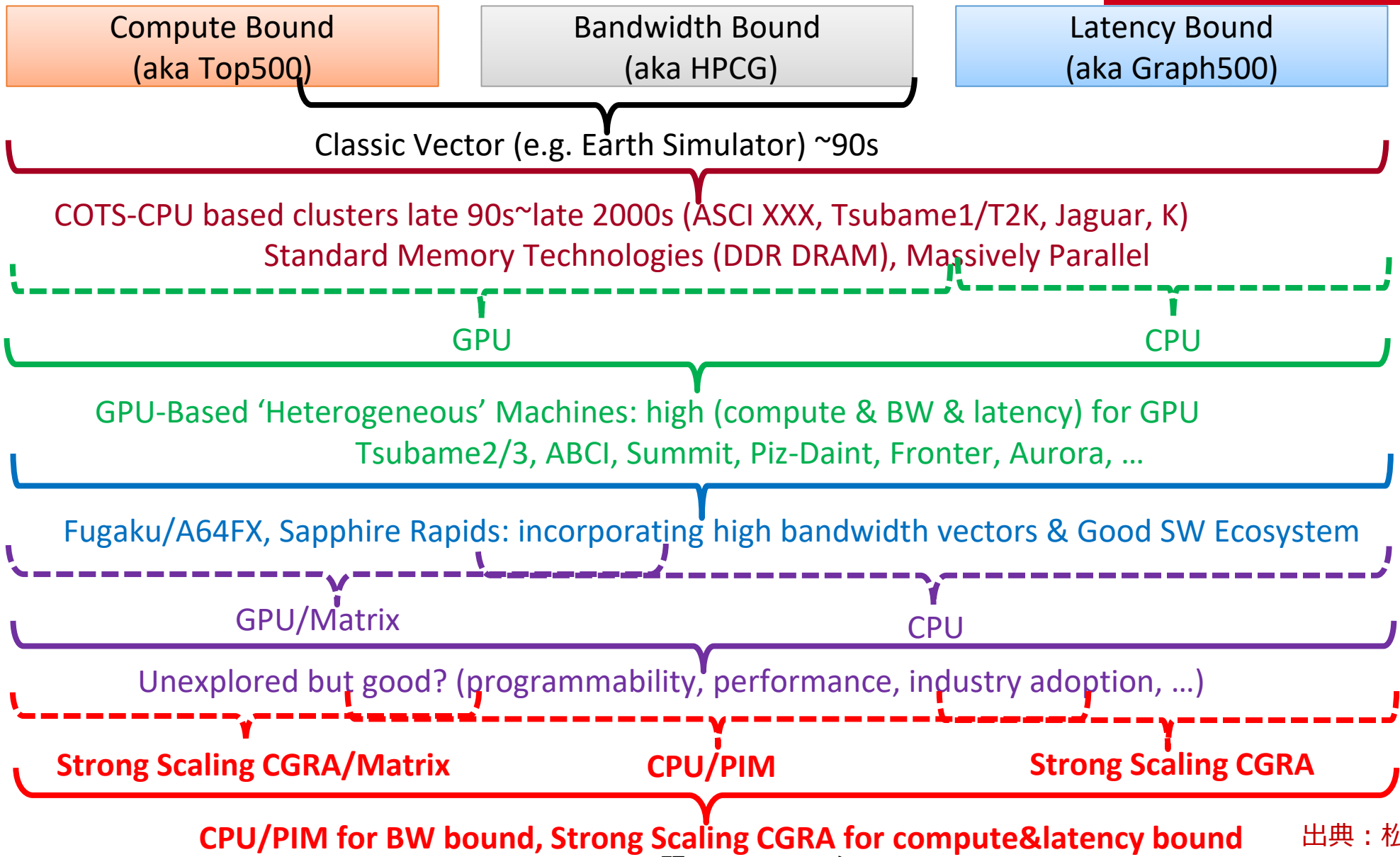


TSMCプロセス



  - A64FXベースで見積もると160～240コア程度は搭載可能
  - メモリバンド幅とバランスは未考慮 → 考慮した上でシステムを検討する必要あり

ワークロードの分類とそれぞれをカバーするアーキテクチャ



出典：松岡R-CCSセンター長

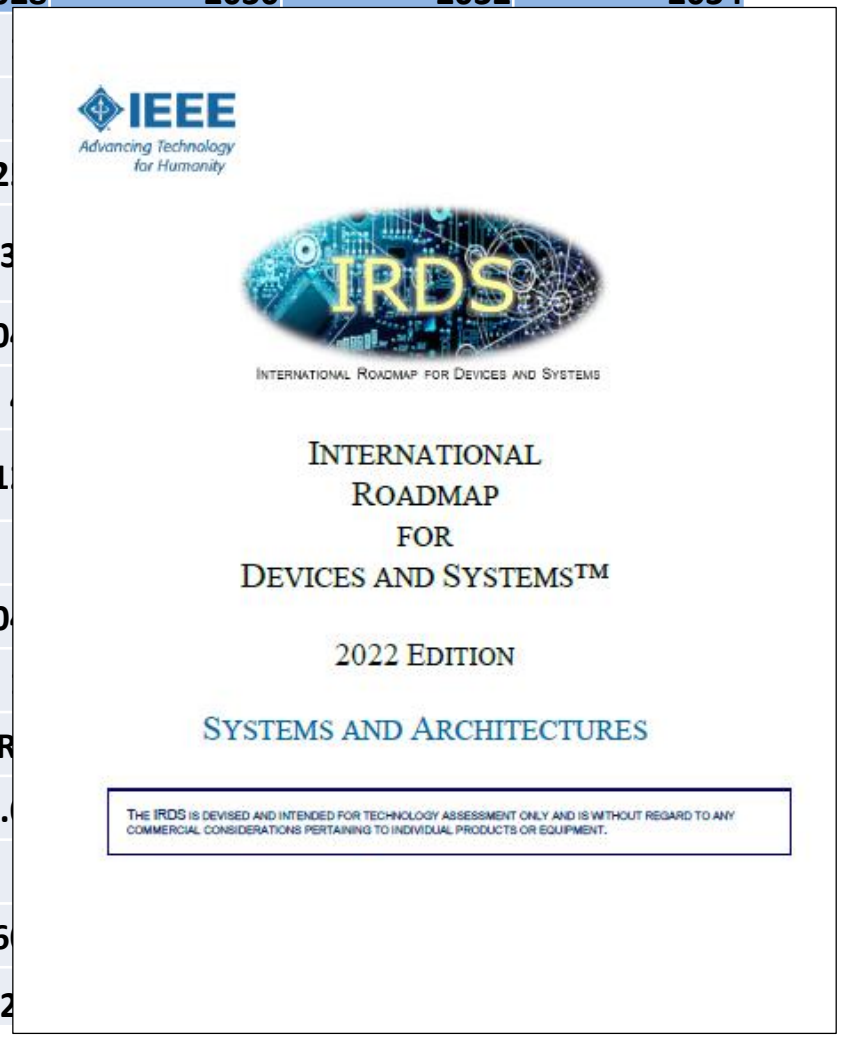
NGACIにおけるプロセッシングエレメントの分類

- 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.

Latency Sensitiveプロセッシングエレメントのトレンド予測

- IEEE IRDS Roadmap System and Architecture (2022年版)に提案し掲載

	2019	2022	2024	2026	2028	2030	2032	2034
# chiplet per socket	8	8	8	8				
# core per chiplet	8	8	12	16				
# core per socket (max)	64	64	96	128	2			
Processor base frequency (GHz)	2.2	2.5	2.8	3	3			
SIMD width (bit)	1024	1024	1024	1024	20			
L1 data cache size (in KB)	36	40	40	42				
L1 instruction cache size (in KB)	48	96	96	128	1			
L2 cache size (in MB)	1	1.5	2	2				
LLC cache size (in MB)	64-128	64-800	128-1024	256-1536	256-20			
# of DDR channels	8 (DDR4)	12 (DDR4)	12 (DDR5)	12 (DDR5)	(DDR			
DDR bandwidth (TB/s)	0.2	0.31	0.61	0.76	1.			
DDR size per socket (in TB)	1	3	4.5	6				
Socket max TDP (Watts)	280	300	400	450	6			
Socket performance (TFLOPS)	4.5	5.1	8.6	12.3	52			



Bandwidth Centricプロセッシングエレメントのトレンド予測

- IEEE IRDS Roadmap System and Architecture (2022年版)に提案し掲載

	2019	2022	2024	2026	2028	2030	2032	2034
# chiplet per socket	1	1	4	16	16	16	16	20
# core per chiplet	108	132	36	36	42	48	54	54
# core per socket (max)	108	132	144	576	672	768	864	1080
Processor base frequency (GHz)	1.2	1.4	1.4	1.4	1.6	1.6	1.6	1.6
Vector length (Ops)	32	64	64	64	128	128	128	128
# of HBM ports	6	6	8	8	8	10	10	10
HBM bandwidth (TB/s)	1.6	3	4	4	6.6	8	8	10
HBM size per socket (in GB)	40	80	256	512	800	1024	1024	1024
Socket max TDP (Watts)	400	400	400	440	500	500	500	500
Socket performance (TFLOPS)	8.3	23.7	25.8	103.2	275.3	314.6	353.9	442.4

2028年のシステムの予測性能

- ソケットあたりの性能比較

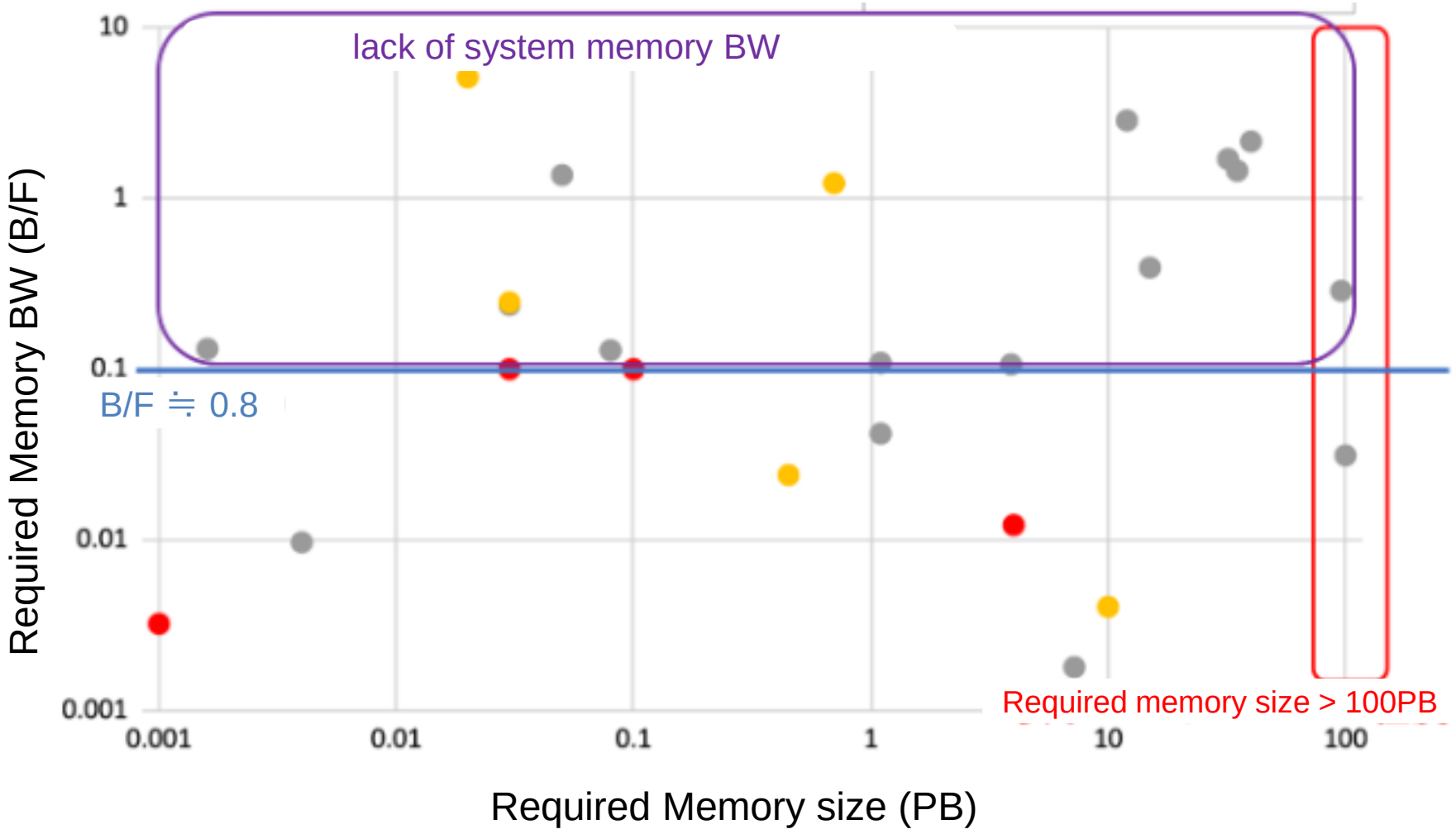
	LS model	BC model
FP64 Peak Performance	13.1 TFlops	137.6 TFlops
Memory Bandwidth	1.02 TB/s	6.6 TB/s
Memory Capacity	8 TB	800 GB
B/F	0.08	0.05
Power Consumption	600 W	500 W

- システム全体の性能見積（40MWの電力を仮定した場合）

	LS model	BC model
Num. of Socket	48,485	58,182
PFLOPS	615	8,007
Total Memory BW (PB/s)	49.5	384
Total Memory Capacity(PB)	379	44.4

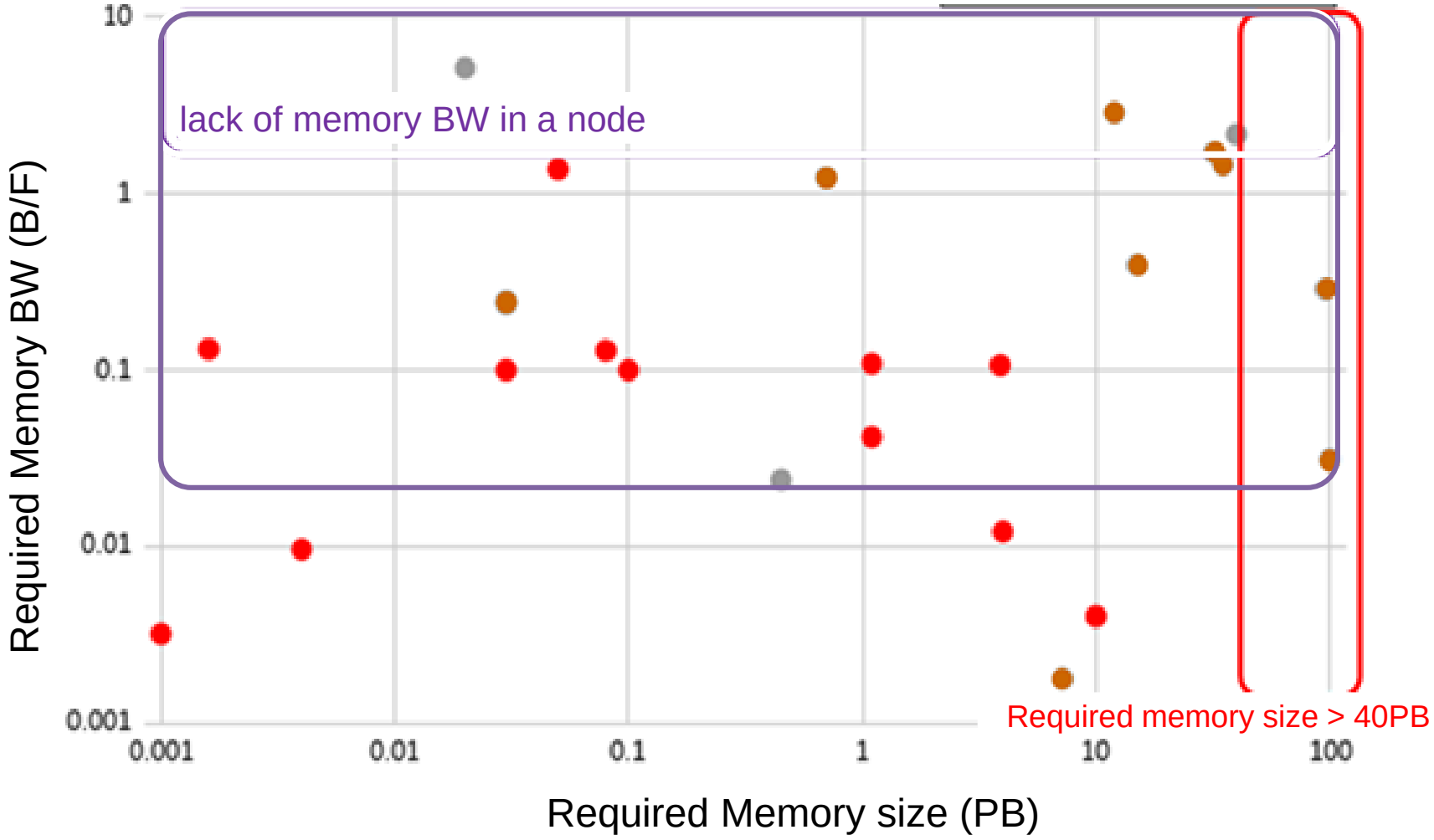
アプリケーションの要求性能との比較

- Latency Sensitive（LS）型プロセッサの場合



アプリケーションの要求性能との比較

- Bandwidth Centric（BC）型プロセッサの場合



世代計算基盤に係る調査研究（FS）の概要

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

- ◆ **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、**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。**
- ◆ 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おり、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。

次世代計算基盤検討部会　中間まとめ（令和３年８月）

◆ **次世代計算基盤検討の留意事項**

技術動向や周辺状況が急速に進化・変化
ムーアの法則の終焉等、関連技術が転換期にある、性能の向上に伴い要求される電力量も増大
⇒　半導体やネットワーク等**国内外の周辺技術動向や利用側のニーズの調査**、**要素技術の研究開発**等必要な調査研究を行い、**多角的な検討**が必要。

◆ **次世代計算基盤の在り方**

次期「フラッグシップシステム」及び国内の主要な計算基盤、データ基盤、ネットワークが一体的に運用され、総体として持続的に機能する基盤
⇒　**調査研究（FS）を通じ**、技術的課題や制約要因を抽出しつつ、**実現可能なシステム等の選択肢を提案**

<ポスト「富岳」時代に目指すべき姿>

次世代計算基盤に係る調査研究

◆ 具体的には以下の取組を実施。

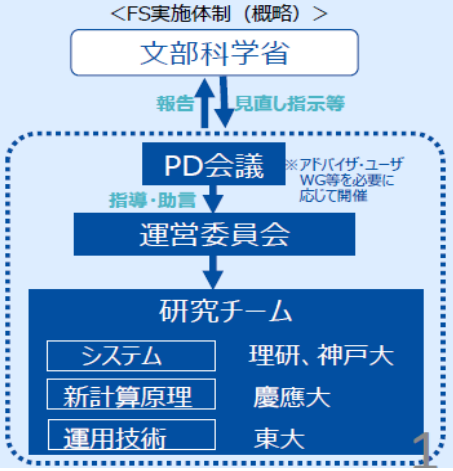
- **要素技術の研究開発**（併せて、我が国として独自に開発・維持するべき技術を特定）
- **評価指標**の検討（例：演算性能、電力性能比、I/O性能、コスト、運用可能性、生産性（アプリ開発のしやすさ）、商用展開・技術展開、カーボンニュートラルへの対応　等）
- **技術的課題や制約要因**の抽出　　等

◆ 実施期間：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

令和５年度の取組：システム候補の性能評価、アプリケーションのコデザイン、新たな計算原理を適用すべき領域・分野の検討、多様な計算基盤の一体的運用、これらにおいて必要な要素技術の研究開発　等

令和４年度の取組：技術や利用分野の動向調査、評価項目・手法の検討　等

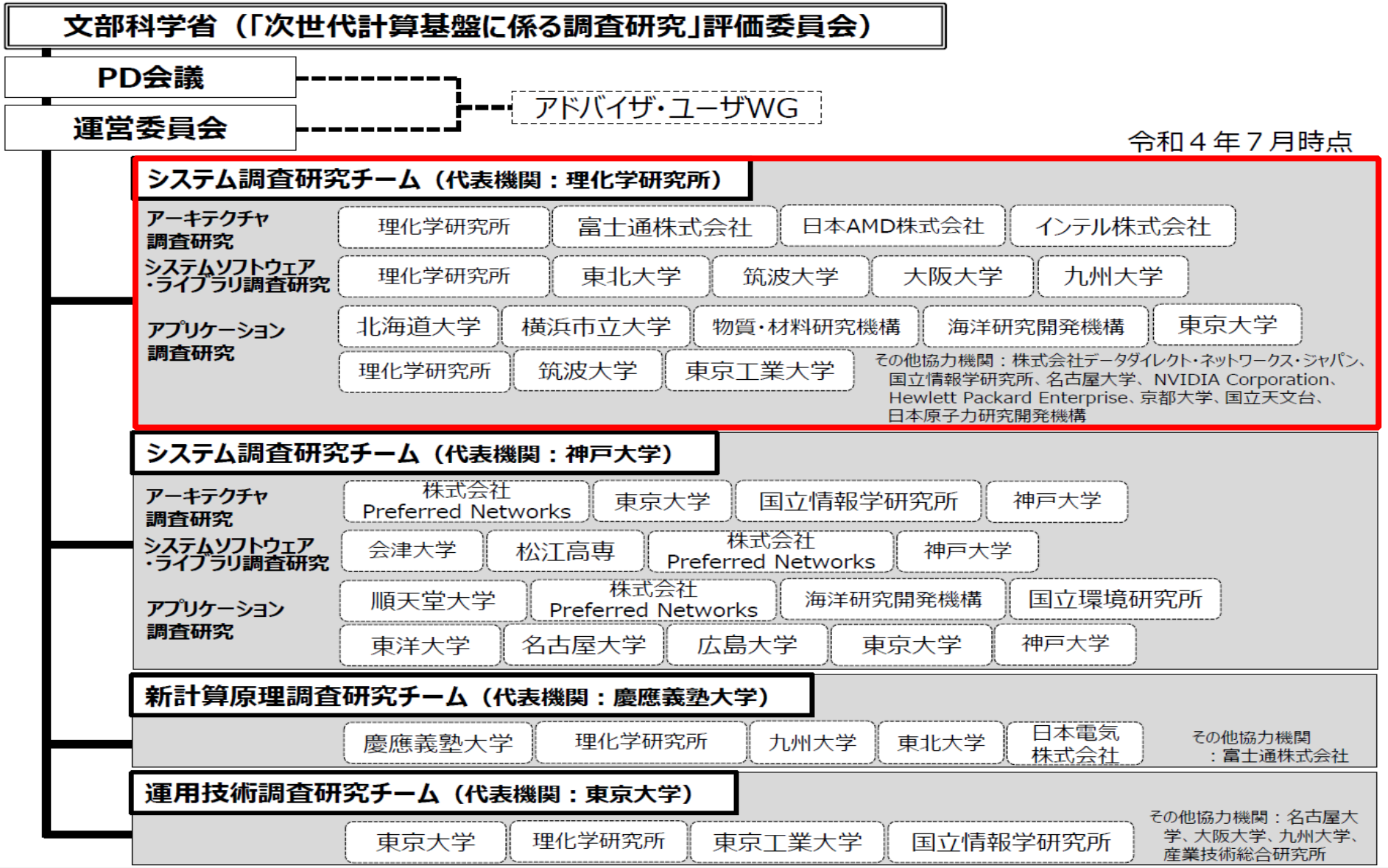
<FS実施体制（概略）>



出典：文部科学省

世代計算基盤に係る調査研究（FS）の全体推進体制

「次世代計算基盤に係る調査研究」実施体制



Source: MEXT

調査研究の全体概要（理化学研究所：システム調査研究）

取組概要

次世代計算基盤には、SDGs・Society5.0の実現に向けた課題解決のためのプラットフォームとしての役割が求められる。そこで、今後の科学に「研究DX」をもたらす**高度なデジタルツイン実現の基盤**として、広範な計算手法・シミュレーション技法や大規模データを駆使しつつ、それらが密に連携しながら全体のワークフロー実行が可能な汎用性の高い計算基盤の実現を目指し、あるべきアーキテクチャやシステムソフトウェア・ライブラリ技術について、アプリケーションとのコデザインを通じた調査研究を行う。

特に、システム設計の基本理念として演算精度も考慮しながら必要な計算性能を確保し、**電力制約の下でデータ移動を高度化・効率化する"FLOPS to Byte"指向のシステム構築**を、アーキテクチャ開発からアルゴリズム設計、アプリケーション技術に至るまで実践する。

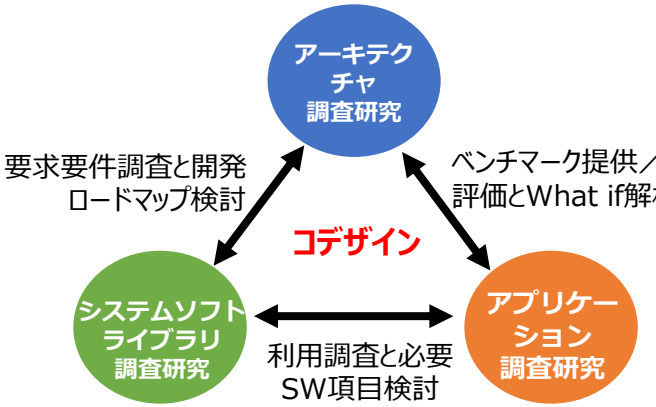
ALL Japan体制のもと、実効的な性能を向上させる次世代計算基盤のシステム構成や要素技術の調査検討、要素技術の開発を、アーキテクチャ・システムソフトウェアとアプリケーションとのコデザインを通じて実施する。

調査内容

アーキテクチャ調査研究
- 半導体技術・パッケージング技術動向を踏まえつつ、システム全体やその構成要素について考え得る技術的可能性や総合性能を調査
  （新規要素技術として三次元積層メモリ技術、強スケーリング・計算インテンシブ向けアクセラレータ、チップ間直接光通信技術などを意識）
- アプリケーション調査研究G提供のベンチマークセットの性能解析に基づき将来システムの性能を予測、また次世代アプリ開発へとフィードバック

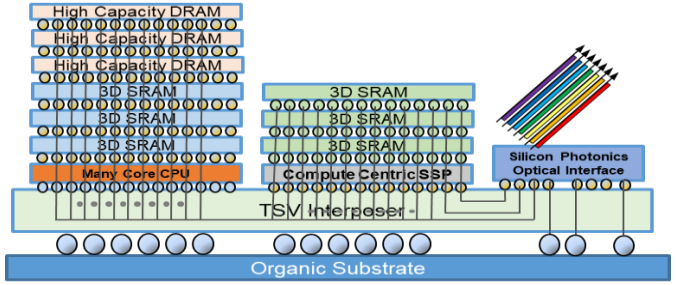
システムソフトウェア・ライブラリ調査研究
- 従来のソフトウェアに加え、データ利活用促進、機械学習技術と第一原理シミュレーションや大規模リアルタイムデータ処理との高度な融合、高セキュリティの担保などを主要検討項目と据え、国内で開発すべきソフトを優先度も含めて明らかにしつつ今後の開発ロードマップを策定

アプリケーション調査研究
- 複数アーキテクチャを統一的に評価するための広範なベンチマークセットを構築し、それを利用したアーキテクチャ評価結果を踏まえてアルゴリズムやパラメータの改善を検討しつつベンチマークセットを更新しながら、性能モデルも利用した探索的な"What if"の視点で解析を行う
- 次世代計算基盤に向けてどのようなアルゴリズムのクラスが大幅な進化が見込まれるか、の指標も抽出して今後のアプリ発展につなげる



スケジュール

	2022 Q3	2022 Q4	2023 Q1	2023 Q2	2023 Q3	2023 Q4	2024 Q1
アーキテクチャ	テクノロジ・アーキ技術の調査・検討			ベンチマーキングによる性能解析/予測		アーキ再検討	
システムソフトウェア	既存ツールや利用動向調査			新規開発ソフト項目検討・定量的評価		将来の要件整理	
アプリケーション	アプリ調査とベンチマーク設計			ベンチマーク評価に基づく性能分析		アルゴリズム最適化検討	



プロセッシングエレメントの検討要素技術例

研究体制：ALL Japan＋メジャー国外ベンダーによる実施体制

GL: グループリーダ
AD: アドバイザー
SGL: サブグループリーダ
※赤字は開始後に追加予定機関

システム調査研究チーム
（代表機関）理化学研究所
【代表者：近藤， AD: 松岡(理研)】

アーキテクチャ調査研究グループ

アーキテクチャ調査研究グループ
取りまとめ担当:（代表機関）理化学研究所◎
【GL: 佐野, 副GL: 三輪(電通大), AD: 天野(慶大)】

- **アーキテクチャ調査研究サブG1**
  （理化学研究所）
  【SGL: 泰地】
- **アーキテクチャ調査研究サブG2**
  （分担機関）富士通株式会社
  【SGL: 新庄】
- **アーキテクチャ調査研究サブG3**
  （分担機関）インテル株式会社
  【SGL:矢澤】
- **アーキテクチャ調査研究サブG4**
  （分担機関）日本AMD株式会社
  【SGL:吉田】
- **アーキテクチャ調査研究サブG5**
  （協力機関）NVIDIA Corporation
  【SGL:Wells】
- **アーキテクチャ調査研究サブG6**
  （協力機関）Hewlett Packard Enterprise
  【SGL:根岸】
- **アーキテクチャ調査研究サブG7**
  （協力機関）海外ベンダー【予定】
  【SGL: 未定】

アーキテクチャ調査研究グループ

システムソフト・ライブラリ調査研究グループ

システムソフト・ライブラリ調査研究グループ
取りまとめ担当:（代表機関）理化学研究所◎
【GL: 佐藤賢斗, 副GL: 片桐(理研), AD : 佐藤(理研)】

システムソフト取りまとめ補助
（協力機関）名古屋大学
【機関代表: 片桐】

- **スケジューラ・ランタイムサブG**
  （分担機関）東北大学
  【SGL: 滝沢】
- **IO・ストレージ・ファイルシステムサブG**
  （分担機関）筑波大学
  【SGL: 建部】
  - **ストレージアーキ・パターン調査**
    （協力機関）株式会社DDNジャパン
    【機関代表: 橋爪】
- **OS・仮想化・クラウド連携サブG**
  （協力機関）国立情報学研究所
  【SGL: 竹房】
- **HPC利用環境サブG**
  （分担機関）大阪大学
  【SGL: 伊達】
- **通信ライブラリサブG**
  （分担機関）九州大学
  【SGL: 南里】
- **コンパイラ・プログラミングモデルサブG**
  （理化学研究所）
  【SGL: 辻】
- **数値ライブラリサブG**
  （理化学研究所）
  【SGL: 今村】
- **AIフレームワークサブG**
  （理化学研究所）
  【SGL: Mohamed】

システムソフト・ライブラリ調査研究グループ

アプリケーション調査研究グループ

アプリケーション調査研究グループ
取りまとめ担当:（分担機関）北海道大学◎
【GL: 岩下, 副GL : 高橋(筑波大),深沢(京大), AD: 中島・富田(理研)】

アプリ調査研究取りまとめ補助
（協力機関）京都大学
【機関代表: 深沢】

- **生命科学分野サブG**
  （分担機関）横浜市立大学
  【SGL: 寺山】
- **新物質・エネルギーサブG**
  （分担機関）物質・材料研究機構
  【SGL: 山地, 副SGL 】
- **気象・気候サブG**
  （分担機関）海洋開発研究機構
  【SGL: 小玉】
- **地震・津波防災**
  （分担機関）東京大学
  【SGL: 藤田】
- **ものづくり分野サブG**
  （理化学研究所）
  【SGL: 大西】
  - **ものづくり分野協力**
    （協力機関）宇宙航空研究開発機構
    【機関代表: 未定】
- **基礎科学分野サブG**
  （理化学研究所）
  【SGL: 青木】
  - **宇宙・惑星アプリ分野協力**
    （協力機関）国立天文台
    【機関代表: 瀧脇】
- **社会科学分野サブG**
  （理化学研究所）
  【SGL: 楳本】
- **デジタルツイン・Society5.0分野サブG**
  （分担機関）東京大学
  【SGL: 下川辺】
  - **デジタルツイン・Society5.0分野補助**
    （協力機関）日本原子力研究開発機構
    【機関代表: 小野寺】
- **科学技術計算アルゴリズムサブG**
  （分担機関）筑波大学
  【SGL: 高橋】
- **機械学習アルゴリズムサブG**
  （分担機関）東京工業大学
  【SGL: 横田】
- **ベンチマーク構築サブG**
  （理化学研究所）
  【SGL: 村井】
- **性能モデリングサブG**
  （理化学研究所）
  【SGL: Domke】

アプリケーション調査研究グループ

次世代計算基盤構築に向けたスケジュール予測

富岳Next スケジュール予測

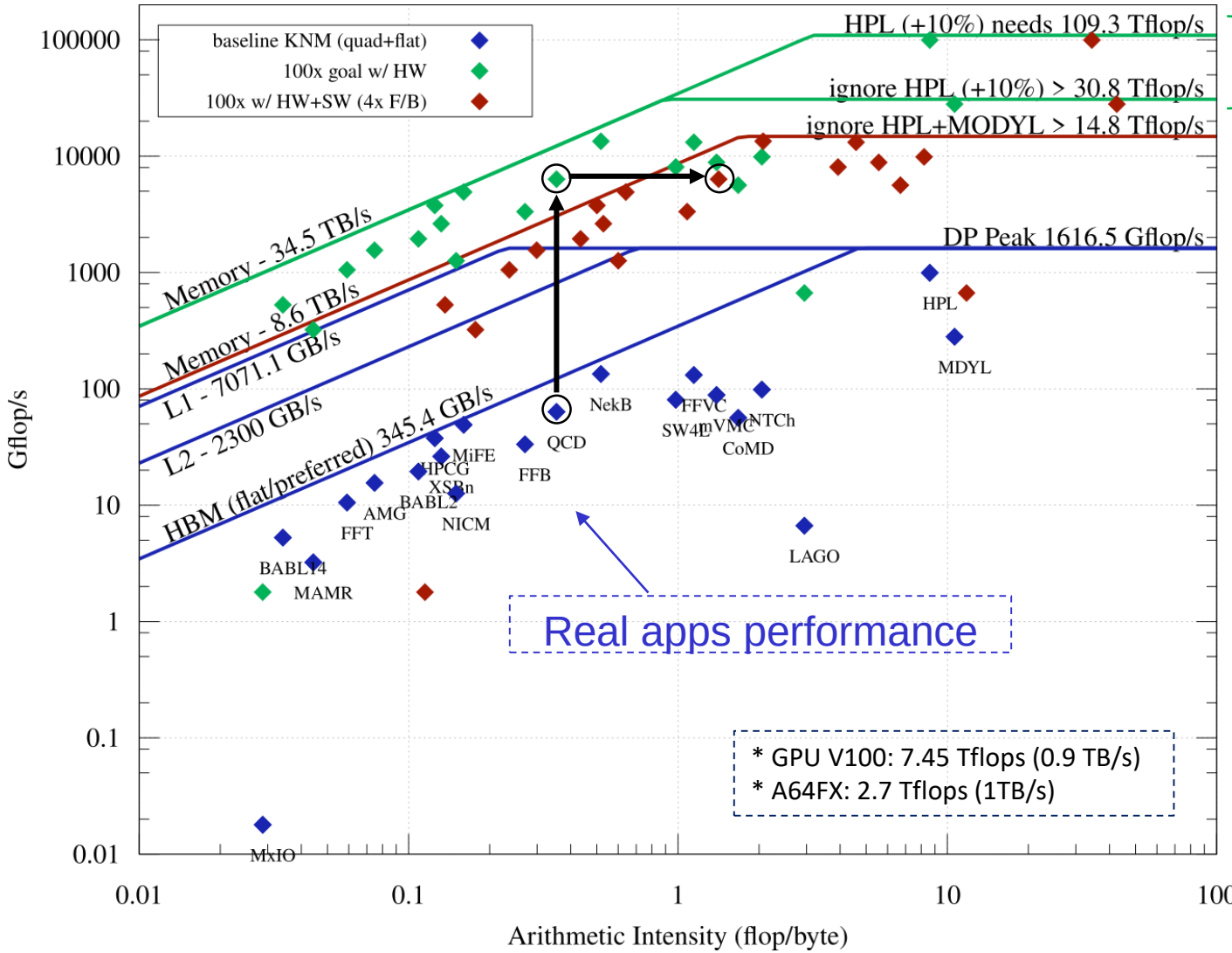
2020 R2	2021 R3	2022 R4	2023 R5	2024 R6	2025 R7	2026 R8	2027 R9	2028 R10	2029 R11	2030 H12	2031 H12
NGACI WP1.0	WP1.1										
FS1.5年版		FS公募	調査研究	開発PJ (基本設計)		開発PJ (詳細設計)			開発PJ (製造・設置・調整)		
FSを2.5年にした場合		FS公募	調査研究		開発PJ (基本設計)	開発PJ (詳細設計)			開発PJ (製造・設置・調整)		
半導体プロセス予測		TSMC N5	TSMC N4	TSMC N3		TSMC 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

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

HPCワークロードのルーフライン解析

- 富岳の100倍の性能を達成するには？



← ハードウェアだけで達成する場合

← ハードウェアおよびアプリ協調

← 現行のA64FXの性能

- ハードウェアへの要求
  - DP性能: 15 Tflops (A64FXの10倍)
  - メモリンド幅: 8.6 TB/s (A64FXの25倍)
- アプリの協調: 低精度演算の積極的利用による演算強度増

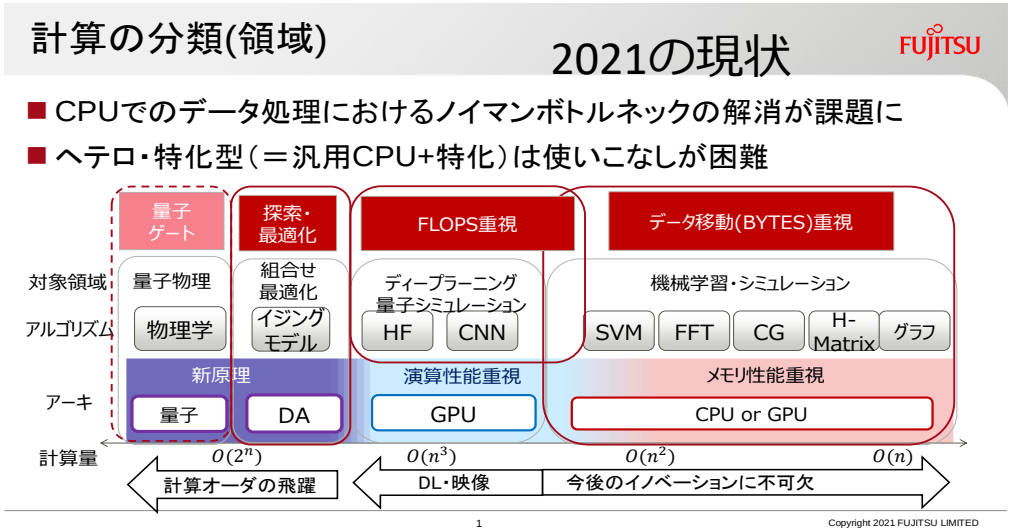
出典：松岡R-CCSセンター長ら

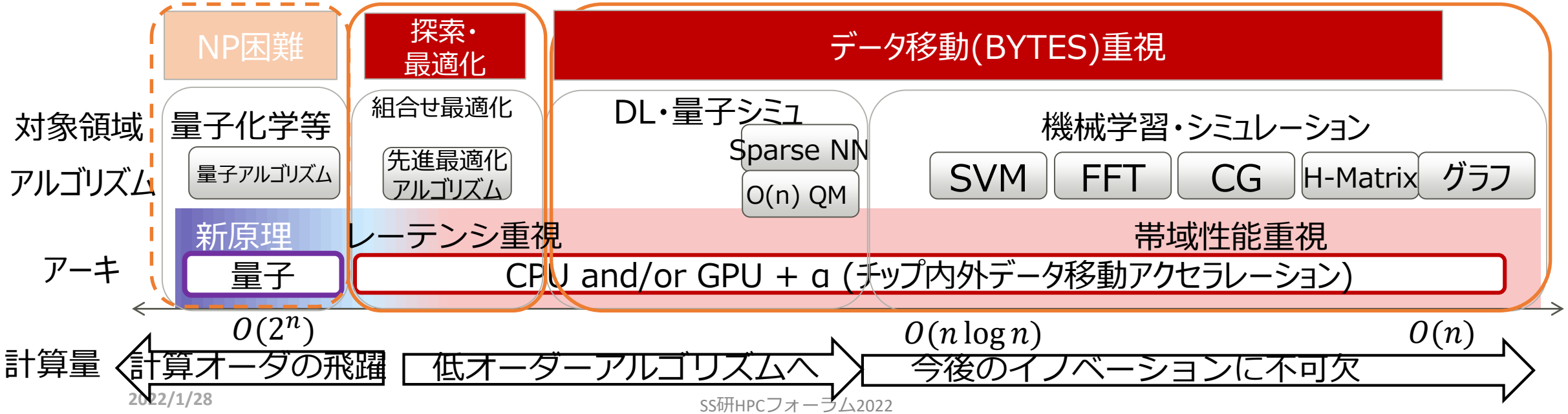
ムーアの法則の終焉に従い計算の主体はO(*n*)データ移動へ

※出典：松岡聡 理研計算科学センター長

- **2030へ向けた変遷 (Post-Mooreの入り口)**

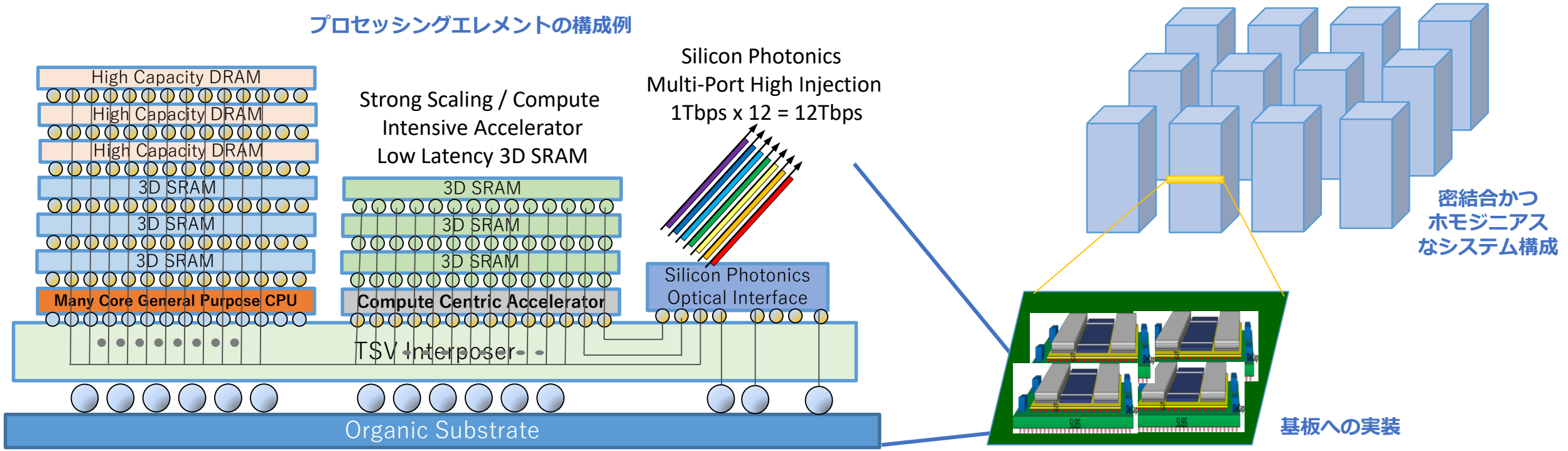
- ムーア則の終焉による、演算性能の進化の終焉
- 新デバイス・パッケージングによるデータ移動コストの革新的削減
- アルゴリズムや利用法の進化による、計算量オーダーの削減
  (＋データ移動の相対的な要求の増加)
- BD/AI/第一原理の融合による本質的なデータ中心の流れ





次世代計算基盤として目指すシステムの方向性

- **データ移動効率化指向のアーキテクチャ・アルゴリズムの一例**
  - 三次元積層メモリ技術を駆使した相対メモリバンド幅の大幅な向上
  - シリコンフォトニクスによるリモートメモリアクセスへの高バンド幅確保
  - 強スケーリング実行での実行効率の確保：データフロー型の導入などによる低遅延実行
  - 混合精度演算の積極的利用とアーキ・ソフトウェアからのサポート



「３次元メモリ技術」＋「光技術」：日本の技術力を結集して破壊的イノベーションへ

技術課題と研究開発ロードマップ（デバイス・アーキ）

- 計算処理ハードウェア
  - 電力効率の改善
  - SIMDベクトル長の拡大への対策
    - テンソル・プロセッシング向けの2次元PEアレイの検討、プログラミング環境の強化
  - アクセラレータアーキテクチャの検討
    - コデザインによる幅広いアプリに適用可能なアクセラレータとそのシステム搭載方式の検討
    - データ転送部分を最適化できるアクセラレータの検討
- メモリ
  - 広帯域と大容量メモリの3D積層に向けたパッケージ技術、歩留まりの改善
  - NVMやSCMといった新しいメモリ技術の導入
- インターコネクト
  - シリコンフォトニクスやコパッケージドオプティクスなどの光伝送技術の実用化

技術課題と研究開発ロードマップ（システムソフトウェア）

- 基盤ソフトウェア
  - 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
  - 不揮発性メインメモリ（NVDIMM）のサポート
- 大規模並列・高性能計算
  - CPUとアクセラレータの効率的に利用と統一的な管理
  - データ転送の最適化、NVMを用いたチェックポインティング、ルーティングや集団通信最適化
  - 非同期通信の進捗処理
- プログラミング環境
  - Oversubscriptionプログラミング、マルチレイヤコンパイラ、ワークフロー向けプログラミング
- 次世代システム向け各種フレームワーク
  - データフレームワーク、プロファイラ、電力管理、スケジューラ、C/R
- 新しいシステムコンセプト向けのシステムソフトウェア
  - 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、ワークフロー型の計算

技術課題と研究開発ロードマップ（ライブラリ・アルゴリズム）

- 自動チューニング
  - 超非均質プロセッサの対応、混合精度演算対応、Society 5.0アプリ向けチューニング
  - AIと自動チューニング技術の融合
- 耐故障機構
  - アプリケーションレベルでのC/R、代替計算継続・縮退リソース計算継続方式
- 機械学習フレームワーク
  - 新アーキ向け各種フレームワーク改良・ライブラリの高速化
  - 外部計算資源との連携、シミュレーションとの連携
- アルゴリズム
  - 特定問題のアクセラレーション（グラフ処理など）
- 新計算原理との融合
  - 従来型計算機と量子計算機のハイブリッド利用

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
  - 大規模データの流通、重要データの保護、ディザスタリカバリ
  - クラウドとの連携
- 設備・管理
  - カーボンニュートラル化検討、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

おわりに：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

新応用分野の開拓の例

- 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

エコシステム構築と長期的な人材育成