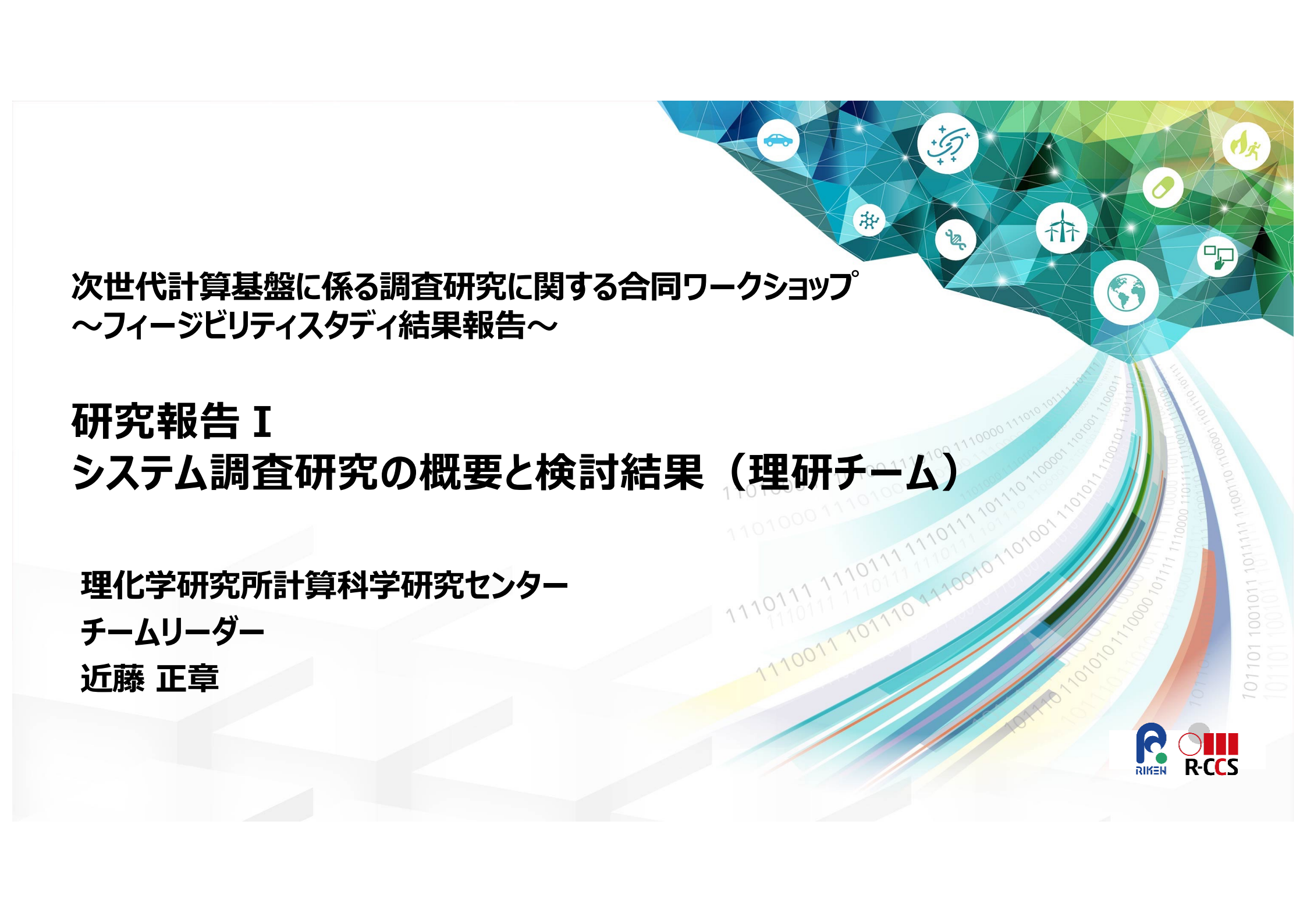




次世代計算基盤に係る調査研究に関する合同ワークショップ° ～フイージビリティスタディ結果報告～

研究報告Ⅰ システム調査研究の概要と検討結果（理研チーム）

理化学研究所計算科学研究センター
チームリーダー
近藤 正章



取組概要

調査研究の全体概要（理化学研究所：システム調査研究）



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

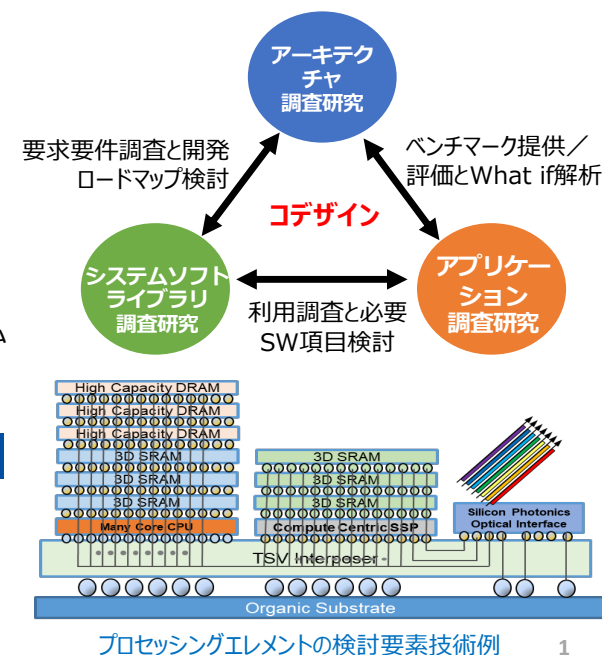
- 半導体技術・パッケージング技術動向を踏まえつつ、**システム全体やその構成要素について考え得る技術的可能性や総合性能を調査**（新規要素技術として三次元積層メモリ技術、強スケーリング・計算インテンシブ向けアクセラレータ、チップ間直接光通信技術などを意識）
- アプリケーション調査研究G提供の**ベンチマークセットの性能解析に基づき将来システムの性能を予測**、また次世代アプリ開発へとフィードバック

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

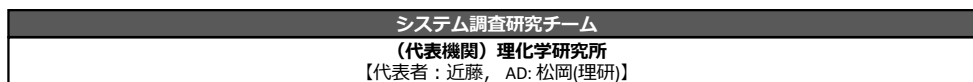
- **複数アーキテクチャを統一的に評価するための広範なベンチマークセットを構築**し、それを利用したアーキテクチャ評価結果を踏まえてアルゴリズムやパラメータの改善を検討しつつベンチマークセットを更新しながら、**性能モデルも利用した探索的な“What if”の視点で解析**を行う
- 次世代計算基盤に向けてどのようなアルゴリズムのクラスが大幅な進化が見込まれるか、の指標も抽出して今後のアプリ発展につなげる



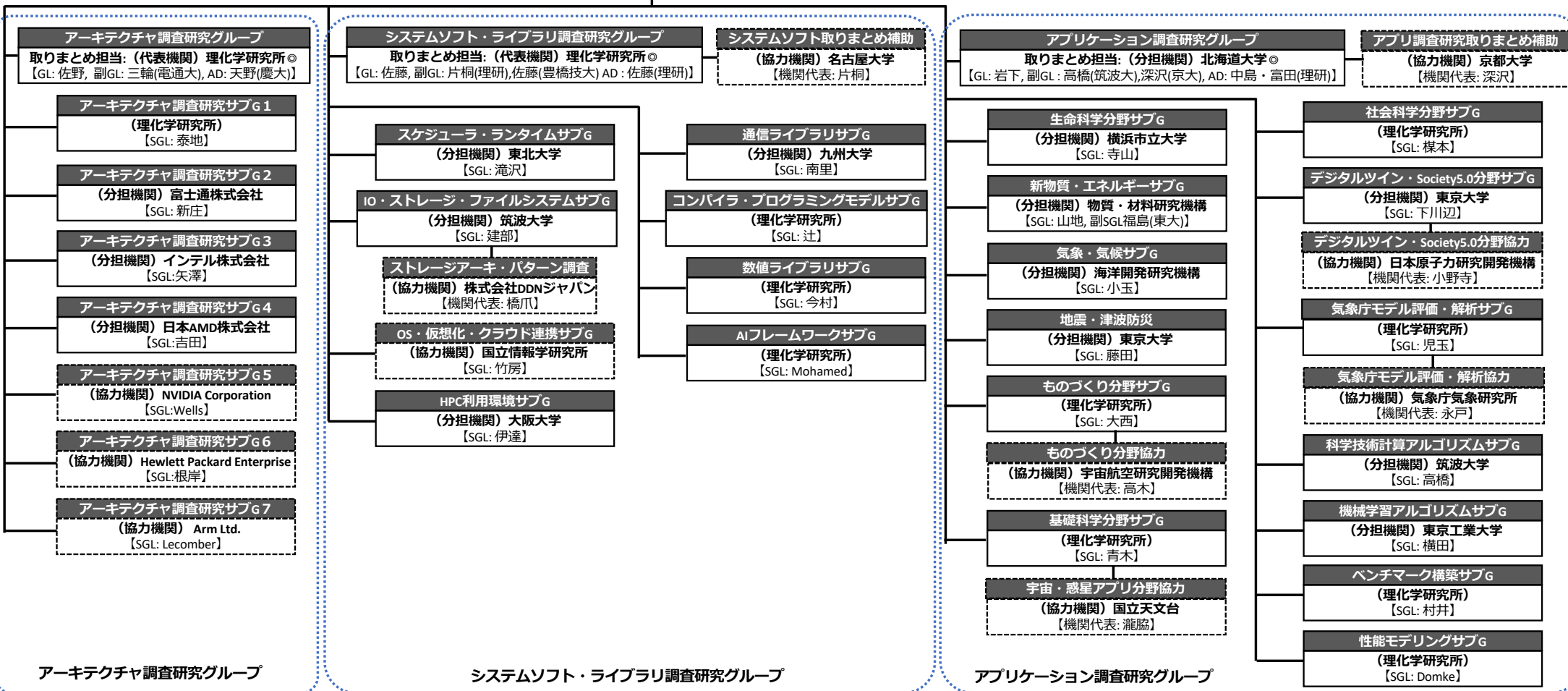
スケジュール

	2022 Q3	2022 Q4	2023 Q1	2023 Q2	2023 Q3	2023 Q4	2024 Q1
アーキテクチャ	テクノロジー・アーキ技術の調査・検討						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	

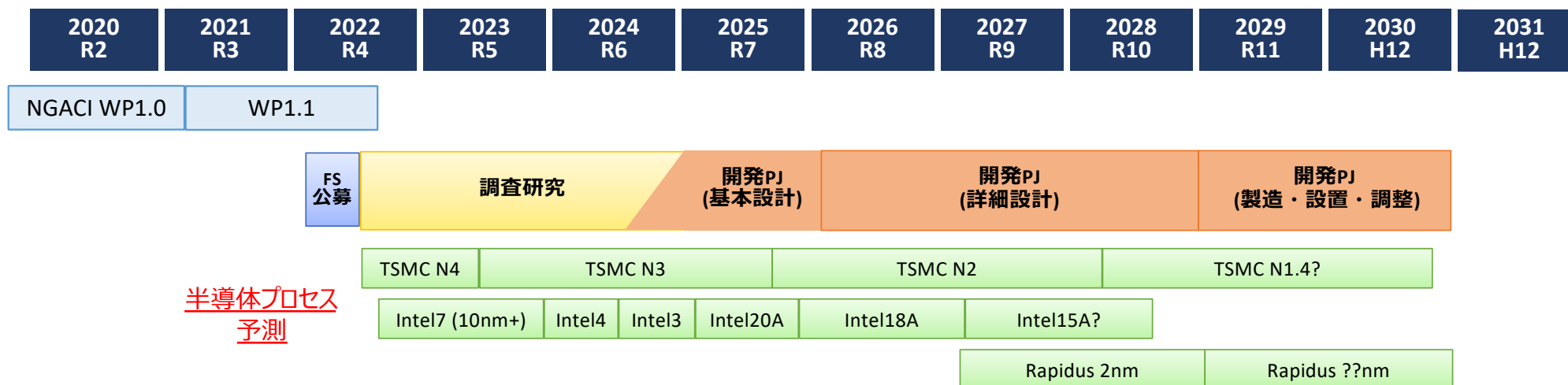
研究体制：ALL Japan+メジャー国内外ベンダーによる実施体制



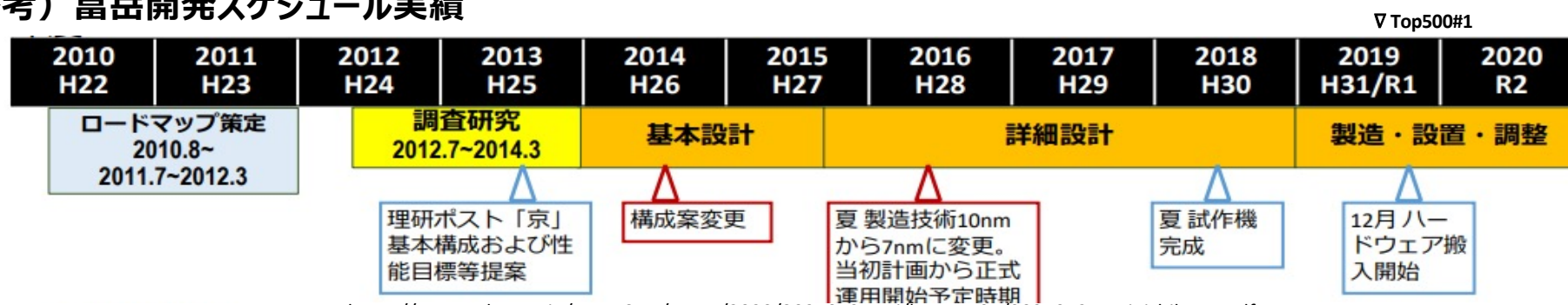
GL: グループリーダー
AD: アドバイザー
SGL: サブグループリーダー



富岳Next スケジュール予測



(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

- 現状で妥当と考えられる満たすべきシステム全体の要求仕様

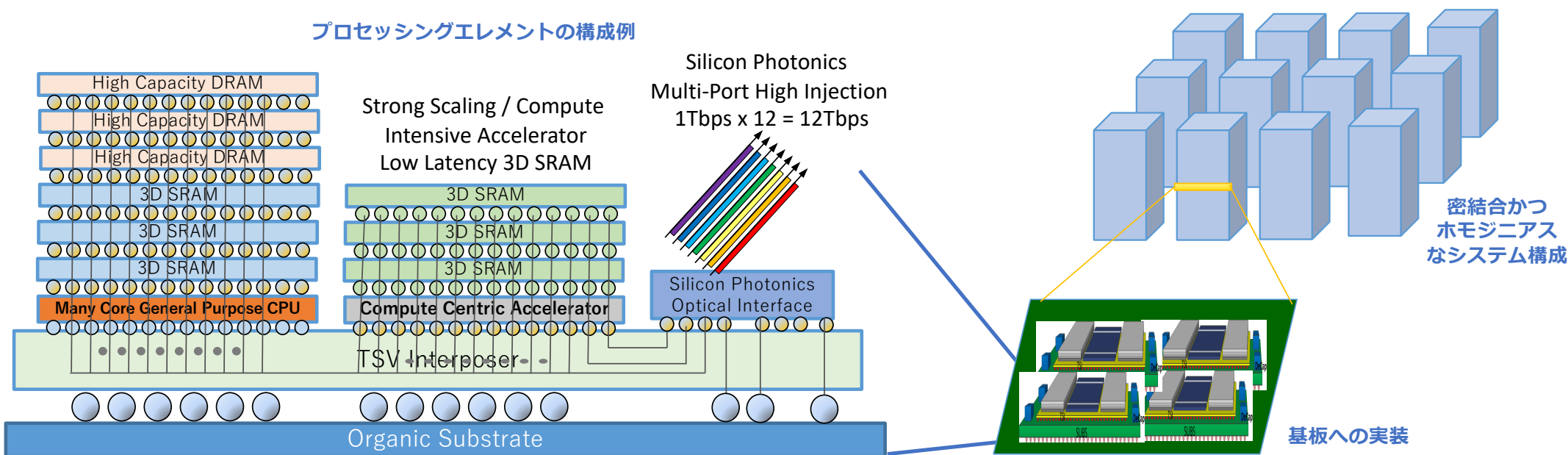
項目	CPU	加速部
合計ノード数	3400ノード以上	
FP64ベクトル性能	48PFLOPS以上	3.0EFLOPS以上
FP16/BF16行列演算性能	1.5EFLOPS以上	150EFLOPS以上
FP8行列演算性能	3.0EFLOPS以上	300EFLOPS以上
FP8行列演算性能（sparsity考慮）	—	600EFLOPS以上
メインメモリサイズ	10PiB以上	10PiB以上
メインメモリバンド幅	7PB/s以上	800PB/s以上
合計消費電力	40MW以下（計算ノードおよびストレージ）	

各グループの調査・検討の概要

当初のビジョン：次世代計算基盤として目指すシステムの方方向性

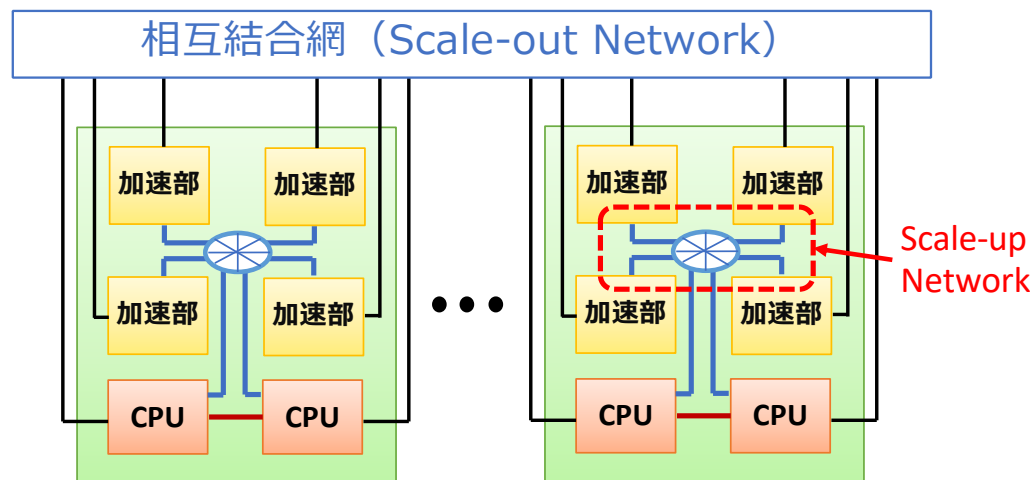
- データ移動効率化指向のアーキテクチャの一例
- 三次元積層メモリ技術を駆使した相対メモリバンド幅の大幅な向上
- シリコンフォトニクスによるリモートメモリアクセスへの高バンド幅確保
- 強スケーリング実行での実行効率の確保：データフロー型の導入などによる低遅延実行
- 混合精度演算の積極的利用とアーキ・ソフトウェアからのサポート

プロセッシングエレメントの構成例



「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

- シミュレーションとAIの融合に向けたシステムアーキテクチャの想定
 - CPUと加速部を高速リンクで接続し計算ノードを構成
 - 高メモリバンド幅のシステム+データ移動効率化かつプログラミングの高生産性が重要

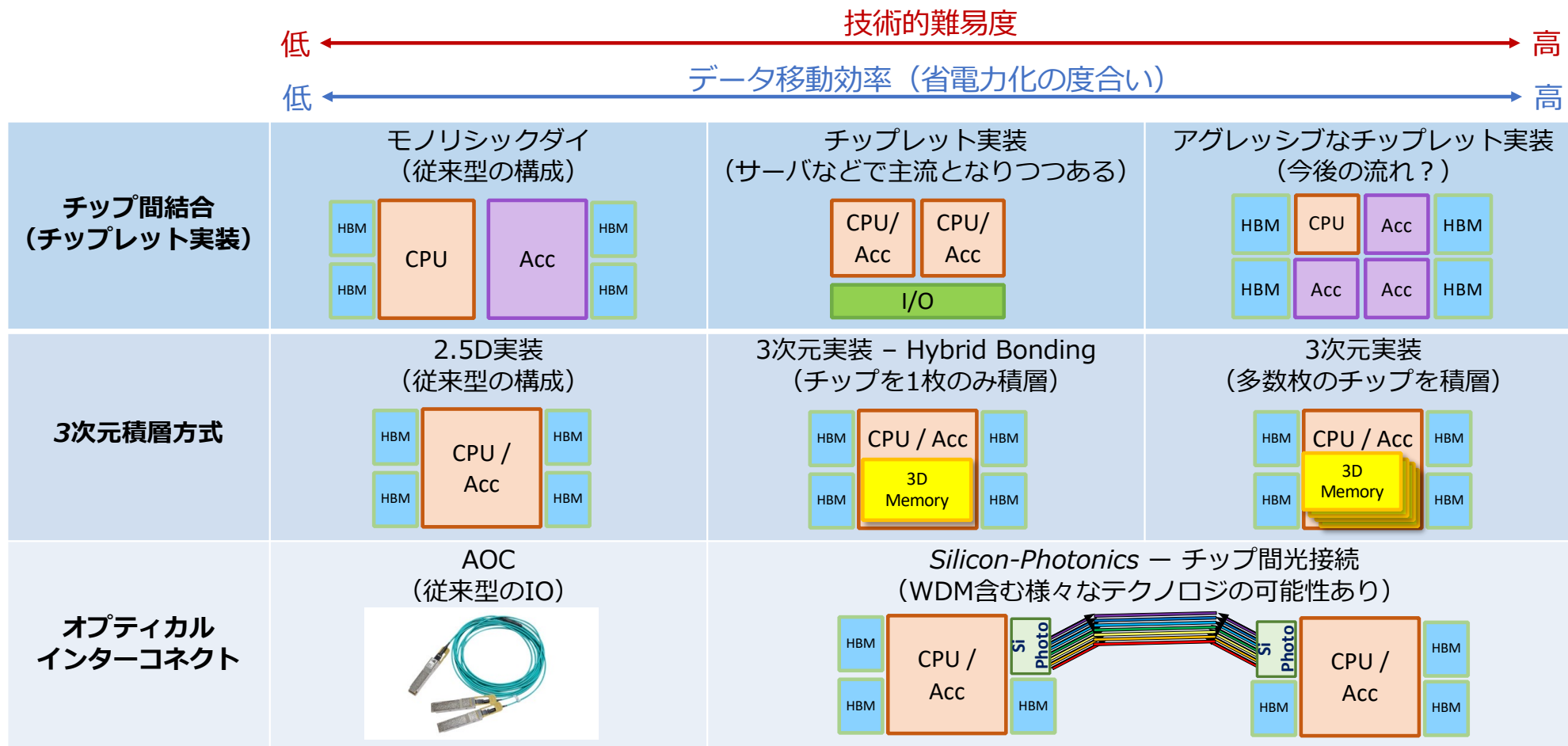


- CPUと加速部のノード内構成
 - 2CPU+4GPUをベース案として今後決定
- Scale-upネットワーク（ノード内外加速部接続）
 - 高速リンクによる全結合のP2P/スイッチ接続
- Scale-outネットワーク（ノード間相互結合網）
 - 多段のスイッチを用いた間接網接続
- メモリ
 - 導入時期における先端メモリ技術を利用
 - DDR系の大容量メモリも搭載

- 強・弱スケーリング双方に適するシステムネットワーク
 - Scale-up/Scale-outネットワークを組み合わせ
 - 加速部間の直接ネットワークリンク接続
- システム全体で数万規模の加速部ソケットを搭載

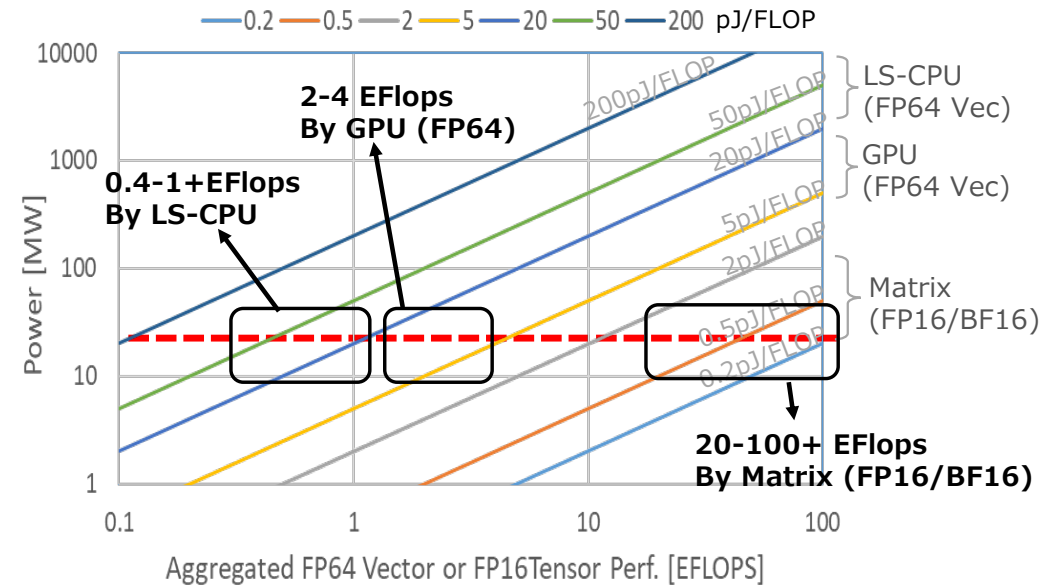
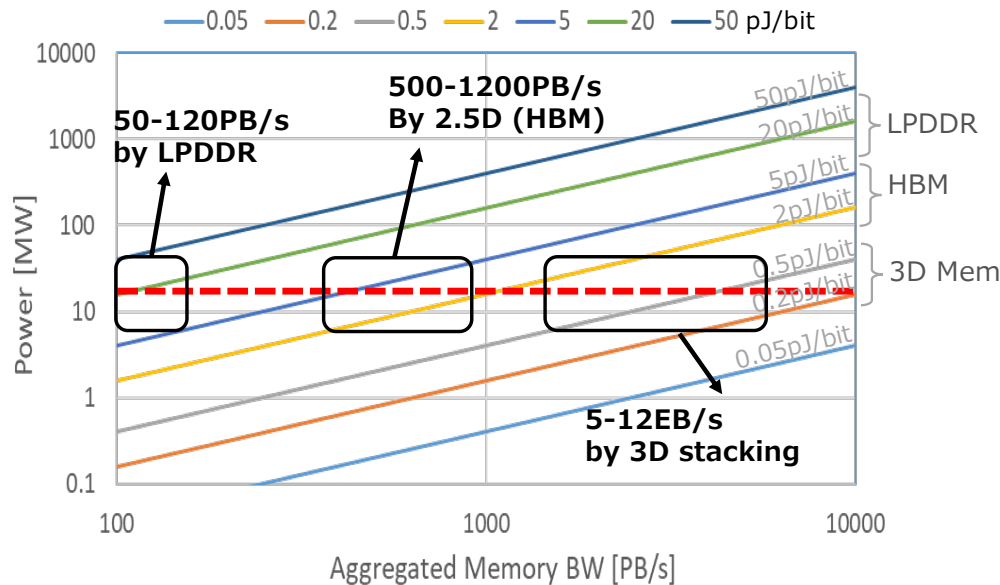
ノードアーキテクチャの実装方式の検討

● パッケージング技術の検討



電力制約下でのシステム性能の見積り

- 将来のデバイステクノロジーにおける演算／データ移動エネルギーの見積り
 - 過去のトレンドおよび公開資料のデータを基に検討
 - 参画ベンダーの予測等とは独立に見積りを実施
- 40MWの電力制約を仮定した場合（メモリアクセス電力資源：20MW、計算処理電力資源：20MW）
 - 相互結合網、ストレージ、冷却などの電力は評価対象外（実際にはそれらも重要）
 - コストや発熱密度など他の制約により実際は見積りよりも厳しくなる可能性あり



調査目的と調査概要

● 調査目的：

- システムソフトウェアの現状と将来の動向を調査し、今後の日本の次世代計算基盤として行うべきソフトウェア開発について調査研究

● 調査概要：

● 【調査項目1】システムソフトウェアの動向調査と利用状況調査

- 移植性、生産性を加味して現在のシステムソフトウェアの動向と将来トレンドの調査
- 大規模システムにおけるシステムソフトウェアの利用状況調査

● 【調査項目2】システムソフトウェア開発戦略に関わる調査

- 国内で独自に開発 or 海外連携により開発
- プロプライエタリソフトウェアとして開発 or オープンソースソフトウェアとして開発
- 新規開発 or 既存のソフトウェアエコシステムを活用（移植、チューニング）した開発

● 【調査項目3】類似ソフトウェアの比較調査

- 類似ソフトウェアに関しては比較調査
- 類似ソフトウェア内で最適なソフトウェア選択と代替可能ソフトウェア群の明確化

システムソフトウェアの動向と利用状況調査

+

候補アーキテクチャ

HPC利用状況調査GやアプリGからの要望に基づき、どのようなシステムソフトウェア・ライブラリを開発/移植・チューニングすべきかを決定

		アーキ候補1 (x86)	アーキ候補2 Arm9	アーキ候補3 Arm9+X(Chiplet)	アーキ候補4 x86+GPU	アーキ候補5 ...
コンパイラ	gcc	(1-a)				
	LLVM/clang	(1-b)	N/A			
	icc					
	新規コンパイラ		(2-a)			
低級 数値ライブラリ	OneMKL	(1-b)				
	OpenBLAS					
	Eigen					
	SSL-III		(2-a)			
高級 数値計算ライブラリ	Numpy	(1-a)	(1-d)			
	Scipy					
	cuNumeric					
	...					

国内で開発すべきか否かを検討するための調査

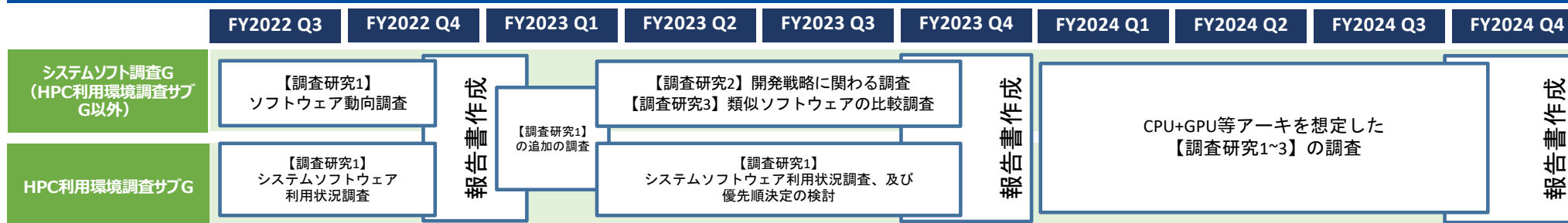
複数の類似システムソフトがある場合、差異の調査

(1-a) OSSコミュニティによる開発
(1-b) 海外企業による独自開発
(1-c) OSS（工数の低い移植：As-is or 軽微な移植）
(1-d) OSS（工数の高い移植：チューニング）
(1-e) 国際共同開発
(2-a) 国内独自開発（重点開発 or 富士ソフトの移植）
(2-b) 開発未定

（主要な）システムソフトに関しては、異なるアーキテクチャでの性能評価

評価指標：
開発コスト、生産性（アプリ開発のしやすさ）
商用展開・技術展開

調査研究計画



- **コンパイラ・プログラミングモデル**：「富岳」におけるコンパイラの利用状況の解析。各候補アクセラレータに対するプログラミング環境についての調査。LLVM のFortran コンパイラであるFlang について開発状況。OSSコンパイラの命令生成性能を調査を実施
- **スケジューラ・ランタイム**：システム大規模化に向けたジョブスケジューラの負荷分散に関してさらに調査。京や富岳におけるジョブスケジューラ開発の実績に基づいて今後に向けて検討すべき課題（コンテナ連携、ワークフロー、外部連携、実行支援ツール）を実施
- **通信ライブラリ**：スケーラブルな並列計算の実現に向けて必要となる新しい通信ライブラリのインタフェースや実装技術を調査。インターコネクトベンダーと連携し今後の動向や通信ライブラリの対応状況とそのアプリケーションレベルでの活用手段の調査を実施
- **I/O・ストレージ・ファイルシステム**：各種アプリケーションのアクセスパターン、I/O要求の調査をものに、次期ストレージアーキテクチャについての考察のまとめを行った
- **数値ライブラリ**：数値計算ライブラリ（密線形代数、疎反復ソルバ、疎直接ソルバもしくは階層行列ソルバ、高速フーリエ変換等）における、数値ライブラリの動向調査、性能評価、開発項目選定に向けた調査検討を実施
- **AIフレームワーク**：性能評価のためのAIモデル学習・推論ベンチマーク開発。AI関連のソフトウェアの調査と実機でのテスト
- **OS・仮想化・クラウド連携**：仮想化技術、特にHPC分野におけるコンテナ仮想化技術の評価および技術要件の整理と欧米スパコンセンターにおけるコンテナ技術の導入状況の調査を実施
- **HPC利用環境**：HPCI構成機関を対象としてスーパーコンピューティングシステムの実際の運用に携わる管理者に対して直接対面のヒアリングによる利用環境の調査を実施
- **分野横断の技術調査（セキュリティ、自動チューニング）**：セキュリティに関し、富岳運用におけるセキュリティに関する調査、高いセキュリティが求められるアプリ調査、機密コンピューティング技術のHPC適応状況の調査、ユーザーから透過的なメモリの暗号化による性能オーバーヘッド調査、自動チューニングに関し分野横断的な自動チューニングの最新技術調査を実施
- **全体**：次世代計算基盤におけるソフトウェア開発戦略策定の一環として、富岳でのソフトウェア利用状況の統計的な解析を行った。また、FugakuNext Software Matrixとしてまとめた

ポスト富岳 Software Matrixの作成

Software/Library				CPU Architectures																GPU												Interconnect												
Role	Standard	Name	Usage	Intel x86				AMD x86				NVIDIA Arm (Grace)				Fujitsu Arm				Intel				AMD				NVIDIA				InfiniBand				SlingShot				ToFu				
役割	標準化API名	ソフトウェア名	利用用途	MJ / NHJ	Prop / OSS	New / CDCI	MJ / NHJ	Prop / OSS	New / CDCI	MJ / NHJ	Prop / OSS	New / CDCI	MJ / NHJ	Prop / OSS	New / CDCI	MJ / NHJ	Prop / OSS	New / CDCI	MJ / NHJ	Prop / OSS	New / CDCI	MJ / NHJ	Prop / OSS	New / CDCI	MJ / NHJ	Prop / OSS	New / CDCI	MJ / NHJ	Prop / OSS	New / CDCI	MJ / NHJ	Prop / OSS	New / CDCI	MJ / NHJ	Prop / OSS	New / CDCI	MJ / NHJ	Prop / OSS	New / CDCI	MJ / NHJ	Prop / OSS	New / CDCI		
#workload-manager	N/A	Slurm	High	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NS	NS	NS	NS	NS	NS	NS	
#workload-manager	N/A	flux	Low	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NS	NS	NS	NS	NS	NS	NS	
#high-level-communication-library	MPI	Fujitsu MPI	High	MJ	Prop	CDCI	UK	UK	UK	UK	UK	UK	MJ	Prop	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	MJ	Prop	CDCI	NS	NS	NS	MJ	Prop	CDCI	NS	NS	NS	MJ	Prop	CDCI	NS	NS	NS	NS	
#high-level-communication-library	MPI	Intel MPI	Mid	NHJ	Prop	CDCI	NHJ	Prop	CDCI	NS	NS	NS	NS	NS	NS	UK	UK	UK	UK	UK	UK	UK	UK	NHJ	Prop	CDCI	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	
#high-level-communication-library	MPI	NVIDIA HPC-X	Mid	NHJ	Prop	CDCI	NHJ	Prop	CDCI	NS	NS	NS	NS	NS	NS	UK	UK	UK	UK	UK	UK	UK	UK	NHJ	Prop	CDCI	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	
#high-level-communication-library	MPI	Open MPI	Mid	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	NS	NS	NS	UK	UK	UK	UK	UK	UK	UK	UK	NHJ	OSS	CDCI	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	
#high-level-communication-library	MPI	HYBRISCH	Mid	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	NS	NS	NS	UK	UK	UK	UK	UK	UK	UK	UK	NHJ	OSS	CDCI	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	
#high-level-communication-library	MPI	MPICH	Mid	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	NS	NS	NS	UK	UK	UK	UK	UK	UK	UK	UK	NHJ	OSS	CDCI	NS	NS	NS	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	
#middle-level-communication-library	UCX	UCX	Mid	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	NS	NS	NS	UK	UK	UK	UK	UK	UK	UK	UK	NHJ	OSS	CDCI	NHJ	OSS	CDCI	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	
#middle-level-communication-library	UCC	UCC	Low	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	NS	NS	NS	UK	UK	UK	UK	UK	UK	UK	UK	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	
#middle-level-communication-library	OFED	OFED	Low	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	NS	NS	NS	UK	UK	UK	UK	UK	UK	UK	UK	NHJ	OSS	CDCI	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	
#light-weight kernel	Linux	McKernel	High	MJ	OSS	CDCI	UK	UK	UK	UK	UK	UK	MJ	OSS	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	
#OS	Linux	TOSS	Low	NHJ	OSS	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	
#container	container	Docker/Podman/Singularity/Charis/Clusify	High	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	NHJ	OSS	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	MJ	OSS	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#compiler	Fortran	GNU Compiler Collection Fortran	High	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	MJ	OSS	CDCI	NHJ	Prop	UK	NHJ	OSS	UK	NHJ	OSS	None (partially supported)		NHJ	OSS	None (partially supported)		UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#compiler	C	GNU Compiler Collection C	High	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	MJ	OSS	CDCI	NHJ	OSS	UK	NHJ	OSS	UK	NHJ	OSS	None (partially supported)		NHJ	OSS	None (partially supported)		UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#compiler	C++	GNU Compiler Collection C++	High	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	MJ	OSS	CDCI	NHJ	OSS	UK	NHJ	OSS	UK	NHJ	OSS	None (partially supported)		NHJ	OSS	None (partially supported)		UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#compiler	Fortran	LLVM/Flang	High	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	MJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#compiler	C	LLVM/Clang	High	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	MJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#compiler	C++	LLVM/Clang++	High	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	MJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#debugger	-	Arm Allinea Studio, Arm Forge	Low	NHJ	Prop	CDCI	NHJ	Prop	CDCI	UK	UK	UK	NHJ	Prop	CDCI	NHJ	Prop	NS	NHJ	Prop	CDCI	NHJ	Prop	CDCI	NHJ	Prop	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#profiler	-	TotalView	Mid	NHJ	Prop	CDCI	NHJ	Prop	CDCI	UK	UK	UK	NS	NS	NS	NHJ	Prop	UK	NHJ	Prop	CDCI	NHJ	Prop	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#compiler	C	Arm Allinea Studio, Arm C-Compiler for Arm Allinea Studio	Mid	NS	NS	NS	NS	NS	NS	NS	NS	NS	UK	UK	UK	NHJ	Prop	CDCI	NHJ	Prop	UK	NHJ	Prop	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#compiler	C++	Arm Allinea Studio, Arm C++-Compiler for Arm Allinea Studio	Mid	NS	NS	NS	NS	NS	NS	NS	NS	NS	UK	UK	UK	NHJ	Prop	CDCI	NHJ	Prop	UK	NHJ	Prop	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#compiler	C	Arm Allinea Studio, Arm Fortran-Compiler for Arm Allinea Studio	Mid	NS	NS	NS	NS	NS	NS	NS	NS	NS	UK	UK	UK	NHJ	Prop	CDCI	NHJ	Prop	UK	NHJ	Prop	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#compiler	oneAPI	Intel® oneAPI C-compiler	High	NHJ	Prop	CDCI	UK	UK	UK	UK	UK	UK	NS	NS	NS	NHJ	Prop	CDCI	NHJ	Prop	UK	NHJ	Prop	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#compiler	oneAPI, SYCL	Intel® oneAPI DPC++/C++-compiler	High	NHJ	Prop	CDCI	UK	UK	UK	UK	UK	UK	NS	NS	NS	NHJ	Prop	CDCI	NHJ	Prop	UK	NHJ	Prop	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#compiler	oneAPI	Intel® oneAPI Fortran compiler	High	NHJ	Prop	CDCI	UK	UK	UK	UK	UK	UK	NS	NS	NS	NHJ	Prop	CDCI	NHJ	Prop	UK	NHJ	Prop	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#compiler	oneAPI, SYCL	Intel® oneAPI DPC++-compiler	Mid	NHJ	Prop	CDCI	UK	UK	UK	UK	UK	UK	NS	NS	NS	NHJ	Prop	CDCI	NHJ	Prop	UK	NHJ	Prop	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#compiler	Python	Intel Distribution for Python	High	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	NS	NS	NS	NHJ	OSS	CDCI	NHJ	OSS	UK	NHJ	OSS	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#compiler	Python	Intel Python	High	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	NS	NS	NS	NHJ	OSS	CDCI	NHJ	OSS	UK	NHJ	OSS	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#communication-tool	SYCL	Data Parallel Control	Low	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	NS	NS	NS	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#debugger	-	GDB	High	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#debugger	-	Intel Distribution for GDB	Mid	NHJ	Prop	CDCI	UK	UK	UK	NS	NS	NS	NS	NS	NS	NHJ	Prop	CDCI	NHJ	Prop	NS	NHJ	Prop	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK
#debugger	-	Intel Inspector	Mid	NHJ	Prop	CDCI	UK	UK	UK	NS	NS	NS	NS	NS	NS	Prop	CDCI	CDCI	NHJ	Prop	NS	NHJ	Prop	CDCI	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	
#profiler	-	perf	Mid	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	NHJ	OSS	CDCI	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK									

AIに関するソフトウェアフレームワークの調査結果

- AIに関するソフトウェア環境についてIntel/AMD/NVIDIA社のGPUを例に調査
- DL-specific premitivesは各社専用のbackendが開発されている
- 主要AIソフトウェアは各アーキ案においてソフトウェア環境の差異は少ない

			Software	Owner	Open/Closed	Languages Supported	Hardware Supported	Link
Training/Evaluation	Frameworks & Distributed Training	Frameworks	PyTorch (Torch)	Meta	Open	Multiple	Various CPUs/GPUs	https://pytorch.org
			TensorFlow	Google	Open	Multiple	Various CPUs/GPUs/TPUs	https://www.tensorflow.org
			Apache MXNet	Apache Foundation	Open	Multiple	Various CPUs/GPUs	https://mxnet.apache.org/
			PaddlePaddle	PaddlePaddle	Open	Multiple	Various CPUs/GPUs	https://github.com/PaddlePaddle/Paddle
			NNABLA	Sony	Open	Multiple	Various CPUs/GPUs	https://nnabla.org
			RAY	Anyscale Inc.	Open	Multiple	Various CPUs/GPUs	https://www.ray.io
			JAX	Google	Open	Multiple	Various CPUs/GPUs/TPUs	https://jax.readthedocs.io/en/latest/#
			Colossal-AI	Colossal-AI	Open	Python	Nvidia GPUs	https://colossalai.org
		Wrapper frameworks	Keras (TF)	Google	Open	Python	Various CPUs/GPUs/TPUs	https://keras.io
			Gluon (MXNet)	Gluon	Open	Python	Nvidia GPUs	https://auto.gluon.ai/stable/index.html
			fastai	fast.ai	Open	Multiple	Various CPUs/GPUs	https://github.com/fastai/fastai
		Extensions	DGL (TF; PyTorch; MXNet)	DGL	Open	Python	Various CPUs/GPUs	https://www.dgl.ai
			Lightning (PyTorch)	Lightning AI	Open	Python	Various CPUs/GPUs	https://lightning.ai
			DeepSpeed	Microsoft	Open	Python	Various CPUs/GPUs	https://www.deepspeed.ai
		Communicators	Pytorch Geometric	Meta	Open	Python	Various CPUs/GPUs	https://pytorch-geometric.readthedocs.io
			Sonnet (TF)	DeepMind	Open	Python	Various CPUs/GPUs	https://sonnet.readthedocs.io/en/latest/#
			Horovod	Uber	Open	Python	Various CPUs/GPUs	https://horovod.readthedocs.io/en/stable/
		Backend (DL-specific primitives)	NCCL	Nvidia	Closed	Python	Nvidia GPUs	https://developer.nvidia.com/nccl
			RCCL	AMD	Open	C	AMD GPUs	https://rocm.docs.amd.com/projects/rccl/en/latest/
			cuDNN	Nvidia	Closed	CUDA	Nvidia GPUs	https://developer.nvidia.com/cudnn
		Numerical Libs & BLAS	oneDNN	Intel	Open	C++	Various CPUs/Intel GPUs	https://oneapi-src.github.io/oneDNN/group_dnnl_api.html
			hipDNN/MIOpen	AMD	Open	C++	Various CPUs/ AMD GPUs	https://github.com/ROCmSoftwarePlatform/hipDNN
			LIBXSMM	LIBXSMM	Open	C++	Various CPUs	https://libxsmm.readthedocs.io
	Model Repository	CUPy	Preferred Networks	Open	Python	Various GPUs	https://cupy.dev	
		Huggingface	Huggingface	Closed	N/A	N/A	https://huggingface.co	
		Keras Applications	Google	Open	Python	Various CPUs/GPUs/TPUs	https://keras.io/api/applications/	
	Model Exchange	ONNX	Meta/Microsoft	Open	Multiple wrappers	N/A	https://onnx.ai	
	Hyperparameter Tuning		SigOpt	Intel	Semi-open	Multiple wrappers	Various CPUs/ Nvidia GPUs	https://sigopt.com
			Determin AI	Determin AI	Open	Python/CLI	Various CPUs/GPUs	https://www.determined.ai
			Optuna	Preferred Networks	Open	Python	Various CPUs	https://optuna.org
			Apache TVM	Apache Foundation	Open	Multiple	Various CPUs/GPUs	https://tvm.apache.org
	Experiment Management		Comet	Comet ML Inc.	Closed	Python	Various CPUs/GPUs	https://www.comet.com/site/
			Neptune	Neptune Labs	Open	Python	Various CPUs	https://neptune.ai
			MLflow	MLflow Project	Open	Python	Various CPUs	https://mlflow.org
			TensorBoard	Google	Open	Python	Various CPUs/GPUs/TPUs	https://www.tensorflow.org/tensorboard
			sacred	IDSIA	Open	Python	Various CPUs	https://github.com/IDSIA/sacred
			Weights & Biases	Weights & Biases	Closed	Python/ Web Interface	Various CPUs	https://wandb.ai/
Resource Management	Slurm	Slurm	Open	C/C++	N/A	https://slurm.schedmd.com/documentation.html		

取組概要

アプリケーション調査研究グループでは、**計算科学・社会科学・データ科学のアプリ分野**から構成するサブグループにおいて、次世代計算基盤により**最先端の研究成果を実現するために必要な計算資源に関する調査**を行う。本調査には、一般的な性能指標に加え、生産性といったアプリケーションプログラムの性質を規定する様々な指標の調査を含む。次に、本調査に基づき、次世代計算基盤におけるアプリ性能を見積もるための代表的なアプリの特性を反映した（マイクロ）**ベンチマークの構築**を行う。



検討状況

- **純アプリグループ（生命科学分野、新物質・エネルギー分野、気象・気候分野、地震・津波防災分野、ものづくりアプリ分野、基礎科学分野、社会科学分野、デジタルツイン・Society5.0分野）**
 - 現行のスーパーコンピュータにおける**アプリ解析に関する調査**を完了
 - **2030年前後において各アプリ分野で期待される成果とそれに必要な計算機資源の調査**を完了
 - 各アプリ分野のプログラムの特性を反映した**ベンチマークプログラム群を整備** → 候補アーキでいくつかを評価
- **CSグループ（科学技術計算/機械学習アルゴリズム、ベンチマーク構築、性能モデリング）**
 - 機械学習のベンチマークとしてMLPerfを活用することとし、モデルの選定を完了
 - 問題サイズやコア当たりのメモリ量を可変とするベンチマークを検討中
- **他グループとの連携**
 - アーキテクチャGやシステムソフトウェアGからの**アンケートへ回答**

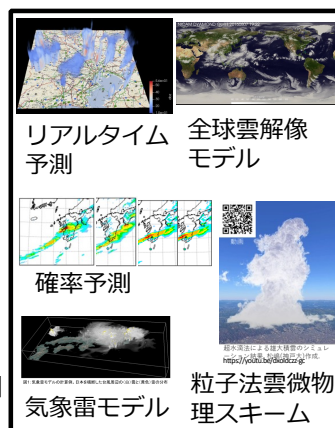
アーキテクチャ評価のためのベンチマークアプリケーションリスト

- ベンチマークアプリケーションのリスト（各分野の代表的なもの）
 - ポスト富岳の開発にあたってはより多くのベンチマークを用いた評価とシステム設計へのフィードバックを行うべき

Area	Application	Type	Language	GPU	Note
Life Science	GENESIS	MD (particle)	Fortran	✓	strong-scalability oriented
New Material & Energy	SALMON	DFT, Stencil, FFT	Fortran	✓	high-precision GEMM required
Weather and Climate	SCALE-LETKF	CFD (structured mesh)	Fortran	✓	
Earthquake & Tsunami Disaster Prevention	EbE-method	FEM (unstructured mesh)	C++	✓	
Manufacturing	FrontFlow/blue	FEM (unstructured mesh)	Fortran	✓	
Fundamental Science	LQCD-HMC-DWF	Stencil, SpMV	C++	✓	
AI	Hugging Face GPT-2 XL	Transformer	PyTorch	✓	1.5B parameters
AI	Megatron-LM DeepSpeed	Transformer	PyTorch	✓	70B parameters

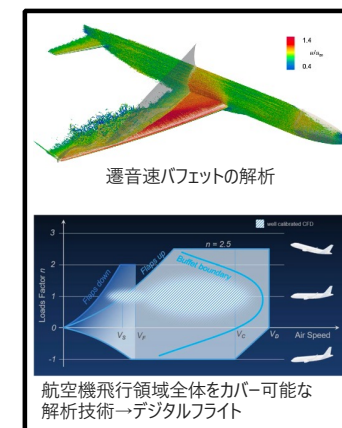
気象・気候

- 集中豪雨・台風予測精度の大幅向上に向けたリアルできめ細やかな気象予測システムの開発。竜巻・雷頻度予測、シナリオを網羅する確率予測、3次元高度観測による高頻度リアルタイム予測の実現。
- 気候変動の理解と予測を質的に高める精緻な地球システムモデルの実現。台風・極端現象・都市気候予測による適応策での利用、温室効果ガス吸排出量推定によるパリ協定への貢献。AI超解像の活用。
- 微小から惑星スケールに至る現象の包括的理解。
- 精緻な物理的表現が可能な次世代型の力学、雲微物理、放射スキームの開発と実証。AI代理モデル等の加速による富岳比5-10倍以上の大規模計算の実現。



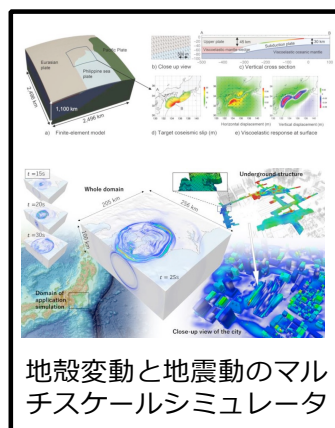
ものづくり

- HPCを活用した現象解明とそれに基づいた最適設計が様々な産業分野の設計・製造プロセスで実用化されることで、製品の更なる高性能化や、多種・多様なニーズに対応した製品設計の迅速化等の実現が期待される（例）航空機の高速飛行条件に関する各種性能の予測評価、および、性能や安全性を統合的に評価する多目的・多設計変数の機体設計
- 上記の実現に向けた取り組み
- ウィークスケリング性能向上のためのプロセス間通信の効率化
 - AIを活用したサロゲートモデルの構築等による設計空間探索の効率化



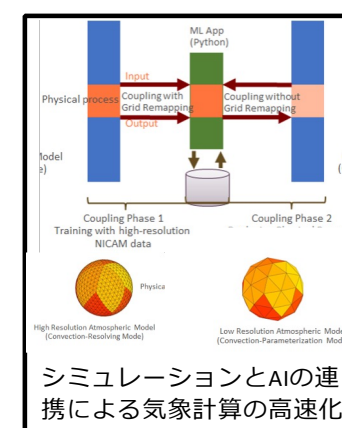
地震・津波防災

- 広域のマクロ地殻変動シミュレータと特定領域のミクロな地殻変動・地震動シミュレータを組み合わせたマルチスケールシミュレータを開発することで、地殻変動・地震動の観測データに整合するモデルを構築。これにより、一定規模の地震が発生した後、周囲での地震発生の推移を予測できるか検証できるようになると期待
- 上記の実現に向けて、富岳の5-10倍程度のシミュレーション性能の実現を目標とする。そのため、次世代計算基盤のシステム特性に合わせた物理シミュレーション、及び、AI等のデータ駆動型のアルゴリズム・実装を開発



デジタルツイン・Society5.0

- 高解像度シミュレーションとAIの連携により、データ駆動型のパラメタリゼーションモデルを構築し、これを活用した超高解像度メッシュを用いずとも高精度な全球気候モデルの計算を実現。富岳の5倍程度の計算性能とAIを併用することで、現行の高空間解像度計算よりも100倍の高速化と数百年規模の長期気候計算の実現。
- 富岳の10倍程度の計算が可能となることで、東京都心部の10km四方に対して、IoT機器のデータをリアルタイムに同化をともなった25cm解像度のリアルタイム風況デジタルツインを実現。これは現在のシミュレーションで達成できる典型的な解像度の約8倍の高精細計算。



次世代計算機で期待されるブレークスルー

科学の在り方を根本的に変える

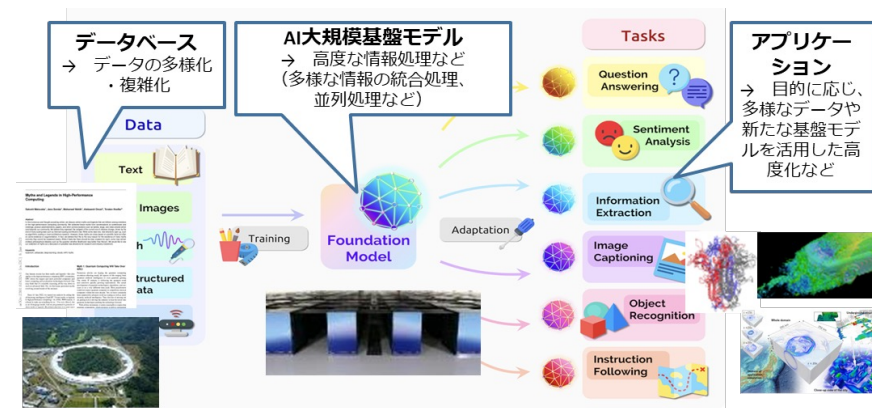
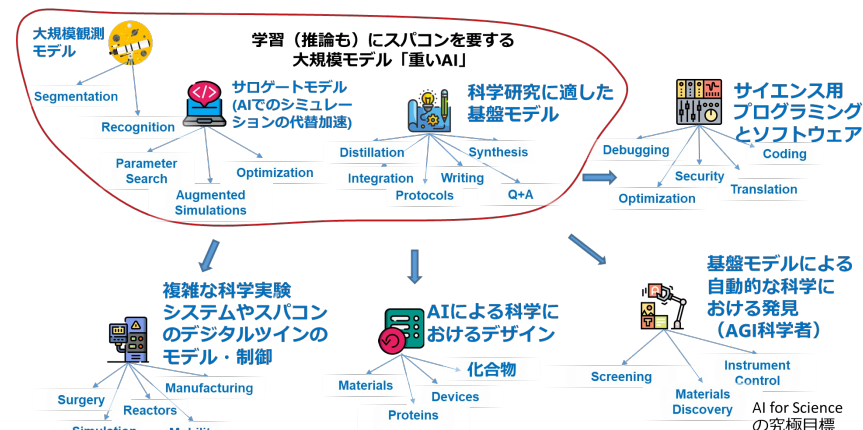
- 高度化した汎用人工知能（Artificial General Intelligence）により、本質的な科学的な発見とイノベーションの達成が行われるなど、**科学分野や産業分野をはじめとするあらゆる分野で大きな変革が進む**可能性がある。
- AIの高い計算能力やデータ解析能力を科学のさまざまな分野に応用することで、いままでの手法では得ることができなかった**新たな発見や問題解決の手法が生まれる**可能性がある。

ブレークスルーに必要な計算環境、解決すべき課題

AI for Science技術の保有および、基盤の整備

- 技術や基盤を保有**しないと、利用に制限がかかる可能性があり、将来AI分野にとどまらず、科学技術・イノベーション分野全体で、我が国が世界のトップ集団から遅れを取り、挽回が不可能となる。
（EUは域内のAIの利用について、一部を制限または禁止する規制案を発表）

スパコンによる新たな AI for Science の機能



**理研チームの調査研究結果：
あるべき次世代計算基盤の開発に向けて**

● ハードウェア

- CPU部としては、これまで富岳で蓄積されたアプリケーションやシステムソフトウェア資産が活用できるよう、可能な限り**富岳とバイナリレベルで互換性を持つべき**
- 加速部としては、プログラミングや性能最適化の観点からユーザに使いやすいものとなるよう**オープンなソフトウェアエコシステムが整備されており**、また富岳NEXTの稼働前からコード移植を効率的に実施できるよう、加速部アーキテクチャが現状で広く利用可能であることが必要。そこで、**大規模なスーパーコンピュータにおいて既に活用実績を持つGPUに基づくアーキテクチャを演算加速部として導入すべき**
- 計算ノードには複数のCPU部ソケットと加速部ソケットが搭載され、**CPU部と加速部同士はキャッシュコヒーレンスを有する高速リンクで接続されるべき**。また、ノード内の加速部同士は、複数の加速部を利用した並列処理が高速に実行できるよう**高帯域なスケールアップネットワークで接続されるべき**
- 各計算ノード間は、大規模なアプリケーションを効率的に並列処理できるよう、スケールアウトネットワークにより接続され、また一つの計算ノードが複数ジョブで共用される際にお互いに影響を及ぼさないよう、CPUソケットのみならず**GPUソケットもスケールアウトネットワーク向けのネットワークインターフェースに直接接続されるべき**
- 計算ノードまたはジョブローカルで利用できる高速ストレージまたはキャッシュと、全計算ノードで共有するストレージシステムを持ち、**各階層において要求される帯域、IOPS、容量を実現するための構成にすべき**。また、実運用で**長時間稼働させたとしても安定した性能が維持される**ことが必要

● ソフトウェア

- コンパイラやライブラリを含むシステムソフトウェアについては、開発終了後の継続的な機能拡張やコードのメンテナンスの観点から、**できる限りオープンソースのソフトウェアを導入すべき**
- 開発したソフトウェアに関しては、ユーザや開発関係者による改変や機能拡張を通じて「富岳」の次世代となる新たなフラッグシップシステムやコミュニティの発展に資するべく、**原則としてオープンソースソフトウェアとして公開すべき**
- 商用のソフトウェアを導入する場合においても、過去の開発実績や今後のロードマップを有するなどを参考に将来を含めて定期的にアップデートされ得るものを採用すべき

● 設置環境

- 運用にかかるコストの適正化、およびカーボンニュートラルの観点も踏まえ、**計算ノード部（ストレージ含む）の最大消費電力は40MW以下であるべき**
- 計算ノードの**殆ど全ての構成要素は水冷とし、ASHRAE W4（W45）およびASHRAE A4の水冷・空冷条件が望ましい**

● アプリケーションとシステムの協調設計

- ベンチマーキングによるシステム開発（ハードウェアやシステムソフトウェア）へのフィードバックについて
 - 想定されるポスト富岳アーキテクチャと構成に近いシステムを用意しベンチマーキングを行っていくべき
 - AI-for-Scienceを含む幅広いアプリケーションとの協調設計に向け、ベンチマークのためのフレームワーク（ツールや環境）を構築し多くのアプリ開発者がアクセスできるようにすべき
 - 以下の例のような幅広いアプリケーションをベンチマーク対象として考えるべき
 - FSベンチマークアプリやFSで開発されたベンチマーク候補アプリ、HPCで利用頻度が高いOSS群、AI関連アプリ、商用アプリ、その他ポスト富岳で重要と考えられるアプリ
- 候補システムを対象としたアプリケーションの加速部向け移植やチューニングについて
 - システム設計者とシステムアプリ開発者と協調してポスト富岳向けのチューニング方法の検討、また可能なものについては移植・チューニングを開発初期段階から進めていくことが望ましい
 - チューニング方法や必要なハードウェア特性など、開発者に有用な情報やシステム設計の進捗は、機微情報を除きできるだけオープンに発信していくべき
- 上記を実現するための体制やロードマップを開発初期から明らかにしていくことが望ましい

- 現状で妥当と考えられる満たすべきシステム全体の要求仕様

項目	CPU	加速部
合計ノード数	3400ノード以上	
FP64ベクトル性能	48PFLOPS以上	3.0EFLOPS以上
FP16/BF16行列演算性能	1.5EFLOPS以上	150EFLOPS以上
FP8行列演算性能	3.0ELOPS以上	300EFLOP以上
FP8行列演算性能（sparsity考慮）	—	600EFLOPS以上
メインメモリサイズ	10PiB以上	10PiB以上
メインメモリバンド幅	7PB/s以上	800PB/s以上
合計消費電力	40MW以下（計算ノードおよびストレージ）	

今後の次世代計算基盤開発・整備の在り方への展望

● 今回の次世代計算基盤に係る調査研究の良かった点

- 様々なベンダーと協業して調査研究を実施することで最先端技術に関し充実した調査が実施できた
- 機微情報を適切に管理しながら検討結果等が報告できる体制を構築できた
- 国際連携を通じて、他国からも参考にされるような調査研究体制を構築することができた
- AI for Scienceに関してロードマップ執筆など、コミュニティの形成のきっかけを作ることができた

● 改善すべき点

- 要素技術の開発をしつつ次世代計算基盤へのフィジビリティを調査するには時間が限られている点
- 多くのベンチマークを準備したが、ベンダーに評価してもらったベンチマークを限定せざるを得なかった点
- 特にシステムソフトウェアやアプリケーションの調査・検討についてアウトリーチの機会を拡充すべき

● 今後の次世代計算基盤開発・整備の在り方への意見

- 分野を越えた次世代基盤に関する知識や知見の集約・共有の場の構築
 - 性能最適化方法、AI利用方、システム性能の傾向、ベンチマーキング、・・・
- 10年に一度立ちあげるのではなく構築した調査研究の連携体制が継続できる仕組みづくり