

2025年度「富岳NEXT」プロジェクト ワークショップ

富岳NEXT システム検討状況

佐野 健太郎

理化学研究所 計算科学研究センター (R-CCS)

プロセッサ研究チーム プリンシパル

次世代AIデバイス開発研究ユニット リーダ

次世代高性能計算基盤システム開発ユニット リーダ (富岳NEXT)

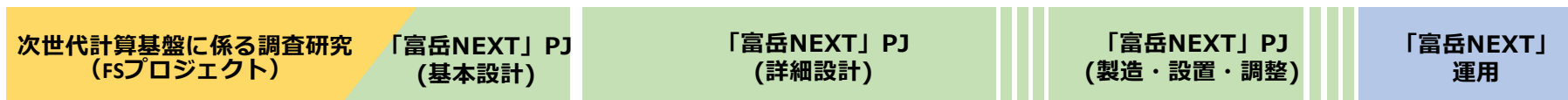
「富岳NEXT」システムの開発・整備のスケジュール（想定含む）

FY2022 R4年度	FY2023 R5年度	FY2024 R6年度	FY2025 R7年度	FY2026 R8年度	FY2027 R9年度	FY2028 R10年度	FY2029 R11年度	FY2030 R12年度	FY2031 R13年度
----------------	----------------	----------------	----------------	----------------	----------------	-----------------	-----------------	-----------------	-----------------

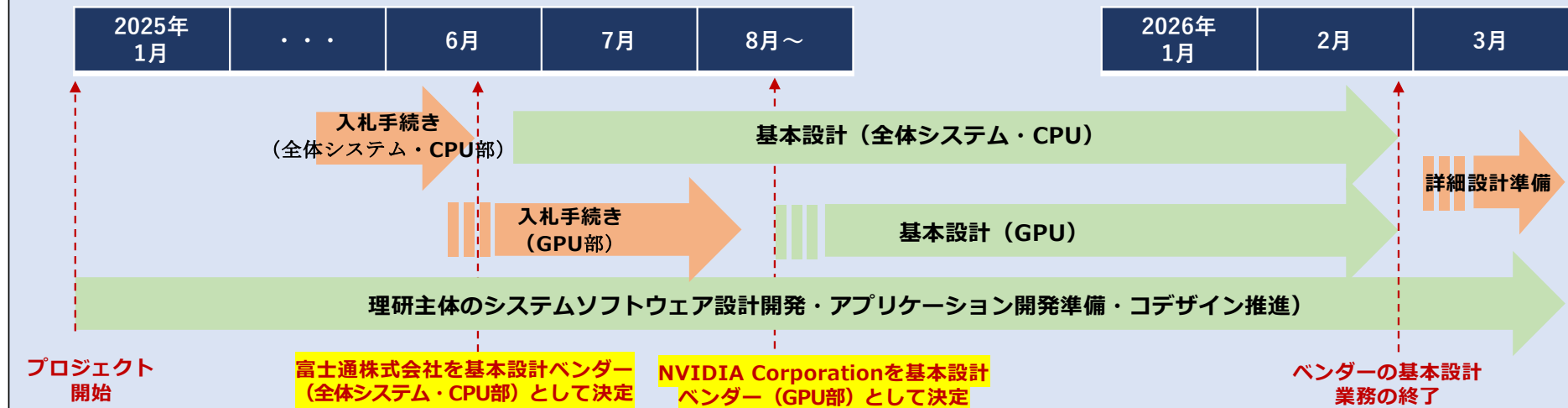
「富岳」



「富岳NEXT」



基本設計に関するスケジュール

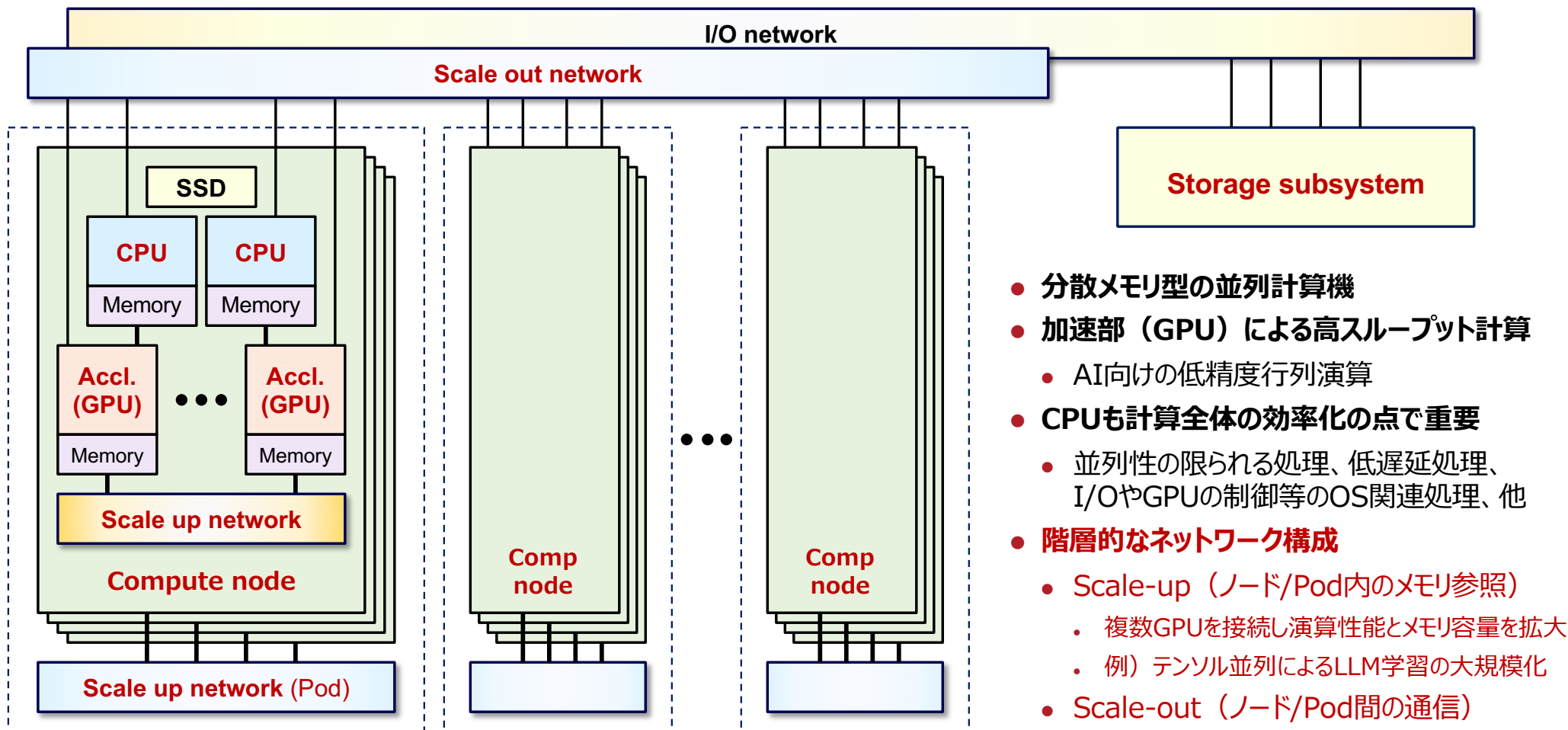


アーキテクチャ検討体制

次世代計算基盤開発部門 システム開発ユニット (リーダー：佐野 健太郎)



近年のスーパーコンピュータシステムの代表的構成



- 分散メモリ型の並列計算機
- 加速部 (GPU) による高スループット計算
 - AI向けの低精度行列演算
- CPUも計算全体の効率化の点で重要
 - 並列性の限られる処理、低遅延処理、I/OやGPUの制御等のOS関連処理、他
- 階層的なネットワーク構成
 - Scale-up (ノード/Pod内のメモリ参照)
 - 複数GPUを接続し演算性能とメモリ容量を拡大
 - 例) テンソル並列によるLLM学習の大規模化
 - Scale-out (ノード/Pod間の通信)

次世代計算基盤システムの指針（ハードウェア）

<p>CPU部</p>	<p>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、可能なかぎり富岳とバイナリレベルで互換性を持つべき</p>
<p>加速部</p>	<p>プログラミングや性能最適化の観点からユーザに使い易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加速部として導入すべき</p>
<p>計算ノード</p>	<p>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</p>
<p>ネットワーク</p>	<p>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、また計算ノード内の複数ジョブを考慮して、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</p>
<p>ストレージ</p>	<p>計算ノードまたはジョブローカルで使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</p>

「富岳NEXT」に求められるシステム性能目標

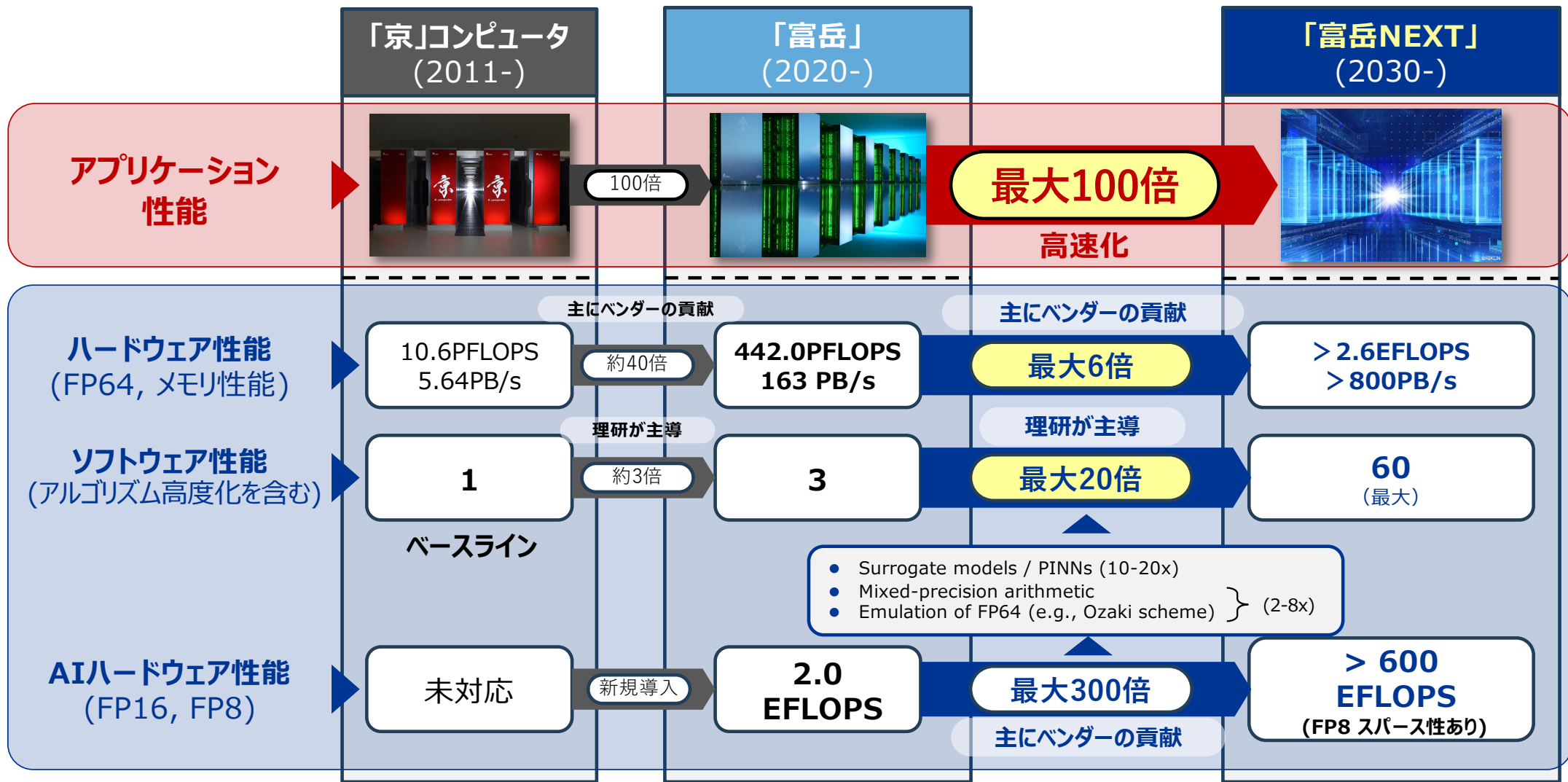
「次世代計算基盤に関する報告書 最終取りまとめ」で示された性能目標

- 既存HPCアプリケーションで現行の5～10倍以上の実効計算性能
- AI処理でゼタ（Zetta） FLOPSスケールのピーク性能を念頭に50EFLOPS以上の実効性能
- シミュレーションとAIの融合により、総合的に数十倍のアプリケーション高速化を目標

当該性能目標をベースに以下の仕様で開発ベンダーと基本設計を実施

項目	CPU	加速部	「富岳」	「富岳」比
合計ノード数	3400ノード以上		158,976	
理論 FP64ベクトル性能	48 PFLOPS以上	2.6 EFLOPS以上	537 PFLOPS	x4.8 以上
理論 FP16/BF16行列演算性能	1.5 EFLOPS以上	150 EFLOPS以上	2.15 EFLOPS	x70.5 以上
理論 FP8行列演算性能	3.0 EFLOPS以上	300 EFLOPS以上	—	
sparsity考慮の 同理論性能	—	600 EFLOPS以上	—	
メインメモリサイズ	10 PiB以上	10 PiB以上	4.85 PiB	x4.1 以上
メインメモリバンド幅	7 PB/s以上	800 PB/s以上	163 PB/s	x4.9 以上
合計消費電力	40 MW以下（計算ノードおよびストレージ）		約30 MW	

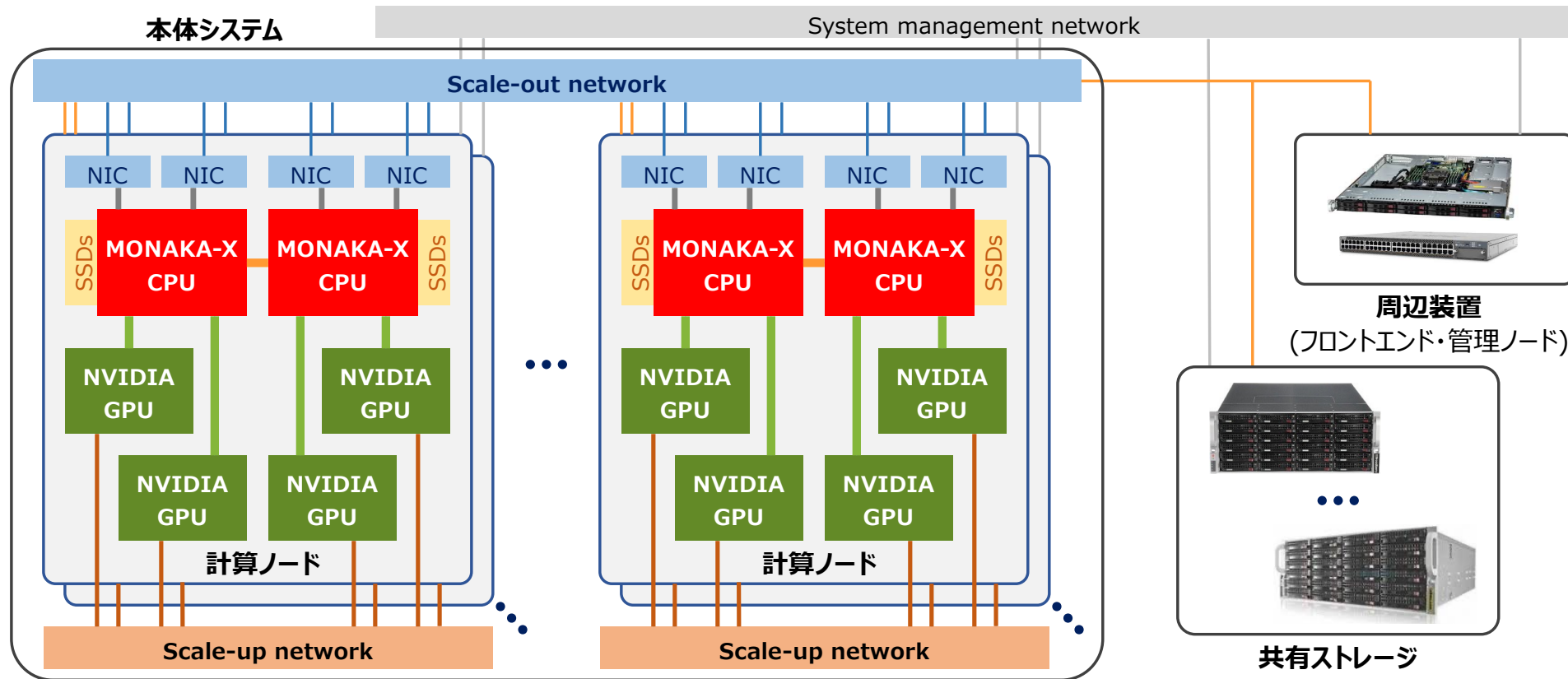
アプリケーション性能 最大100倍への道



基本設計におけるシステム検討の概要

- 「富岳NEXT」に求められるシステム性能目標や設計指針に基づき、基本設計を実施
 - **理研** 要求定義と主導、設計空間設定とコデザイン、性能モデル調査
 - **富士通** CPU、システムインテグレーション（ネットワークを含む）
 - **NVIDIA** GPU、Scale-upネットワーク
- システムの主たる基本設計項目
 - CPUとそのメモリ技術
 - 加速部（GPU）とそのメモリ技術
 - 計算ノード構成（CPU-GPU接続技術を含む）
 - Scale-upネットワーク
 - Scale-outネットワーク
 - 全体システム構成
- 基本設計に基づき設計空間を設定、コデザインを実施
- 性能推定ツール開発に向け、CPUやGPUの性能モデル検討を実施

基本設計による富岳NEXTの全体システム案



本体システム

- 富士通製MONAKA-X CPUとNVIDIA社製GPUから成る計算ノードの超並列計算機
- Scale-outネットワークによるノード間接続

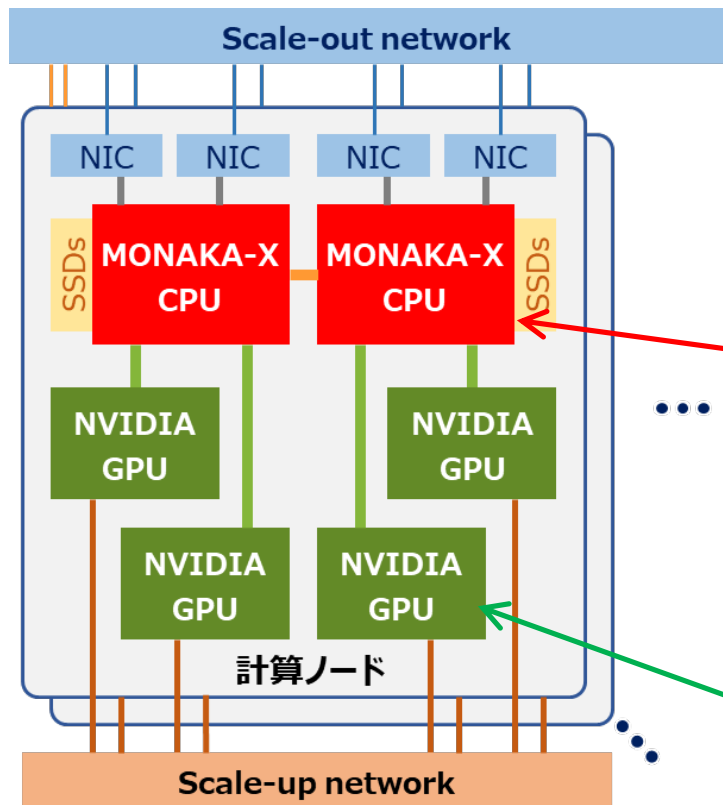
計算ノード

- 広帯域でコヒーレントなNVLink-C2CによるCPU・GPU間接続を有力候補として検討
- NVLinkのGPU scale-upネットワークを検討

CPU, GPU

- 演算性能の基本方針を検討 (例: FP64ベクタ, 低精度マトリックス)
- 容量・帯域の異なるメモリ技術オプションを検討

計算ノードの検討状況



ノード基本構成 (CPU数:GPU数) の検討

- 2 CPU : 4 GPUがノードメモリ容量とCPU間接続帯域の点で最有力

ノード設計空間に対し、コデザインを実施

富士通 NVIDIA 理研

- CPU関連を含む一部については、第一候補となる技術オプションや仕様を選定
- その他の仕様は、引き続きコデザインを実施し絞込み

MONAKA-X CPU 基本仕様の検討

富士通 理研

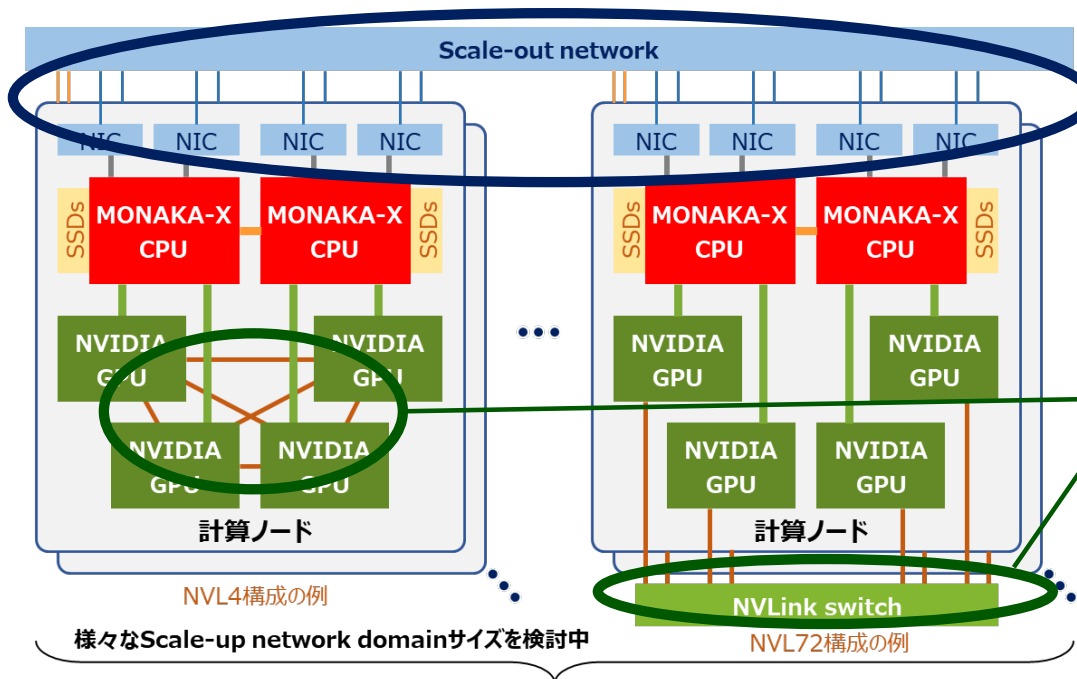
- チップレットによるARM命令セットのメニーコアCPU (高性能・低消費電力・低コストを実現)
- ARM SVE (ベクトル演算)、およびSME (低精度行列演算 : NPU) の仕様を検討
- ノードの実装方式と併せて、容量・帯域の異なるメモリ技術オプションを検討
- 有力なGPU接続方式として、広帯域・コヒーレントなNVLink-C2Cを検討

GPU 基本仕様の検討

NVIDIA 理研

- SM (ベクトル演算)、Tensorコア (低精度行列演算) の仕様を検討
- 容量・帯域の異なるメモリ技術オプションを検討
- 2.5Dメモリに加え、先端技術を採用した積層メモリの可能性を検討

Scale-out / Scale-upネットワークの検討状況



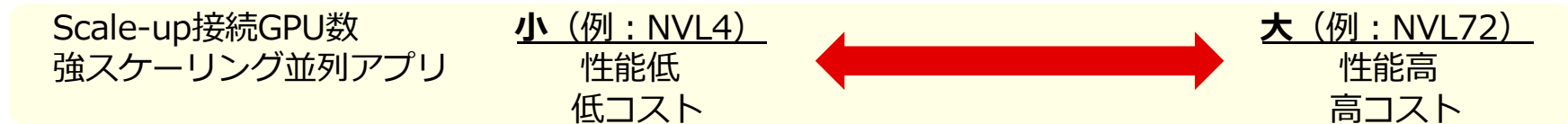
Scale-outネットワーク

- **システム全体**でのノード (or ポッド) の相互接続
- **大域的**/高遅延/狭帯域で粗粒度のメッセージ通信
- 全体の問題サイズを増やして実行時間低下よりも計算性能向上を目指す**弱スケールングに適合**
- **ツリートポロジベース**のネットワーク構成を検討

Scale-upネットワーク

- ノード/ポッドの**限定的な範囲内**でのGPU相互接続
- **局所的**/低遅延/広帯域で分散共有メモリアクセス
- 全体の問題サイズを増やさずに実行時間を低下させる**強スケールングに適合**。しかし**コスト増**
- **異なるScale-upドメインサイズを検討** (混載を含)

Scale-upドメインサイズ (接続GPU数) に対し想定されるアプリ性能・コストのトレードオフ



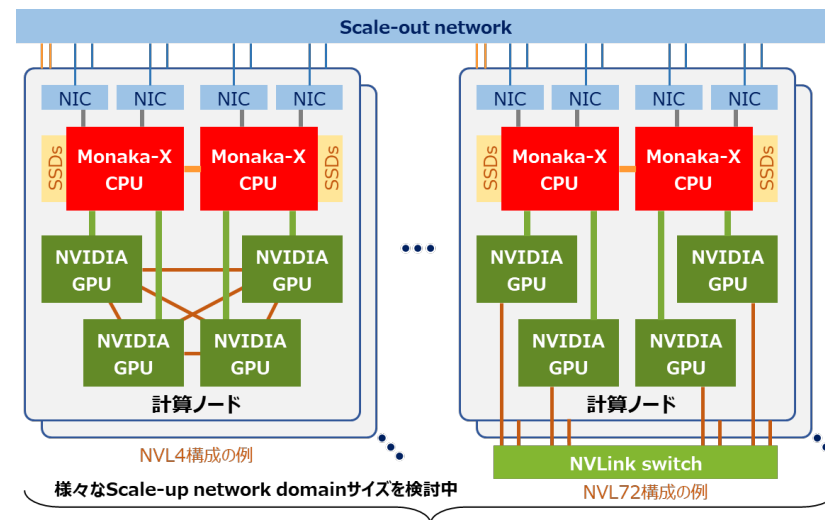
Scale-up設計において、強スケールングを要求するアプリやAIの性能便益の割合や度合いについて調査が必要

コデザイン検討状況

- **基本設計に基づき設計空間を設定して、アプリケーションエリアとの協調設計を実施**
 - 設計空間 = 技術や性能の選択肢の組合せ
- **これまでに設定された設計空間における技術項目の例**
 - CPU性能関連（演算性能、メモリ性能など）
 - GPU性能関連（演算性能、メモリ性能など）
 - CPU-GPU接続技術
 - ノード構成
 - Scale-upネットワーク関連（ドメインサイズなど）
 - Scale-outネットワーク関連（インジェクション帯域など）
- **便益とコスト・リスクを総合的に評価し、第一候補を選定していく予定**
 - 便益 アプリケーション性能など
 - コストとリスク 技術の実現可能性/スケジュールやリスク、コストなど

まとめ（システム検討状況）

- 「富岳NEXT」に求められるシステム性能目標や設計指針に基づき、基本設計を実施
 - 全体システム案
 - 計算ノード検討
 - Scale-up/Scale-outネットワーク検討
- 得られた設計空間に対し、コデザインを進める予定



項目	CPU	加速部	「富岳」	「富岳」比
合計ノード数	3400ノード以上		158,976	
理論 FP64ベクトル性能	48 PFLOPS以上	2.6 EFLOPS以上	537 PFLOPS	x4.8 以上
理論 FP16/BF16行列演算性能	1.5 EFLOPS以上	150 EFLOPS以上	2.15 EFLOPS	x70.5 以上
理論 FP8行列演算性能	3.0 EFLOPS以上	300 EFLOPS以上	—	
sparsity考慮の 同理論性能	—	600 EFLOPS以上	—	
メインメモリサイズ	10 PiB以上	10 PiB以上	4.85 PiB	x4.1 以上
メインメモリバンド幅	7 PB/s以上	800 PB/s以上	163 PB/s	x4.9 以上
合計消費電力	40 MW以下（計算ノードおよびストレージ）		約30 MW	