シンポジウム「ポスト富岳で拓くアプリケーションの未来」

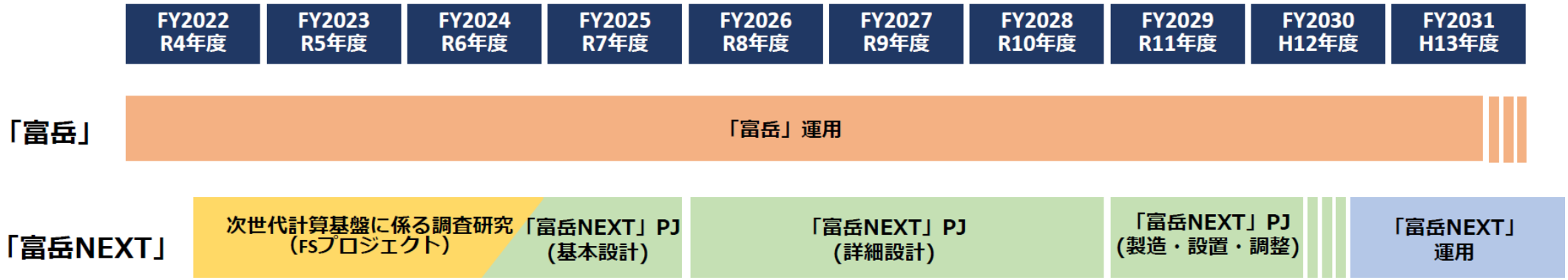
ポスト富岳として想定されるシステム概要（アーキテクチャ・ハードウェア）

佐野 健太郎

理化学研究所計算科学研究センター
プロセッサ研究チーム　チームリーダー
次世代計算基盤開発部門
次世代計算基盤システム開発ユニット ユニットリーダー（就任予定）

次世代計算基盤に係る調査研究事業

- **文部科学省 次世代計算基盤に係る調査研究事業（2022年8月～2025年3月）**
- **理化学研究所チームによるシステム調査研究**
  - 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
- **2030年の運用開始を目標**
  - 次世代計算基盤として目指すべきシステム性能と利用可能な技術、必要な開発項目



背景および求められる性能・機能に関する方針

計算基盤の今後の重要性増大

- 生成AIの進展
- 計算科学だけでなく科学技術やイノベーション全体の推進
- 産業競争力強化

計算資源の需要増大

- シミュレーションとAI技術の高度な融合、量子コンピューティングとのハイブリッド利用など、求められる機能の変遷・多様化

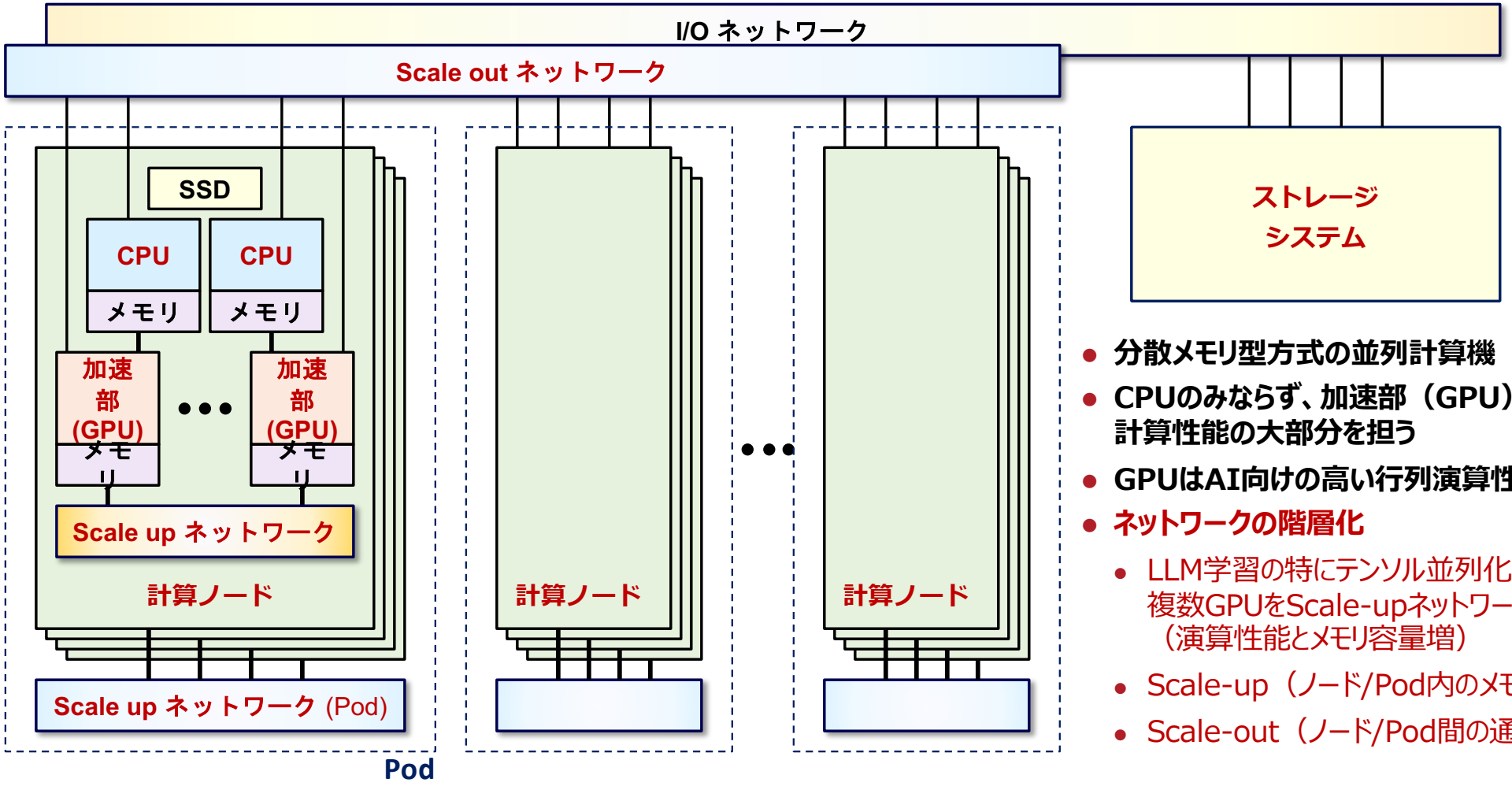
産学官の利用者に対してあらゆる分野で世界最高水準の計算資源を提供する必要性

- 研究開発力・産業力の革新をもたらす「AI for Science」に向けた取組みが加速

- **文部科学省 HPCI計画推進委員会（研究振興局長の諮問機関）による新たなフラッグシップシステムの開発に関する方針の最終とりまとめ（2024年6月）**
  - CPUに加えて、GPUなどの加速部を導入
  - 電力性能の大幅に向上させた計算環境の提供
  - 既存のシミュレーションに関しては、「富岳」の5～10倍以上の実効性能の達成
  - AIの学習・推論に必要となる性能に関しては、世界最高水準の利用環境（実効性能50EFLOPS以上）を実現
- **ポスト「富岳」後継機の開発主体として理化学研究所を選定（2024年6月）**

次世代計算基盤の調査研究を行いつつ、開発主体の立場でポスト富岳として想定されるシステムを検討

近年のスーパーコンピュータシステムの構成



- 分散メモリ型方式の並列計算機
- CPUのみならず、加速部（GPU）が計算性能の大部分を担う
- GPUはAI向けの高い行列演算性能を持つ
- ネットワークの階層化
  - LLM学習の特にテンソル並列化のために、複数GPUをScale-upネットワークで接続（演算性能とメモリ容量増）
  - Scale-up（ノード/Pod内のメモリ参照）
  - 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）

CPU部	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
加速部	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
計算ノード	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
ネットワーク	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
ストレージ	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

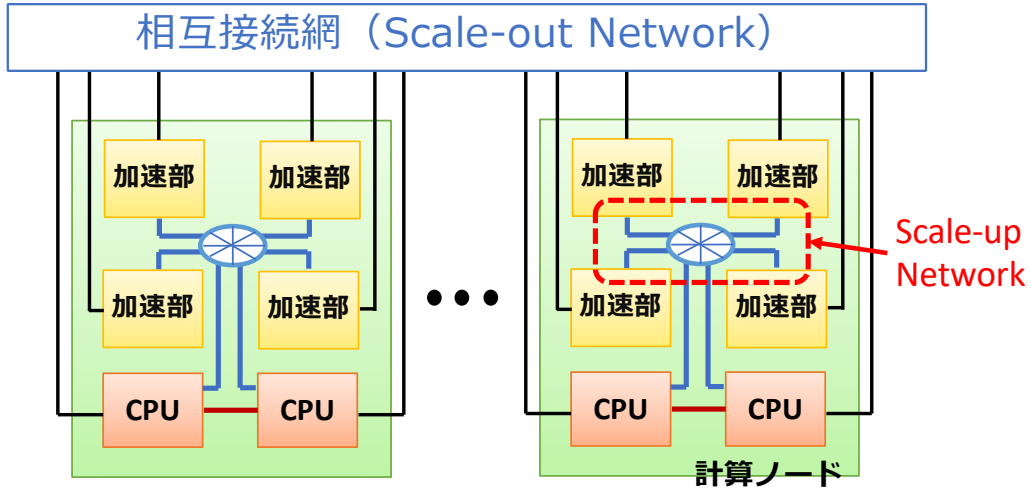
次世代計算基盤に求めるべきシステムの性能

- 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

- アプリケーションの実行性能を最優先とする
  「アプリケーションファースト」の理念
- 電力制約下でも上記目標を達成するため、「富岳」で
  培ったアプリケーションソフトなどの資産を有効活用できる
  電力効率の高いCPU部と帯域重視の演算処理加速部
  を組み合わせた、高帯域およびヘテロジニアスなノード
  アーキテクチャを基本構成としたシステム

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	3.0 EFLOPS以上	488-537 PFLOPS	x5.7 以上
理論 FP16/BF16行列演算性能	1.5 EFLOPS以上	150 EFLOPS以上	1.95-2.15 EFLOPS	x70.5 以上
理論 FP8行列演算性能	3.0 ELOPS以上	300 EFLOP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	

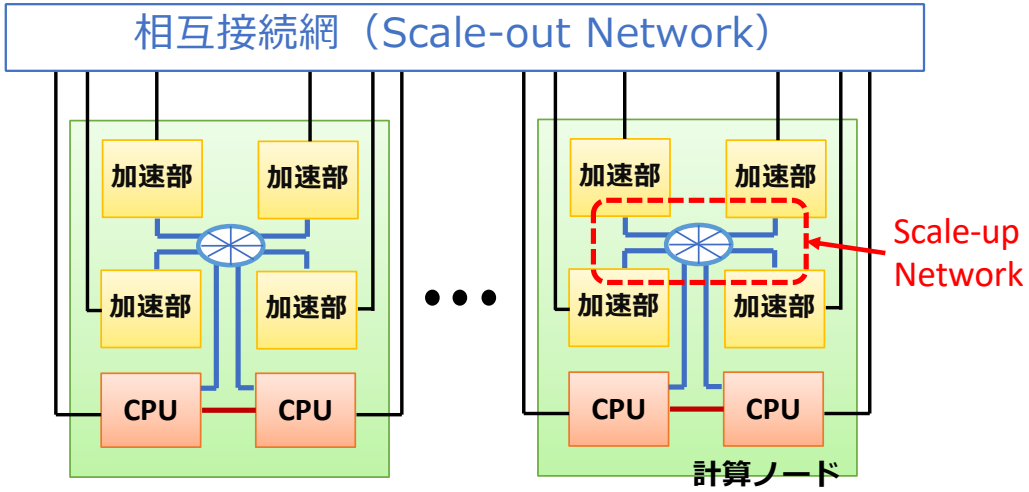
システム・アーキテクチャ概要（1/2）



- **CPUと加速部を接続した計算ノード**
  - 富岳とバイナリレベルで互換性を持つCPU
  - 大規模スパコンで活用実績のあるGPUに基づくアーキの加速部
- **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**

- **分散メモリ型方式の並列計算機**
- **CPUと加速部（GPU）から構成される計算ノード**
  - 計算ノード： 分散共有メモリ & シングルOS
  - CPU-GPU間はキャッシュコヒーレンスを有する高速リンクにより、低レイテンシかつ高バンド幅で接続
- **CPU**
  - ARM命令セットのメニーコアアーキテクチャ
  - FP64のベクトル演算（HPC向け）に加え、AIの特に推論の高速化のために低精度行列演算（FP16/BF16/FP8/INT8等）をサポート
- **加速部（GPU）**
  - 高いFP64ベクトル演算性能、および FP16/BF16/FP8/INT8等の行列演算性能
  - 高バンド幅のメモリ（導入時期における先端メモリ技術）
  - DDR系の大容量メモリも検討
  - 複数のGPUインスタンスへの分割や仮想化をサポート

システム・アーキテクチャ概要（2/2）



- **CPUと加速部を接続した計算ノード**
  - 富岳とバイナリレベルで互換性を持つCPU
  - 大規模スパコンで活用実績のあるGPUに基づくアーキの加速部
- **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**

- **高性能並列計算のためのネットワーク**
  - Scale-up ネットワーク：加速部の相互接続
  - Scale-out ネットワーク：ノードの相互接続
- **Scale-up ネットワーク（ノード内またはノード間を検討）**
  - ノード内 直接網：
    ノード内の複数のGPUを相互に接続（全結合, P2P）
  - ノード間 関節網：
    複数ノードのGPUをスイッチで相互接続し、Podを構成
- **Scale-out ネットワーク**
  - IPやRDMAをサポートするオープンな仕様のものを検討
  - CPUのみならずGPUソケットもインタフェースを持つ
  - 多段のスイッチを用いた間接網等を検討

- **性能とコストのバランスを考慮し、製造時に利用可能と予想される技術を検討**

検討するストレージシステムと仕様策定に向けた取組み

「富岳NEXT」ストレージシステムの方向性	• データサイエンス、大規模チェックポインティング等の**従来型のI/O**や、「AI for Science」等の新たなI/O要求に対応可能な**最先端のストレージシステム**に進化 • **I/Oインテンシブアプリ開発者へのヒアリング**に基づくストレージシステム、性能および容量の検討 • 「富岳」から「富岳NEXT」への円滑なデータ移行の実現（**利用の継続性・利便性の確保**）
富岳からの改善点と対応	• **安定した性能を実現するためのハードウェア／ソフトウェア設計** ⇒ ストレージノードのメモリ枯渇（接続情報、メタデータ保持等）によるI/O性能の不安定性を解決 • **持続可能なファイルシステム・ソフトウェアの開発** ⇒ ユーザのフィードバックに対して柔軟に改修ができる “OSS” を基本とした継続的な開発

● 要件

*SSF: Single Shared File

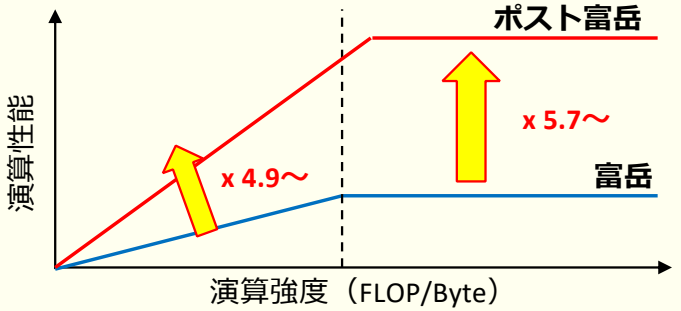
	アーキテクチャ	ファイルシステム	バンド幅 (SSF*の実効性能)	IOPS	容量
第1階層	(ニア)ノードローカルストレージ	検討中 (CHFS等)	総メモリダンプ時間: 1分以下	最大I/Oプロセス数からの メタデータ処理時間：1秒以下	総メモリサイズの 2倍以上
第2階層	共有ストレージ	Lustre, DAOS等	総メモリダンプ時間: 5分以下	第1階層の1/10のIOPS	総メモリサイズの 30倍以上

アプリケーション開発時に考慮すべきシステム特性（１）

- **演算性能・メモリ帯域共に約5倍以上、FP64のB/Fは同程度を目指す**
  - 計算バウンド、メモリバウンドのいずれのアプリでも、恩恵を得ることが期待

総メモリ帯域	富岳 163 PB/s	ポスト富岳 800 PB/s以上	（x 4.9 ～）
FP64ベクトル演算性能	富岳 537 PFLOPS	ポスト富岳 3.0 EFLOPS以上	（x 5.7 ～）
倍精度 Byte/FLOP	富岳 0.30	ポスト富岳 0.27前後	（x 0.9 前後）

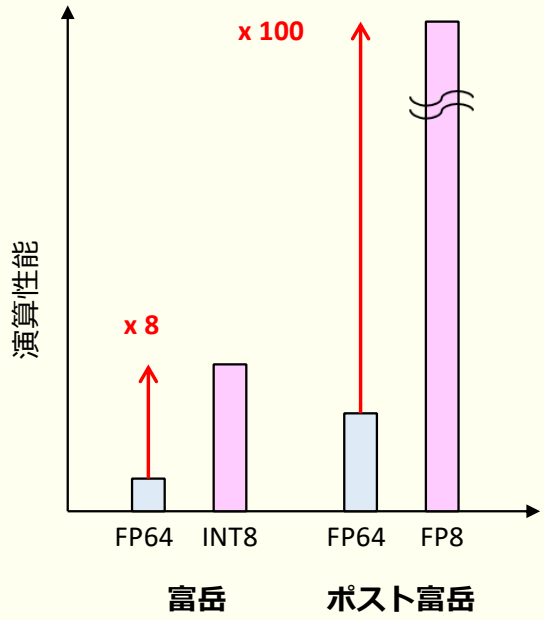
- **メモリ容量あたりの演算性能**
  - 同じ演算性能に対して使えるメモリ容量は小さくなる傾向
  - メモリ容量は貴重な資源となる。
    主メモリ容量に対し、相対的に演算のコストが下がる
    （例：主メモリに保存せずに再演算するのがお得になる可能性）
    - 富岳　537 PFLOPS/4.85 PiB ~= 98.4 FLOPS/Byte
    - ポスト富岳　3000 PFLOPS/10 PiB ~= 266.7 FLOPS/Byte



Roofline**モデル**

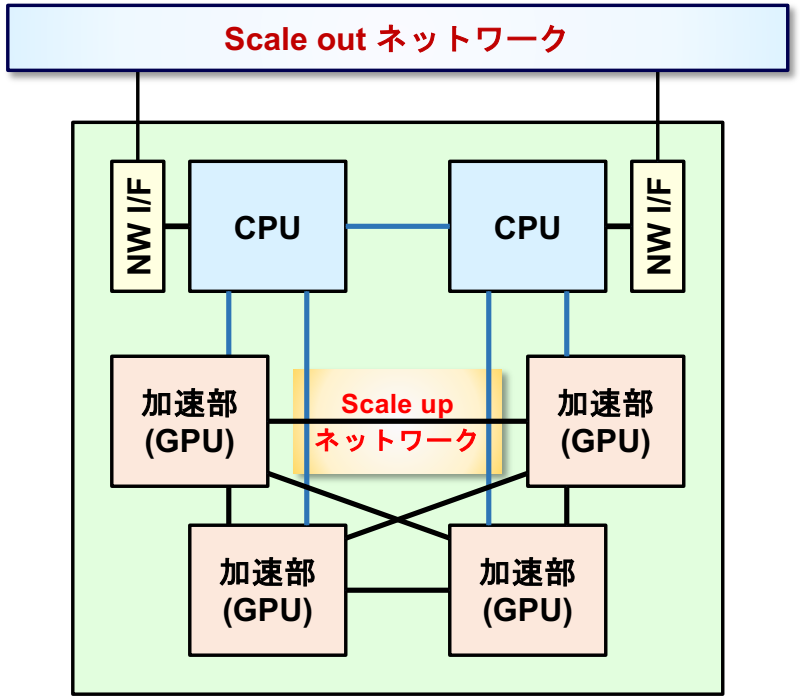
アプリケーション開発時に考慮すべきシステム特性（２）

- **AI向け低精度行列演算性能が高い（FP16/BF16, FP8など）**
  - (FP64演算性能)：(INT8またはFP8演算性能)
    - 富岳　　FP64ベクトル：INT8ベクトル　= 537 PFLOPS：4.30 EOPS　= 1：8
    - ポスト富岳　FP64ベクトル：FP8行列　　= 3.0 EFLOPS：300 EFLOPS　= 1：100
  - 倍精度ベクトル演算性能は、相対的に低下
  - 圧倒的に高い低精度行列演算性能をどう使いこなすか
    - 倍精度行列演算には、尾崎スキームによる
      低精度行列演算を用いたエミュレーションが有望



アプリケーション開発時に考慮すべきシステム特性（３）

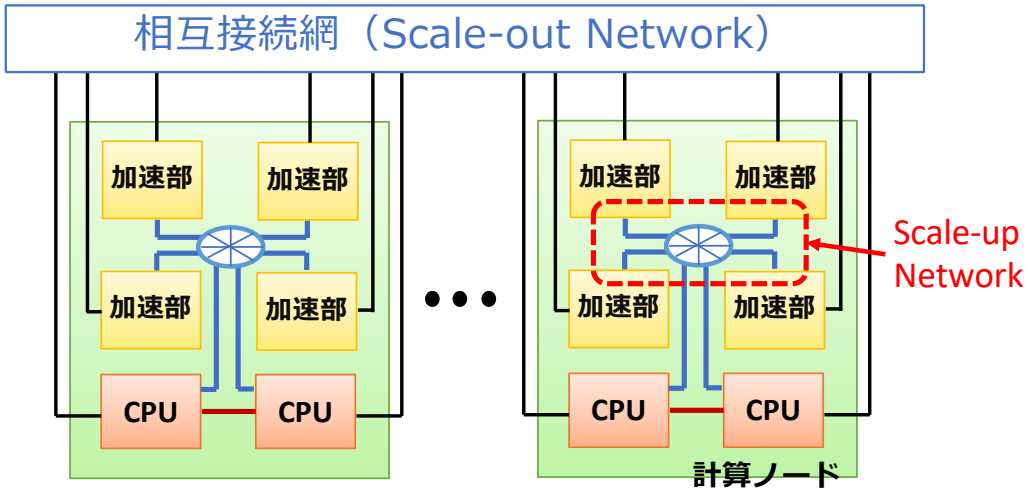
- **計算ノード構成やネットワークの詳細は今後の検討課題であるが、Scale-upネットワークにより計算ノード内の加速部を相互に接続する見込み**
  - （場合により、ノードを超えて接続しPodを構成）
- **計算ノードの「粒度」が増大**
  - 計算ノード：分散共有メモリ計算機、シングルOS
  - 複数GPU：演算性能、メモリ容量が格段に増加
  - まずは、ノード内の複数GPUの性能を使い切ることが重要
    - CPUやCPU-GPU間データ転送がボトルネックとならないこと
    - 複数GPUによる適切な並列処理を行うこと（プログラミング環境によるところが大きいかもしれない）
- **計算ノードを超えた並列計算**
  - Scale-out, scale-upネットワークそれぞれの特性を生かす並列プログラミングが必要となる（システムソフトの開発次第かもしれない）



ノード内の加速部のみをScale upネットワークで接続した計算ノード構成例

まとめ（アーキテクチャ・ハードウェア）

- **ポスト富岳として想定されるシステム**
  - CPU + 加速部（GPU）
- **アプリ開発で考慮すべき事項**
  - 演算性能あたりメモリ容量の低下
  - AI向け低精度行列演算性能の利用
  - 計算ノードの「粒度」増大
  - Scale-out, scale-upネットワークそれぞれの特性を生かす並列プログラミング



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	3.0 EFLOPS以上	488-537 PFLOPS	x5.7 以上
理論 FP16/BF16行列演算性能	1.5 EFLOPS以上	150 EFLOPS以上	1.95-2.15 EFLOPS	x70.5 以上
理論 FP8行列演算性能	3.0 ELOPS以上	300 EFLOP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	