

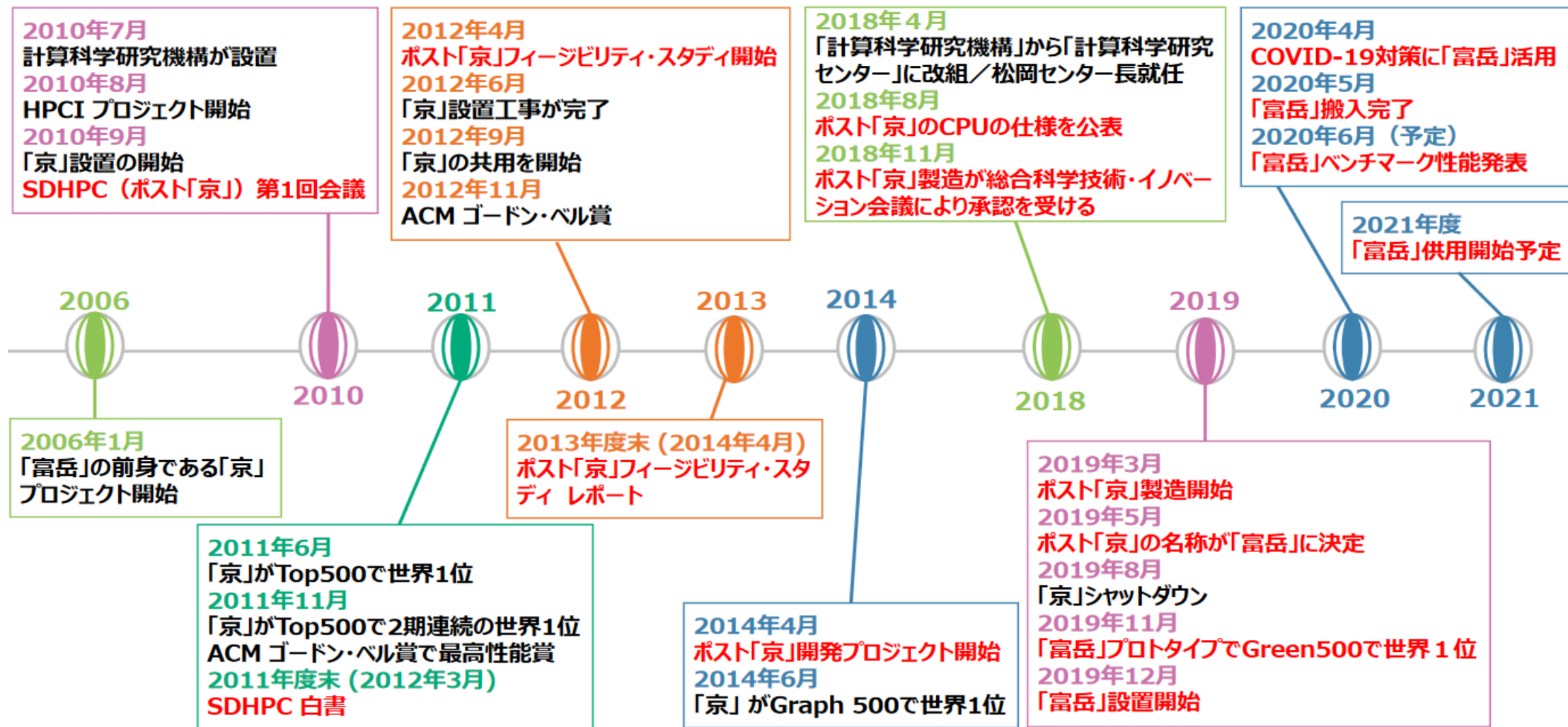
# 次世代高性能計算基盤の開発に向けた 計算機科学コミュニティでの活動紹介

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(兼務)

近藤 正章

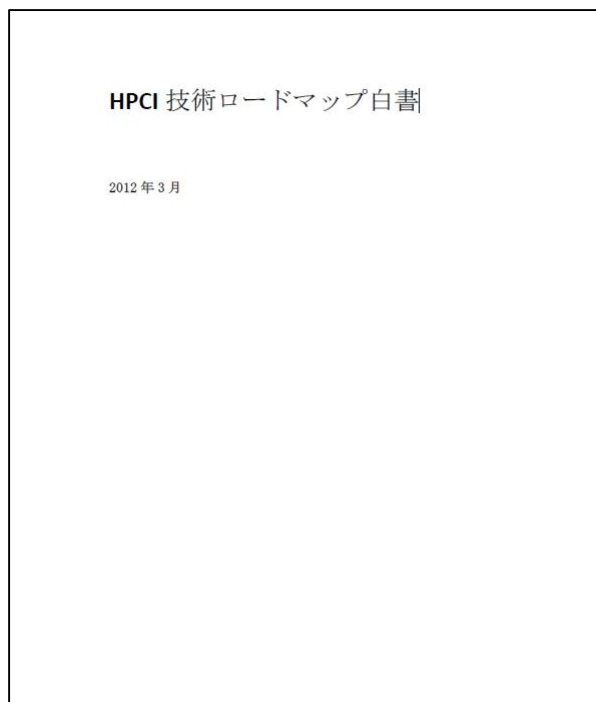
# 富岳開発への道のり

- 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



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<br>PetaFLOPS | 総メモリ帯域<br>PetaByte/s | 総メモリ容量<br>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

# 現在のCS側のコミュニティ活動の紹介

## • 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。

### – これまでの実績

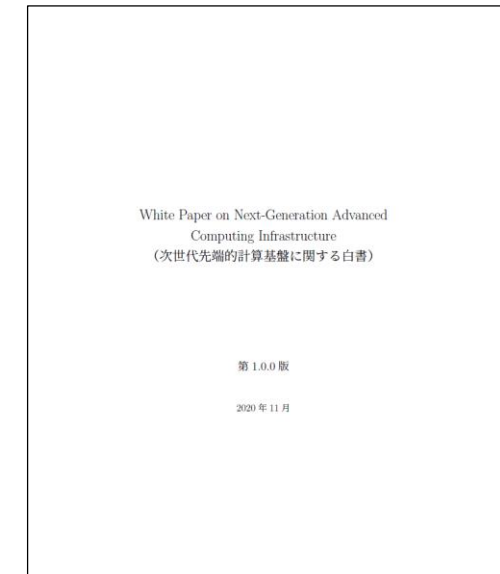
- 本活動に登録して頂いているコミュニティのメンバー数:104人
- 7回の全体ミーティングと3回のセミナーを実施
- 4つのWGにより将来のシステム像や課題を集中的に議論
  - アーキテクチャWG、システムソフトWG、アプリ/ライブラリWG、システム運用WG

### – 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>)
- 現在更新版の公開に向けて準備中



<https://sites.google.com/view/ngaci/home>



# White Paperの執筆協力者

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リーダー: 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リーダー: 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), ゲローフィ(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リーダー: 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), 棕木(理研), 村上(都立大)
- **システム運用WG**
  - WGリーダー: 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)

# White Paperの章構成

## 1. はじめに

## 2. スーパーコンピュータの技術動向

### 2.1 ハードウェア技術の動向

#### 2.1.1 デバイス

#### 2.1.2 プロセッサ

#### 2.1.3 メモリ技術

#### 2.1.4 データ転送技術

#### 2.1.5 ASIC/FPGA

#### 2.1.6 その他

### 2.2 システムアーキテクチャの技術動向

#### 2.2.1 ノードアーキテクチャ

#### 2.2.2 インターコネクト

#### 2.2.3 ストレージ

### 2.3 システムソフトウェアの技術動向

#### 2.3.1 基盤ソフトウェア

#### 2.3.2 大規模並列/高性能計算

#### 2.3.3 プログラミング環境

#### 2.3.4 性能解析ツール

#### 2.3.5 利用高度化ツール

#### 2.3.8 資源管理

#### 2.3.9 外部資源連携

### 2.4 数値計算ライブラリ/ミドルウェア/ アルゴリズムの技術動向

#### 2.4.1 数値計算ライブラリ

#### 2.4.2 数値計算ミドルウェア

#### 2.4.3 数値計算・アプリケーションを支える重要技術

### 2.5 運用に関する技術動向

#### 2.5.1 スパコン利用の枠組み

#### 2.5.2 従来のスパコン利用方式

#### 2.5.3 クラウドとHPC

#### 2.5.5 新しい利用形態

#### 2.5.6 設備と運用技術

## 3. アプリケーションの要求性能分析

### 3.1 アプリケーションの次世代システムに対する要求性能

### 3.2 要求性能に対するアプリケーション分析

#### 3.2.1 汎用システム型要求アプリケーション

#### 3.2.2 メモリ性能要求アプリケーション

#### 3.2.3 演算性能要求

#### 3.2.4 ネットワーク性能要求

#### 3.2.5 ポスト処理性能要求

## 4. 次世代(2028年頃)システムの検討

### 4.1 汎用システム型

#### 4.1.1 メニーコアCPU型

#### 4.1.2 メニーコアCPU & GPU混載型

#### 4.1.3 その他(ベクトルプロセッサ)

### 4.2 専用システム混載型および新たな可能性

#### 4.2.1 CPU拡張型

#### 4.2.2 アクセラレータ主体型 / ヘテロジニアス型

#### 4.2.3 Processing-In-memory主体型

## 5. 次世代型運用への要求

### 5.1 新しい利用形態とシナリオ

### 5.2 設備・管理

### 5.3 ユーザ利用・課金モデル

## 6. 技術課題と研究開発ロードマップ

### 6.1 デバイス・アーキテクチャ

#### 6.1.1 汎用システム型

#### 6.1.2 専用システム混載型

#### 6.1.3 PIM混載型

### 6.2 システムソフトウェア

#### 6.2.1 基盤ソフトウェア

#### 6.2.2 大規模並列/高性能計算

#### 6.2.3 プログラミング環境

#### 6.2.4 データフレームワーク

#### 6.2.5 性能解析ツール

#### 6.2.6 利用高度化ツール

#### 6.2.7 資源管理

#### 6.2.8 外部資源連携

### 6.3 数値計算ライブラリ・アルゴリズム

#### 6.3.1 数値計算ライブラリ

#### 6.3.2 数値計算ミドルウェア

#### 6.3.3 数値計算・アプリケーションを支える 重要技術

## 7. おわりに



# 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-IO DC 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                      | $3.3 \times 10^6$ | $3.8 \times 10^6$ | $4.4 \times 10^6$ | $4.4 \times 10^6$ | $5.1 \times 10^6$ | $5.8 \times 10^6$ | $5.4 \times 10^6$ | $6.3 \times 10^6$ | $7.3 \times 10^6$ |
| PFLOPS                    | 815               | 950               | 1086              | 1086              | 1267              | 1448              | 1358              | 1584              | 1810              |
| DDR 総<br>BW (PB/s)        | 102               | 120               | 137               | 137               | 160               | 182               | 171               | 200               | 228               |
| HBM 総<br>BW (PB/s)        | 307               | 358               | 410               | 410               | 478               | 547               | 512               | 598               | 683               |
| DDR 総容量<br>(PB)           | 17                | 20                | 23                | 23                | 27                | 31                | 29                | 34                | 39                |
| HBM 総容量<br>(PB)           | 4                 | 5                 | 5                 | 5                 | 6                 | 7                 | 7                 | 8                 | 9                 |
| インジェク<br>ション BW<br>(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<br>(TB/s)        | 34                | 34                | 34                | 34                | 34                | 34                | 34                | 34                | 34                |
| 総ストレ<br>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

158,976  
 $8.3 \times 10^6$   
537  
—  
163  
—  
4.85  
0.33

参考) 富岳の諸元

# 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              | $5.3\times10^8$ | $6.2\times10^8$ | $7.1\times10^8$ | $7.1\times10^8$ | $8.3\times10^8$ | $9.4\times10^8$ | $8.8\times10^8$ | $1.0\times10^9$ | $1.2\times10^9$ |
| PFLOPS            | 1279            | 1492            | 1706            | 1706            | 1940            | 2474            | 2132            | 2487            | 2843            |
| HBM 総BW<br>(PB/s) | 91              | 107             | 122             | 122             | 143             | 163             | 153             | 178             | 204             |
| HBM 総容量<br>(PB)   | 1               | 1               | 2               | 2               | 2               | 2               | 2               | 3               | 3               |

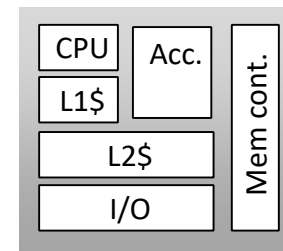
積極的な予測の場合  
(CPUとの性能比の  
トレンドから外挿)

|                   | 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              | $3.4\times10^9$ | $3.9\times10^9$ | $4.5\times10^9$ | $4.5\times10^9$ | $5.2\times10^9$ | $6.0\times10^9$ | $5.6\times10^9$ | $6.5\times10^9$ | $7.5\times10^9$ |
| PFLOPS            | 8083            | 9431            | 10778           | 10778           | 12574           | 14371           | 13472           | 15718           | 17963           |
| HBM 総BW<br>(PB/s) | 334             | 390             | 445             | 445             | 520             | 594             | 557             | 650             | 743             |
| HBM 総容量<br>(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

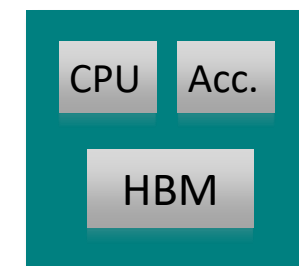
- チップ内拡張型 (SoC)

- 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



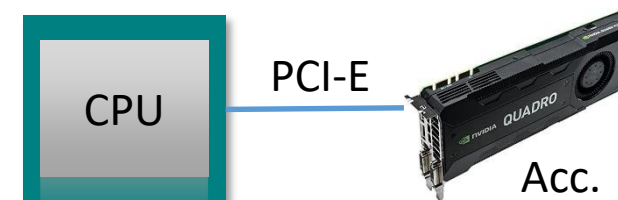
- チップ内拡張型 (マルチチップパッケージ)

- チップレットとして実装されたアクセラレータをインターポーザなどでCPUと結合
- 主記憶を共有



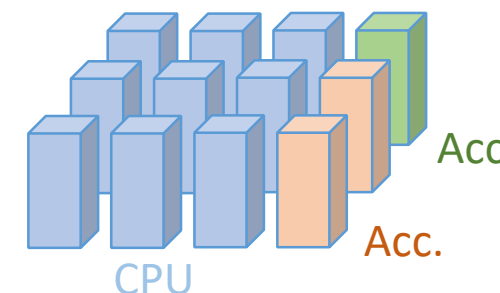
- ノード内拡張型

- 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



# アクセラレータの構成法式の種類

- 専用アクセラレータ

- 特定の計算問題または計算ドメインのみを高速に処理可能な構成方式
- 最も高性能・高電力効率であるが、柔軟性やプログラマビリティが低い
- 例) ディープニューラルネットワークアクセラレータ、FFTアクセラレータ

- 準専用アクセラレータ

- 複数の計算モデル・計算問題を高速に処理可能＋ある程度のプログラマビリティを持つ構成
- 専用アクセラレータに比べ若干性能や電力効率で劣る
- 例) データフロープロセッサ、コンボリユーションプロセッサ

- 汎用アクセラレータ

- 特定のアクセラレータを再構成可能デバイス上に構成する方式
- 例) FPGA, Coarse-Grain Reconfigurable Arrays (CGRA)

- 特定計算ドメインの候補

- 疎行列計算、N体問題、FFT、ソート、グラフ処理、脳型計算、量子計算シミュレーション、エージェントシミュレーション

# 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)

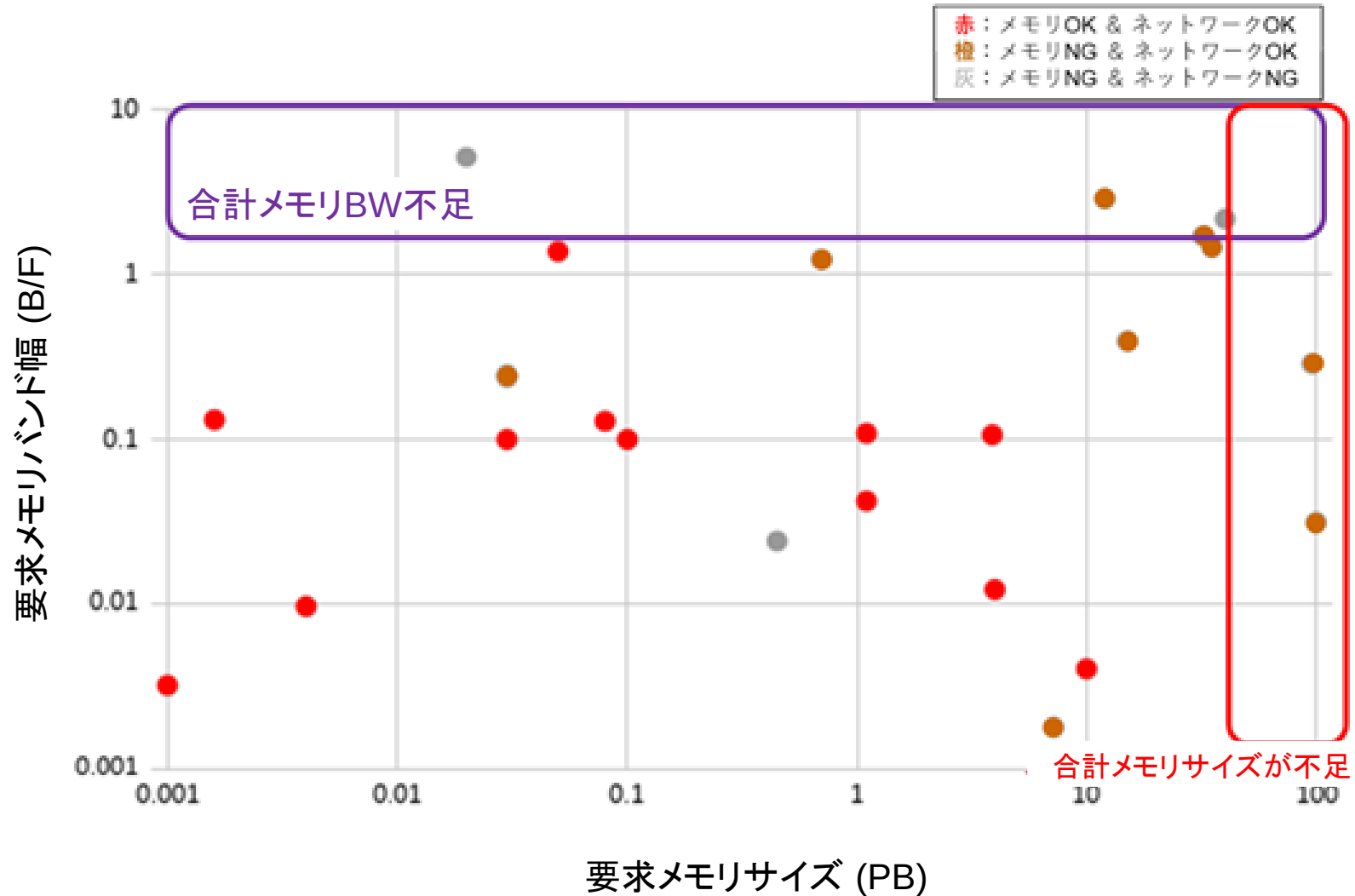
|                          | Manycore (50MW, CPU80%) | GPU (50MW, CPU80%) |
|--------------------------|-------------------------|--------------------|
| # of CPU Sockets or GPUs | 103,600                 | 112,580            |
| # of total cores         | 7,252,000               | $1.2 \times 10^9$  |
| 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

| 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) |
|-----------|---------------------------|--------------------|---------------------------|-----------------------|------------------|-----------------------------|-------------------------|-------------------|
| 素粒子・原子核   | (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               |
| 物質科学      | HPhi                      | 1,000,000          | 0.03                      | 0.001                 | 100              | 1.025                       | 0.009                   | 340               |
| エネルギー・資源  | 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               |
| 脳科学・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               |
| 地震・津波     | GAMERA                    | 200,000            | 2.8                       | 1.1                   | 75               | 0.025                       | 0.005                   | 1                 |
| 気象気候      | 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                |
|           | GreeM                     | 100,000            | 6.9                       | 0.028                 | 100              | 25                          | 0.3                     | 300               |
| 宇宙・天文     | (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                |

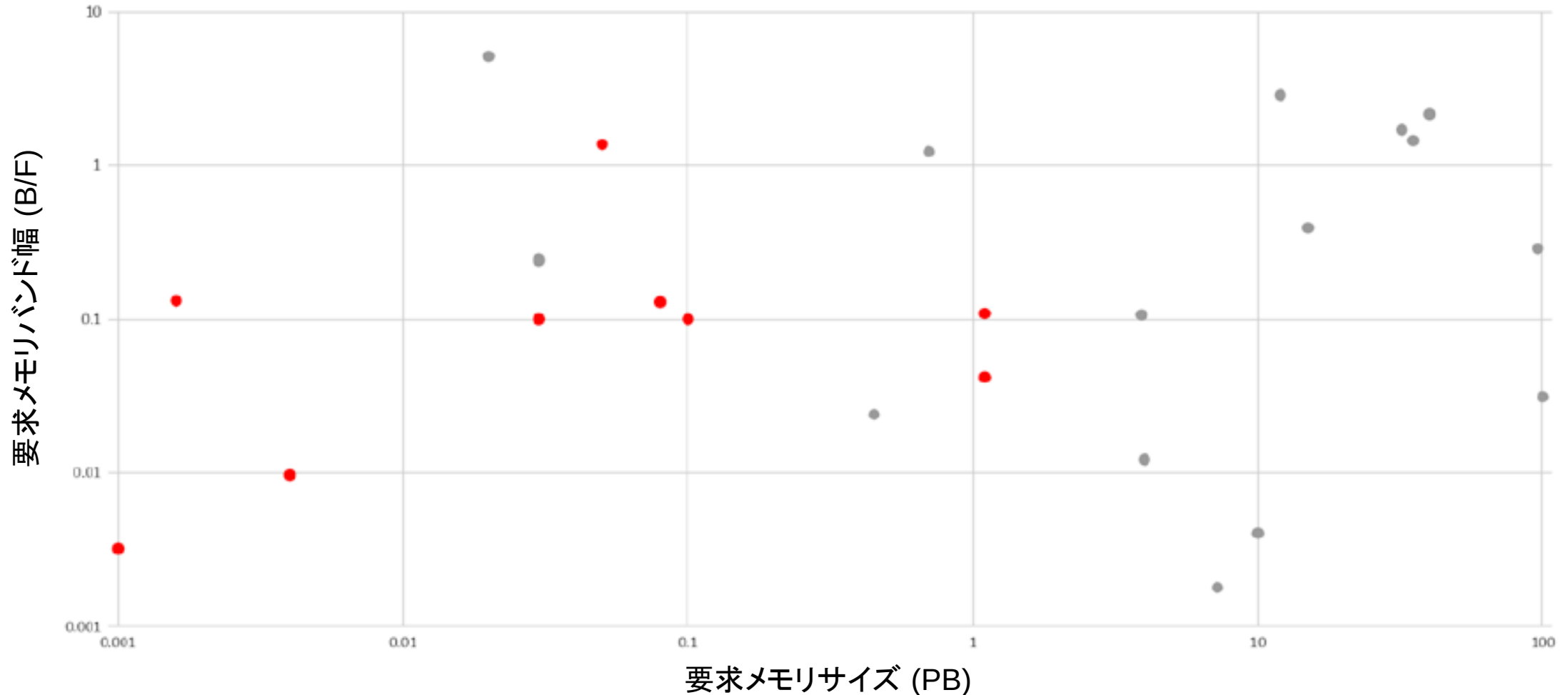
# アプリケーションの要求性能との比較

- 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



# 各アプリケーションの要求性能との比較

- メニーコア型システム (○および×は各コンポーネントが要求を満たしているかどうかを表す)

| App. Area | Name of Application       | Node | CPU | Memory | Interconnect | Storage | Total |
|-----------|---------------------------|------|-----|--------|--------------|---------|-------|
| 素粒子・原子核   | (unknown)                 | ○    | ○   | ○      | ○            | ○       | ○     |
|           | (unknown)                 | ○    | ○   | ○      | ○            | ○       | ○     |
|           | (unknown)                 | ○    | ○   | ○      | ○            | ○       | ○     |
|           | rmcsm                     | ○    | ○   | ○      | ○            | ○       | ○     |
|           | (unknown)                 | ○    | ○   | ×      | ×            | ○       | ×     |
| 物質科学      | HPhi                      | ×    | ○   | ○      | ○            | ○       | ×     |
| エネルギー・資源  | NTChem                    | ○    | ○   | ×      | ○            | ○       | ×     |
|           | SMASH                     | ○    | ○   | ×      | ○            | ○       | ×     |
|           | paraDMRG                  | ○    | ○   | ○      | ○            | ○       | ○     |
|           | GELLAN                    | ○    | ○   | ○      | ○            | ○       | ○     |
|           | MODYLAS                   | ○    | ○   | ○      | ○            | ○       | ○     |
| 脳科学・AI    | WHC                       | ○    | ○   | ○      | ○            | ○       | ○     |
|           | Realtime cerebellum       | ○    | ○   | ○      | ○            | ○       | ○     |
|           | NEURON K+ Stochastic      | ○    | ○   | ○      | ○            | ○       | ○     |
|           | CNN (Forward & Back-prop) | ○    | ×   | ×      | ○            | ○       | ×     |
|           | CNN (Forward)             | ○    | ×   | ×      | ○            | ×       | ×     |
| 地震・津波     | GAMERA                    | ×    | ○   | ○      | ○            | ○       | ×     |
| 気象気候      | NICAM                     | ×    | ○   | ○      | ○            | ○       | ×     |
|           | SCALE-RM                  | ○    | ○   | ○      | ○            | ○       | ○     |
|           | CHASER-LETKF              | ○    | ○   | ×      | ×            | ×       | ×     |
|           | NICAM-LETKF               | ×    | ○   | ○      | ○            | ×       | ×     |
| 宇宙・天文     | GreeM                     | ○    | ○   | ○      | ×            | ○       | ×     |
|           | (unknown)                 | ○    | ×   | ○      | ○            | ○       | ×     |
|           | EM-PIC                    | ○    | ○   | ×      | ○            | ×       | ×     |
|           | P3T                       | ○    | ○   | ○      | ○            | ○       | ○     |
|           | AmaTeRAS                  | ×    | ○   | ○      | ×            | ○       | ×     |

# 各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 (○および×は各コンポーネントが要求を満たしているかどうかを表す)

| App. Area | Name of Application       | Node | CPU | Memory | Interconnect | Storage | Total |
|-----------|---------------------------|------|-----|--------|--------------|---------|-------|
| 素粒子・原子核   | (unknown)                 | ○    | ○   | ○      | ○            | ○       | ○     |
|           | (unknown)                 | ○    | ○   | ○      | ○            | ○       | ○     |
|           | (unknown)                 | ○    | ○   | ○      | ○            | ○       | ○     |
|           | rmcsm                     | ○    | ○   | ○      | ○            | ○       | ○     |
|           | (unknown)                 | ○    | ○   | ×      | ×            | ○       | ×     |
| 物質科学      | HPhi                      | ×    | ○   | ×      | ○            | ○       | ×     |
| エネルギー・資源  | NTChem                    | ○    | ○   | ×      | ○            | ○       | ×     |
|           | SMASH                     | ○    | ○   | ×      | ○            | ○       | ×     |
|           | paraDMRG                  | ○    | ○   | ○      | ○            | ○       | ○     |
|           | GELLAN                    | ○    | ○   | ×      | ○            | ○       | ×     |
|           | MODYLAS                   | ○    | ○   | ○      | ○            | ○       | ○     |
| 脳科学・AI    | WHC                       | ○    | ○   | ×      | ○            | ○       | ×     |
|           | Realtime cerebellum       | ○    | ○   | ○      | ○            | ○       | ○     |
|           | NEURON K+ Stochastic      | ○    | ○   | ○      | ○            | ○       | ○     |
|           | CNN (Forward & Back-prop) | ○    | ×   | ×      | ○            | ○       | ×     |
|           | CNN (Forward)             | ○    | ×   | ×      | ○            | ×       | ×     |
| 地震・津波     | GAMERA                    | ×    | ○   | ×      | ○            | ○       | ×     |
| 気象気候      | NICAM                     | ×    | ○   | ×      | ○            | ○       | ×     |
|           | SCALE-RM                  | ○    | ○   | ○      | ○            | ○       | ○     |
|           | CHASER-LETKF              | ○    | ○   | ×      | ×            | ×       | ×     |
|           | NICAM-LETKF               | ×    | ○   | ○      | ○            | ×       | ×     |
| 宇宙・天文     | GreeM                     | ○    | ○   | ×      | ×            | ○       | ×     |
|           | (unknown)                 | ○    | ×   | ×      | ○            | ○       | ×     |
|           | EM-PIC                    | ○    | ○   | ×      | ○            | ×       | ×     |
|           | P3T                       | ○    | ○   | ○      | ○            | ○       | ○     |
|           | AmaTeRAS                  | ×    | ○   | ○      | ×            | ○       | ×     |

# システムへの要求の分類(1/2)

- 汎用システム要求
  - 従来型の延長である汎用的なメニーコア型システムを要求するアプリ
- メモリ性能要求
  - 特にメモリに対して容量・転送速度・アクセス手法の面での拡張を要求するアプリ
- 高演算性能要求
  - ノード数やコア数の増加を要求、あるいはSIMD演算機構への要求を持つアプリ
- ネットワーク性能要求
  - 通信バンド幅や特殊なネットワークトポロジーに関する要求を持つアプリ

| システムのタイプ | アプリケーション                                                                                                                                     |
|----------|----------------------------------------------------------------------------------------------------------------------------------------------|
| 汎用システム   | Rmcs, paraDMRG, GELLAN, MODYLAS, 生命科学, 創薬医x6, FFB, ADVENTURE, Front-Comp, GT5D, WHC, Realtime cerebellum, NEURON + Stochastic, SCALE-RM, P3T |
| メモリ性能    | (BW) NTChem, SMASH, バリオン2体波動関数計算, NICAM, LQCD, GAMERA, SCALE-RM                                                                              |
|          | (size) NTChem, CHASER-LTKF, EM-PIC                                                                                                           |
| 高演算性能    | (#node) GAMERA, NICAM, NICAM-LETKF, AmaTeRAS, (#core) CNN, ニュートリノ輻射輸送<br>(others) LQCD, SCALE-RM, NICAM                                      |
| ネットワーク性能 | GreeM, 精密格子QCD, AmaTeRAS                                                                                                                     |



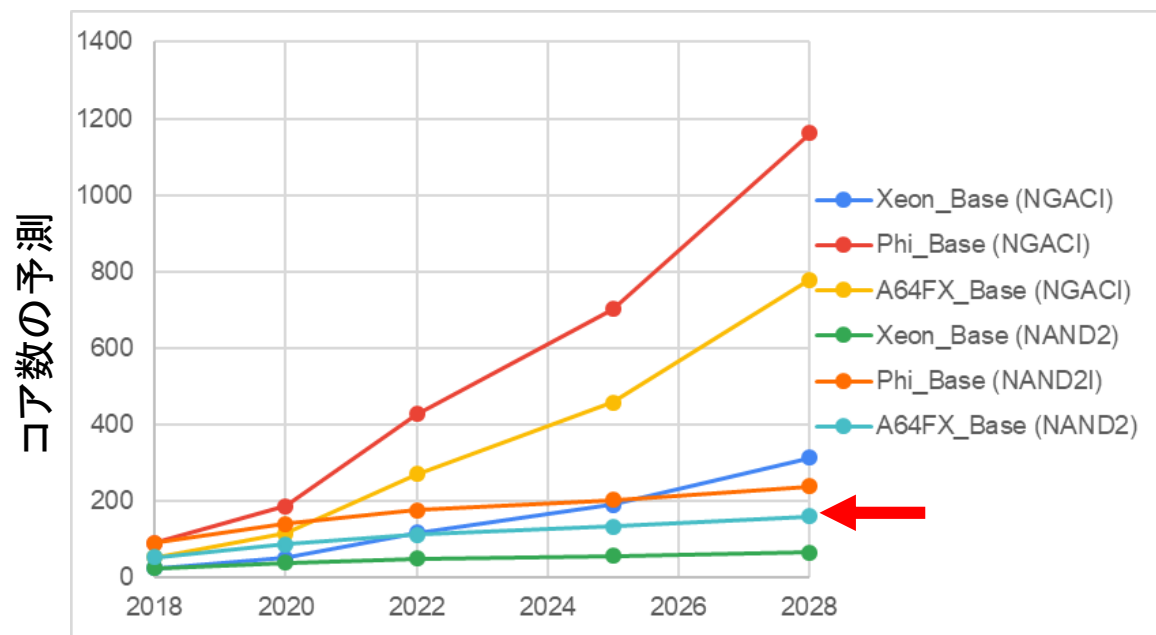
# システムへの要求の分類(2/2)

- 個別の要求
  - SCALE-SMD: 配列へのリストアクセス
  - 精密格子QCD: Allrecue専用NW、4次元構造格子を並列化しやすいNW
  - GENESIS: リアルタイム可視化のための高いポスト処理性能
- 専用アクセラレータの要求
  - 負荷分散支援ハードウェア(粒子計算)
  - エージェント処理のハードウェア実装(エージェントシミュレーション)
  - 疎行列専用ハードウェア(疎行列計算)
  - メモリアクセスのスケジューリング支援機構(モンテカルロ計算、グラフ処理)
  - アニーリング(イジングモデルによる最適化問題)
  - データフロー型かつ大量のレジスタを持つハードウェア(FFT)
  - ソーティングのハードウェア実装(ソーティング)

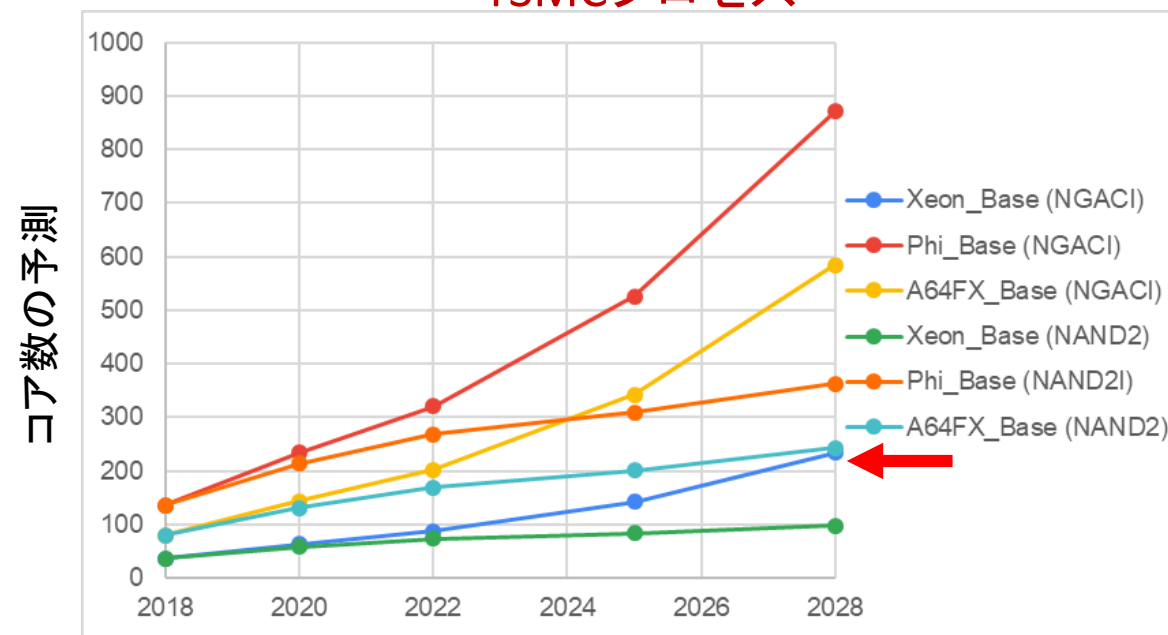
# より高性能なシステム構成へ向けて

- 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<sup>2</sup>、コア部割合=33%
  - TSMCプロセスでの仮定: チップサイズ450mm<sup>2</sup>、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

# 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

WP未公開データのため削除

- 50MW、80%におけるシステム性能の見積り

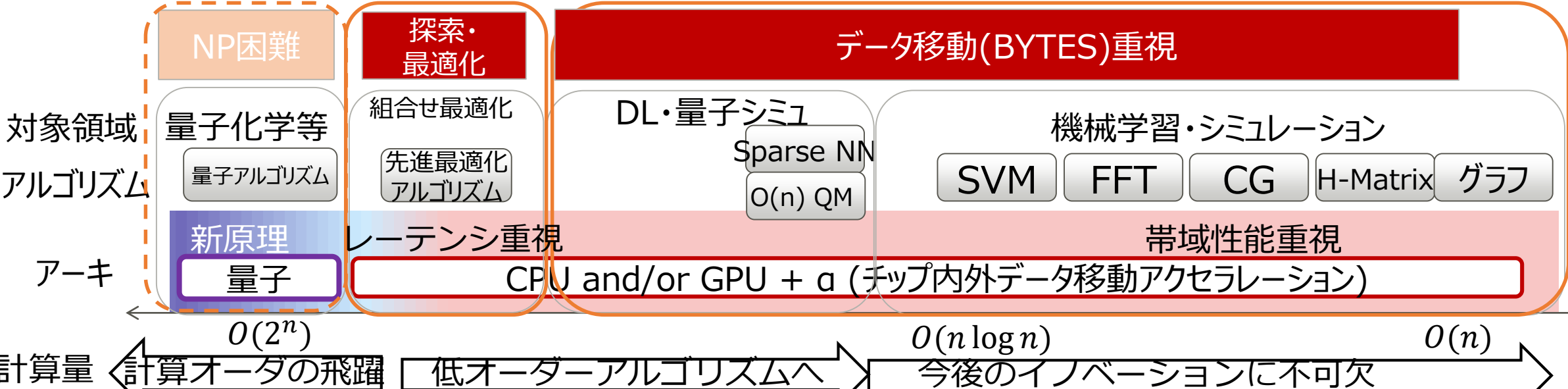
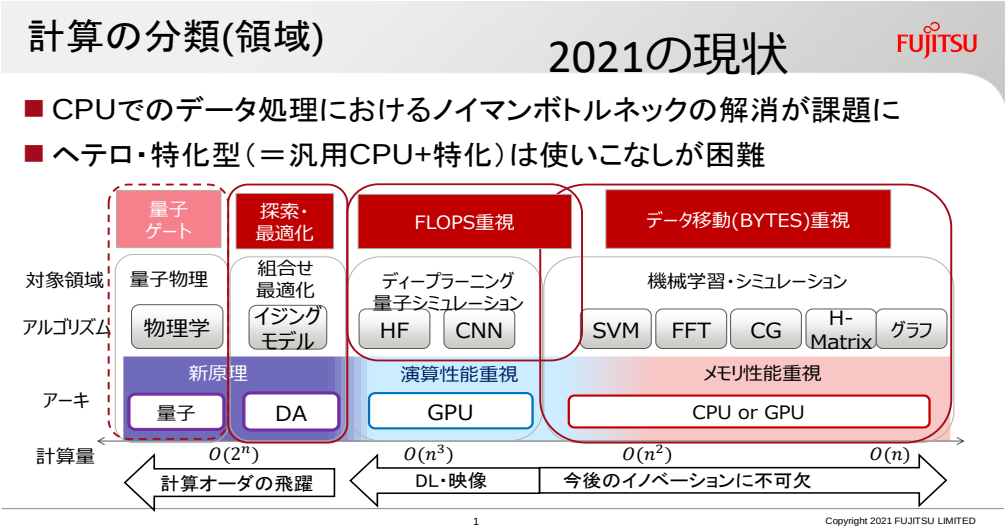
WP未公開データのため削除

# 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



# 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(350Wでノード当たり35TFLOPSの達成は簡単ではない)
  - 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
- メモリ
  - 積層メモリの広帯域化と大容量化、歩留まりの改善
  - ロジックとメモリの3次元積層化
  - 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トティング、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(TensorFlowなど)、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・ライブラリ的高速化
  - 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
  - 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

- これまでに頂いているWhite Paperへご意見
  - メニーコア型の性能見積りのベースとなるデータが保守的過ぎる
  - より幅広いアプリケーションで性能要求を調査することが必要
  - 新計算原理と古典コンピューティングのハイブリッド構成も検討すべき
  - 将来的な産業展開を当初から見据えた検討をすべき
- 現在White Paper 1.1.0版の公開に向けて作業中
- アプリケーション・システムのコデザインがこれまで以上に重要
- 計算科学の研究者の皆様と密に協力・連携させて頂ければ幸いです