将来の高メモリバンド幅アプリケーションに適したシステムについて

板倉　憲一
海洋研究開発機構

2013年8月27日

SS研HPCフォーラム2013

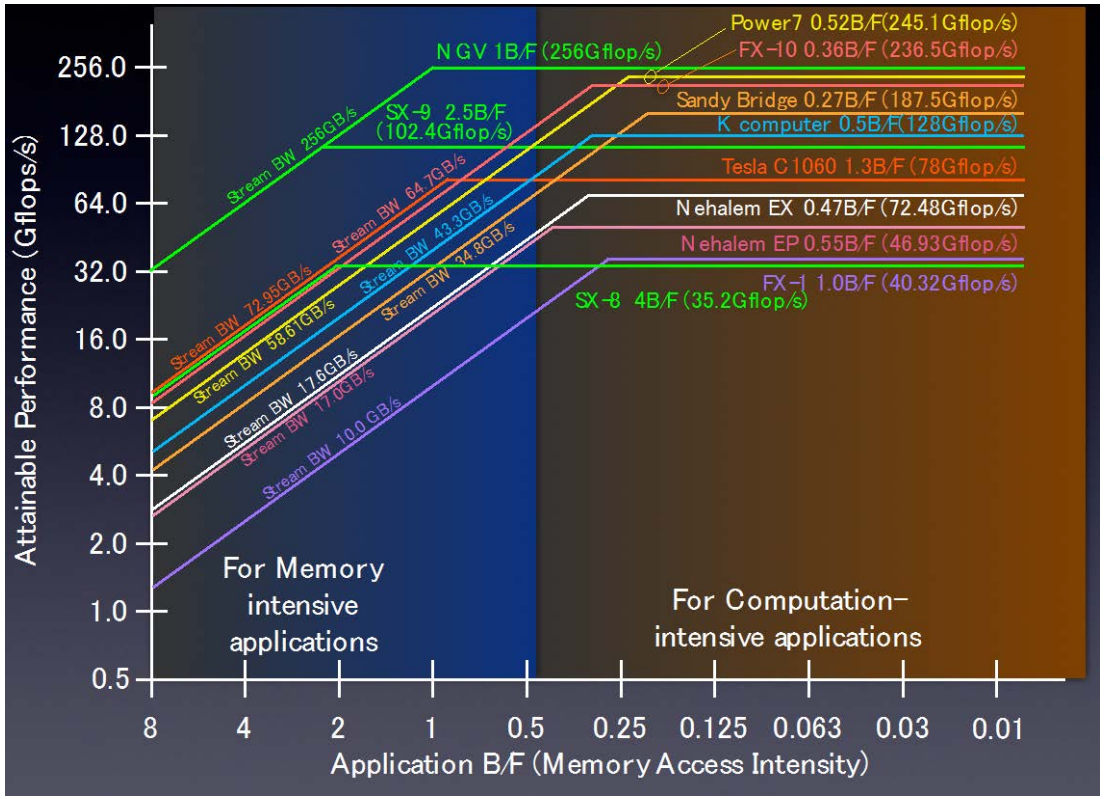
東北大学　海洋研究開発機構　日本電気株式会社

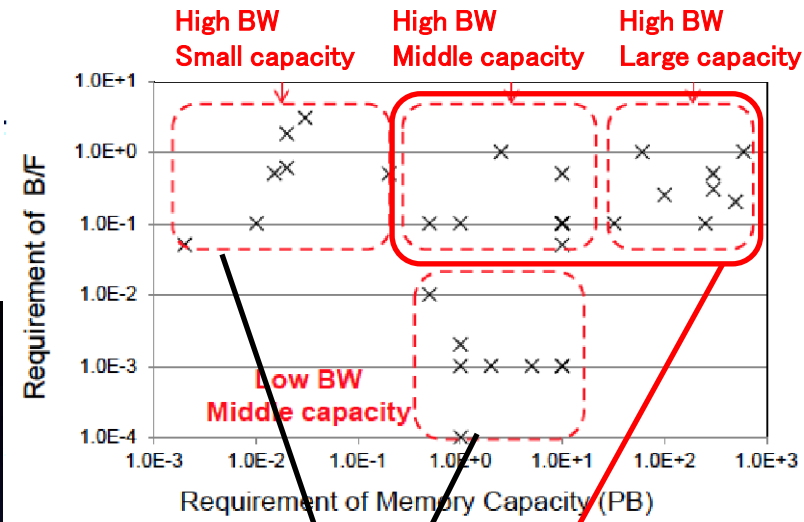
講演内容

- 研究の概要と目的
- 対象するアプリケーション
  - 総合防災
  - ものづくり
- 要素技術・アーキテクチャ検討
  - デバイス
  - 高速・低消費電力チップ間配線
  - ノード間ネットワーク
  - ストレージシステム
  - システムソフトウェア
- 性能推定・実現性
  - アプリケーションの要求性能
  - システム諸元
  - コモディティクラスタとの性能比較
- まとめ

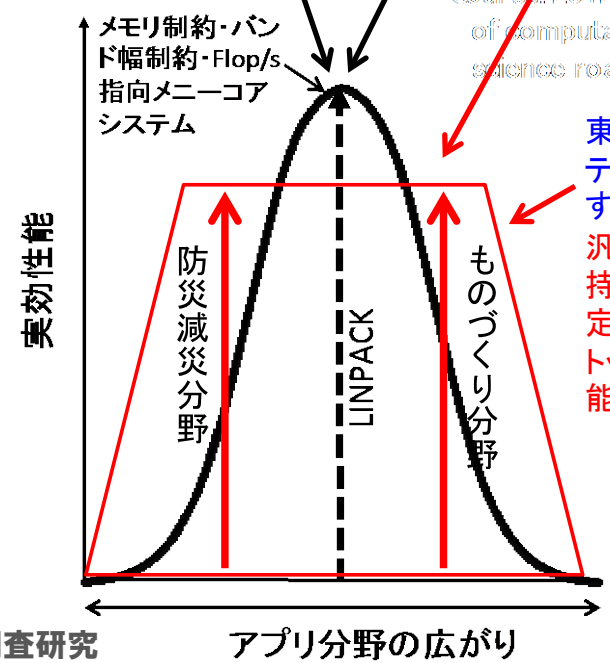
東北大チームの基本的な考え方

Flop/s性能向上にあわせてメモリバンド幅を高めることにより，可能な限り眠っている演算器を活用し，ノードレベルの超並列性に頼ることをできるだけおさえ，広汎なアプリで高い実効性能を実現する





Source: 2012 report of computational science roadmap



東北大システムが目指すところ

汎用性を維持しつつ，特定分野ではトップ実効性能を実現

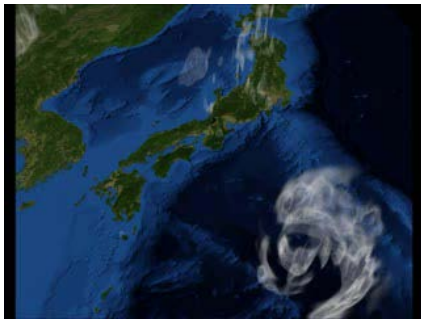
東北大チームがめざすこと

✧ **2020年頃に解決が求められる社会的・科学的課題の明確化とその解決のためのアプリケーションの検討・開発ロードマップの策定**

– 自然災害，とりわけ地震津波による災害軽減を目指すため，沿岸データを使用したリアルタイム津波予測・地震シナリオを含む津波ハザード予測・地震津波に伴う複合被害予測・集中豪雨等を含めた都市災害と避難の予測といった課題に取り組む．これらは津波が共通のキーワードであり，特に本事業においては，津波計算に関わるターゲットアプリケーションを用いたシステム評価や課題検討を中心に行う．

1) 自然災害に対する防災・減災，および先進ものづくり分野での社会的・科学的課題の明確化
2) 社会的・科学的課題を解決するための地震・津波・気象・気候,およびプラズマ・CFD等に関するターゲットアプリケーションの調査研究
3) ターゲットアプリケーションの実行に適したHPCI システムの機能・性能要件の明確化
4) システム評価用ベンチマーク開発

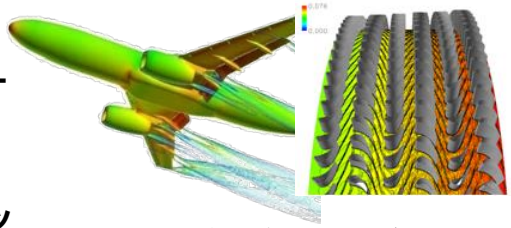
地震・津波に伴う被害想定


気象・気候に伴う災害想定

✧ **上記ターゲットアプリケーションに適したHPCIシステムの検討・開発ロードマップの策定**

– 地震・津波，集中豪雨など都市災害の被害予測と避難予測に不可欠なアプリケーションがHPCIシステムに必要とする性能・機能を明らかにし，その実現のためのアーキテクチャおよびデバイス技術の検討を通じて，2018年頃にシステムを実現するための開発ロードマップの策定を行う．

1) ターゲットアプリケーションの性能要件（例えば，1B/F以上のメモリバンド幅・演算スループット比）を低消費電力で実現するためのアーキテクチャの調査研究
2) 大容量・低消費電力・階層型メモリサブシステムアーキテクチャ検討
3) アーキテクチャ，およびデバイスレベルでの低消費電力技術や耐故障技術と，OS,アプリケーションが協調してその効果の最大化を可能とするシステムソフトウェア技術の検討
4) オフラインシミュレーションの高速化に加え，様々な時間制約の中でリアルタイムシミュレーションを実現する上で必要な将来のHPCIシステムが備えるべき機能・性能の明確化
5) ターゲットアプリケーションによるシステムの評価と2018年頃の実現を目指す開発ロードマップの策定


次世代ものづくり

体制と役割

東北大学
(事業・機関代表：小林広明)

- サイバーサイエンセンター
- 災害科学国際研究所
- 情報科学研究科
- 未来科学技術共同研究センター
- 電気通信研究機構
- 電気通信研究所

海洋研究開発機構
(機関代表　金田義行)

日本電気株式会社
(機関代表　橋本ユキ子)

アプリケーション検討サブグループ
- 東京大学
- 宇宙航空研究開発機構

役割：社会的・科学的課題を明確にし，その解決アプリの検討と開発ロードマップの策定

システム検討サブグループ
- 大阪大学
- 北陸先端科学技術大学院大学
- 理化学研究所

役割：ターゲットアプリにマッチしたHPCIシステムの検討とその開発ロードマップの策定

デバイス技術検討サブグループ
- 明星大学
- 東北マイクロテック株式会社
- 国外メモリメーカ

役割：アーキテクチャを実現する要素デバイス技術の検討とその開発ロードマップの策定

アプリケーション検討サブグループ体制

リーダー　金田　義行(海洋機構)　　取りまとめ　渡邉　國彦（海洋機構）

地震検討グループ

堀　高峰　（海洋機構）
兵藤　守　（海洋機構）
堀　宗朗　（東京大）
市村　強　（東京大）
古村　孝志（東京大）
豊国　源知（東北大）

津波検討グループ

今村　文彦　（東北大）
越村　俊一　（東北大）
寺田賢二郎（東北大）
有川　太郎　（港湾技研）
阪口　秀　（海洋機構）
西浦　泰介　（海洋機構）

流体検討グループ

中橋　和博（宇宙機構）
山本　悟　（東北大）
佐々木大輔（金沢工大）
高橋　俊　（農工大）
松岡　浩　（東北大）

気象検討グループ

高橋　桂子　（海洋機構）
石川　洋一　（海洋機構）
木田　新一郎（海洋機構）

性能最適化設計グループ

撫佐　昭裕　（日本電気）

システム検討サブグループ・デバイス検討サブグループ体制

大学・研究機関側は，システムの基本検討とその詳細化・ターゲットアプリによる評価を主に担当

大学・研究機関側総括責任者
小林

アプリ検討・システム要件策定・評価検討WG 板倉，横川，上原，浅野，小松	アプリ検討・システム要件策定・評価検討WG 撫佐　他
連携	
システムソフトウェア検討・システム要件策定WG 滝沢，佐藤，宇野	システムソフトウェア検討・システム要件策定WG 林　他
トップダウン ターゲットシステム ボトムアップ	
アーキテクチャ検討WG 江川，多田	アーキテクチャ検討WG 百瀬，萩原　他
ネットワーク検討WG 佐野，滝沢，荒川，長谷川	ネットワーク検討WG 百瀬，萩原　他
ストレージ・I/O検討WG 横川，板倉，上原，浅野，宇野，村岡，	ストレージ・I/O検討WG 百瀬，萩原　他
協調設計・連携評価	
デバイス検討 サブグループ（WG） 小柳，福島，李，元吉	デバイス検討 サブグループ（WG） 梅澤　他

NEC側総括責任者
橋本

NEC側は，システム実装の観点からシステム検討・シミュレータ設計とシステム評価を主に担当

青文字はWGとりまとめ担当

講演内容

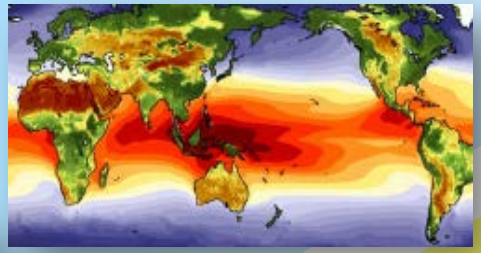
- 研究の概要と目的
- 対象するアプリケーション
  - 総合防災
  - ものづくり
- 要素技術・アーキテクチャ検討
  - デバイス
  - 高速・低消費電力チップ間配線
  - ノード間ネットワーク
  - ストレージシステム
  - システムソフトウェア
- 性能推定・実現性
  - アプリケーションの要求性能
  - システム諸元
  - コモディティクラスタとの性能比較
- まとめ

将来のHPCIに対する社会的・科学的ニーズ

安全安心な社会実現に向けた防災・減災

集中豪雨

地震



気候変動

津波

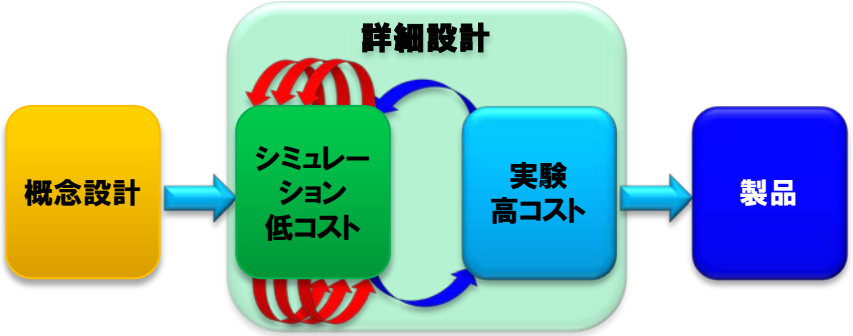
輸送技術

発電技術

信頼性・安全性・生産性の革新

総合防災：地震・津波・気象による災害の科学的な予測

ものづくりにおける革新的設計



ディジタルデザインの活用が産業界でのイノベーションの創出を推進し，製品開発における我が国の国際競争力を強化

シミュレーションの積極的利用による設計コストの大幅削減と設計期間の短縮

設計空間の拡大
模型実験の縮小

⇒

信頼性・安全性・生産性の向上
環境配慮機器・省エネ化の実現

産業界へのフィードバック

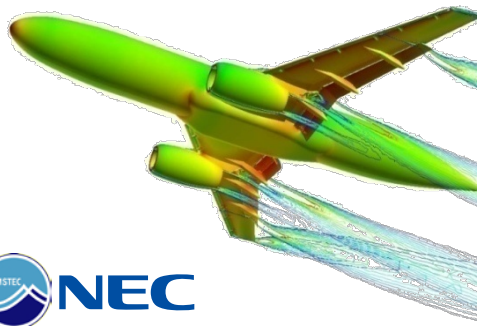
① Digital Flightの実現
（定常から非定常現象の再現）
②静粛航空機の設計
（空力音響解析の実現）

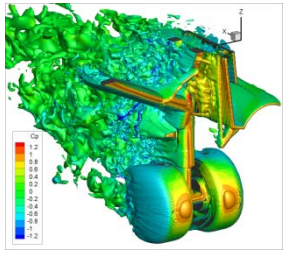
① 高効率タービンの実現
（タービンをまるごと熱流動解析）
② マルチフィジックスCFDの実現
（相変化，腐食，破壊の実現）

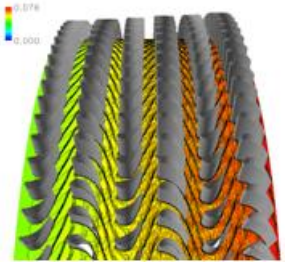
航空機設計

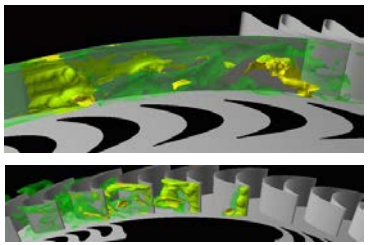
発電機器設計

マクロな流れと共にミクロな現象のマルチスケールシミュレーションが実現

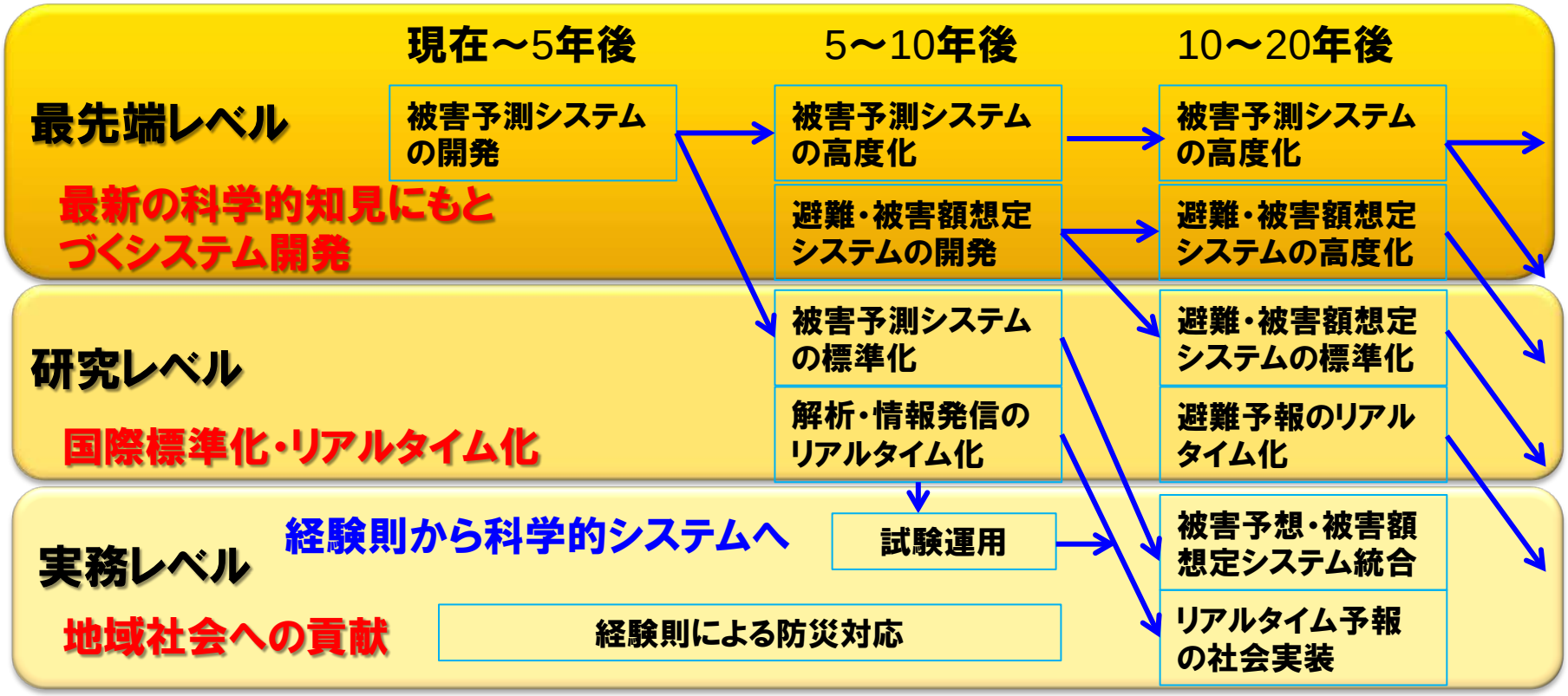






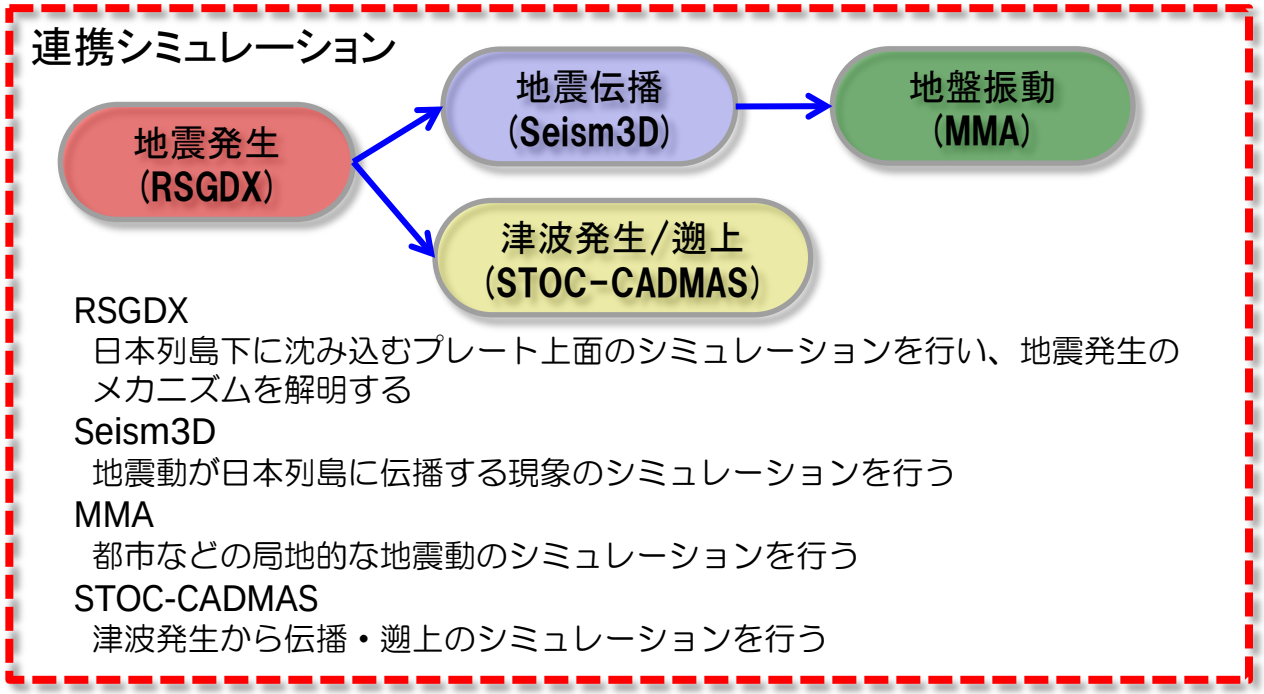


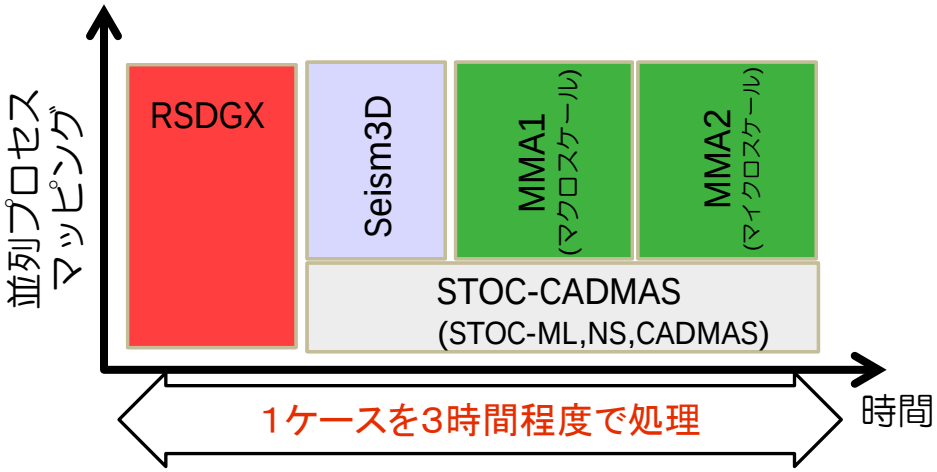
科学的知見による予報・被害想定のシステム化

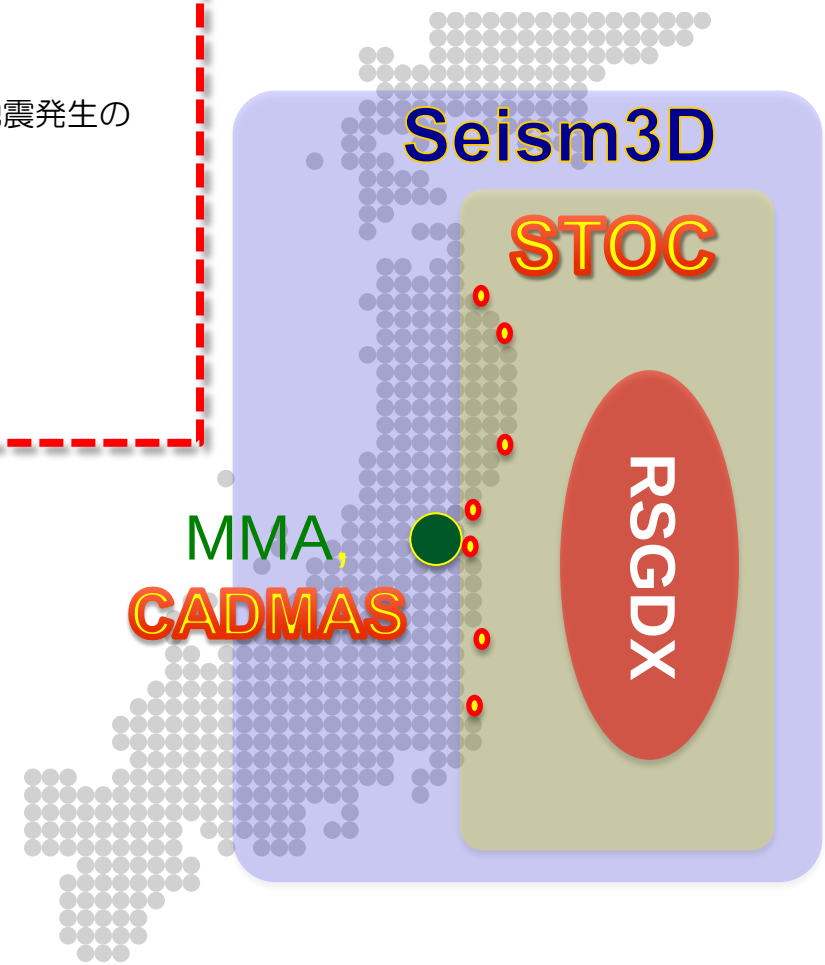


－国や自治体で使われる標準化（開発から社会実装までの流れを作る）
－リアルタイム予報（警報，注意報），避難情報の高度化，および予警報解除のための予測技術の高度化と迅速化
－発災直後の被害状況推計および把握・救助・復旧対策のための被害予測の迅速化
－科学的知見とその不確かさにもとづいた想定幅の明確化

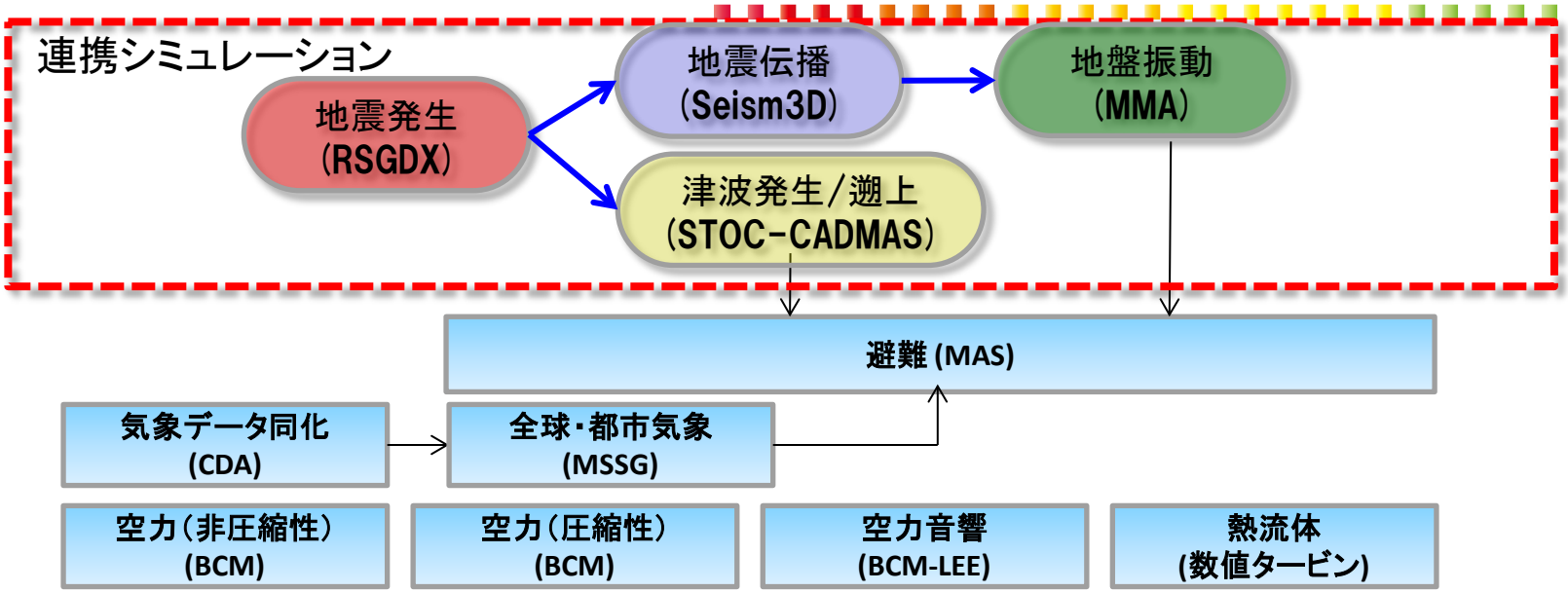
総合防災アプリケーション連携イメージ

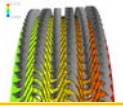






評価アプリケーションと性能要求要件



目的	アプリケーション	B/F	必要メモリ(TB)	必要FLOP数 (×10^{18})	要求実行時間(H)
超高解像度単独計算	RSGDX	8.00	14	520	24
	Seism3D	2.14	2,900	1,000	8
	MSSG	4.00	175	720	6
	BCM	5.47	13.6	1	0.5
高解像度アンサンブル計算	総合防災の連携シミュレーション (1000ケース)	2.1～8.0	98	25x1000	3×1000 ＝約4ヶ月
	数値タービン (20ケース同時実行) 	2.33	163.5	140	20

アプリケーションが求めるシステムを目指して

高メモリバンド幅を要するアプリケーションのためのアーキテクチャ検討

メモリ性能
■より高いB/Fのメモリバンド幅
■配列の間接参照の高速化

→ **実効2B/F以上の高メモリバンド幅**
■ 新デバイス技術によるメモリカスタム設計
＋オンチップベクトルロードストア支援機構

演算性能
■短ベクトル長にも対応

→ **ベクトル処理制御機構の高度化**

ノード性能
■ノード性能が高く、並列数が少なくて良いシステム

→ **高いコア性能**

ネットワーク性能
■近接通信を重視しつつも、大域通信も低レイテンシ

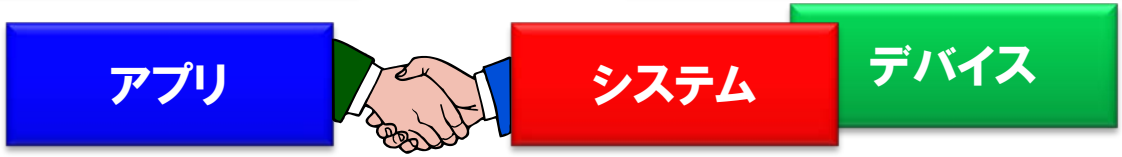
→ **High-Radixスイッチによる小直径・低遅延トポロジ**

ストレージI/O性能
■データ同化に耐えうる高容量・大規模ストレージ

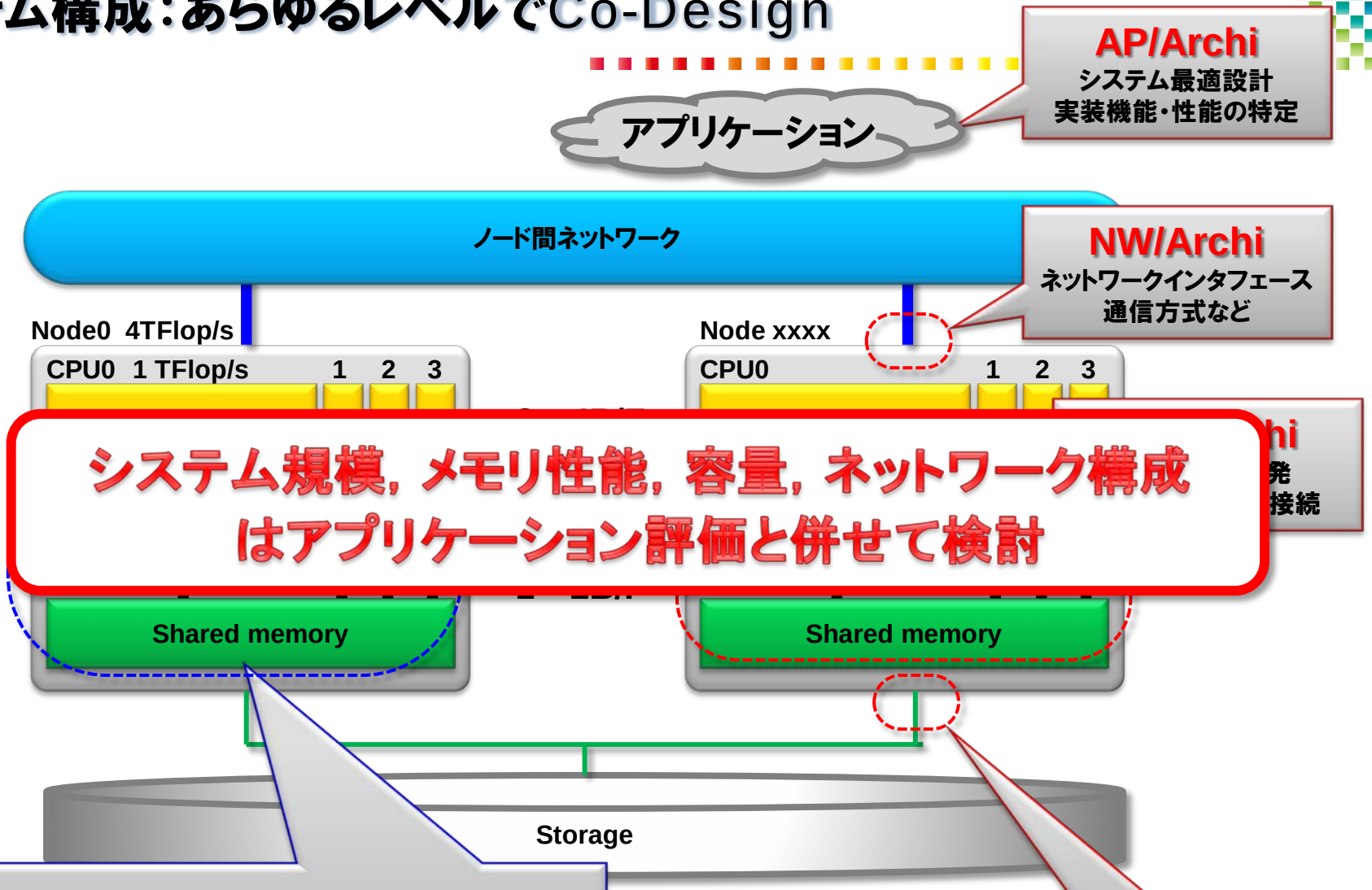
→ **階層型ストレージシステム**

システムソフトウェア
■ソフトウェアスタック全体の標準との互換性
■多様かつ多数のジョブの高効率実効

→ **Linuxをベースにシステムの性能を引き出すシステムソフトウェア開発**



システム構成：あらゆるレベルでCo-Design



Architecture

- VLSBやMSHRを中心としたCPUアーキテクチャによるメモリアクセス効率化
- アーキテクチャ技術，デバイス技術による省電力化
- CPU/メモリの高速接続・SMP構成・実現性検討

IO/Archi

File Systemインタフェース
実装方式など

講演内容

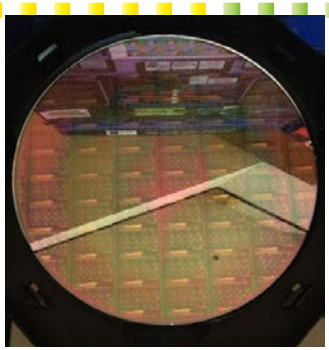
- 研究の概要と目的
- 対象するアプリケーション
  - 総合防災
  - ものづくり
- 要素技術・アーキテクチャ検討
  - デバイス
  - 高速・低消費電力チップ間配線
  - ノード間ネットワーク
  - ストレージシステム
  - システムソフトウェア
- 性能推定・実現性
  - アプリケーションの要求性能
  - システム諸元
  - コモディティクラスタとの性能比較
- まとめ

要素技術抽出の妥当性：デバイストレンド

■ 半導体加工技術のトレンド

- ■ 2018年に向けた製品開発
  - ■ 2016年の詳細設計，検証が必須
  - ■ 2016年は14nmクラスの加工技術が主流に
    - – Global Foundry, TSMCは2014年を目標に 14nmプロセスのサービスを提供予定 (@IEDM2012)

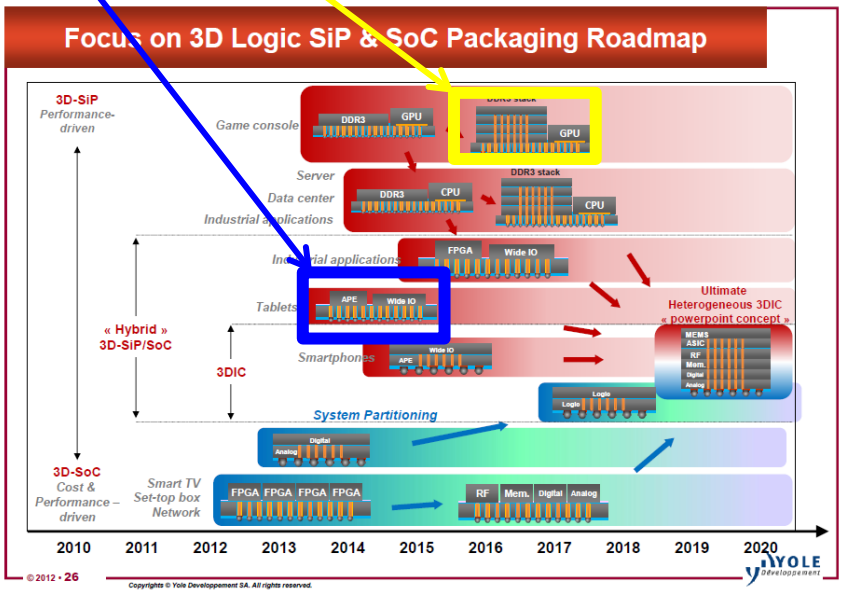
10 – 16nmのプロセスを想定



A・STAR社の12インチ(300mm)ウェハ 25mmx40mmのインタポーザ@3DAsiP

■ 2.5D，3D実装のトレンド

- ■ 3次元積層技術
  - ■ 積層型DRAM(HMC)は2013～2014年に市場へ
  - ■ メモリ・ロジック積層は2018年頃 に展開されることが予想
- ■ 2.5次元積層技術
  - ■ 2011年にVirtex7が出荷
  - ■ 2013年には2.5D実装のモジュールが発表予定
    - – GPU with memory on Si Interposer
  - ■ 2014年には主要CADベンダーが EDAツールを提供開始予定
  - ■ 2018年には50mmx50mmのSiインタポーザが実現
- ■ 我が国においても，文科省，経産省の支援のもと，2.5D/3D技術の実用化研究が急ピッチで進んでいる



L. Cadix, "3DIC & 2.5D Interposer market trends and technological evolutions", 3DASIP (2012)

性能と電力要件を満たすには微細化とLSI積層を考慮したシステム設計が鍵

テストデバイス試作評価環境

東北大学3Dイノベーション拠点設備
仙台市多賀城市桜木3-4-1 みやぎ復興パーク内

以下のプロセスモジュール設備を所有し、この組み合わせにより3次元積層チップ/TSV/光導波路付高機能インターポーザー等の最先端試作評価が可能である

TSV形成プロセスモジュール
ウェハ研磨装置
TSV穴あけ用エッチャー
絶縁膜形成装置
メタル堆積装置
Cuメッキ装置

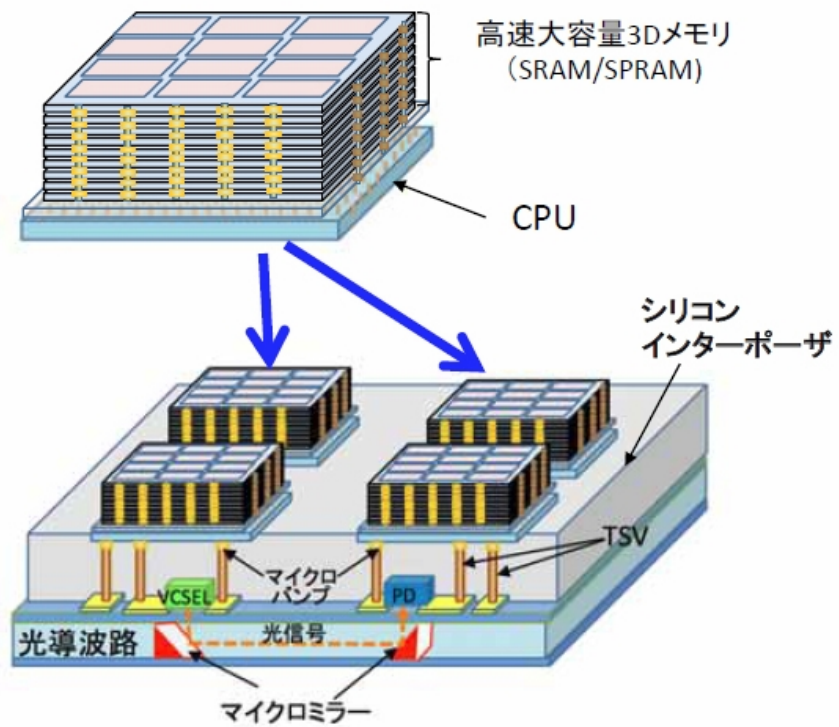
マイクロバンプ形成プロセスモジュール
メタルバンプ堆積装置
Cuメッキ装置
リフトオフ装置

微細パターン形成プロセスモジュール
縮小投影露光装置
1:1露光装置
フォトレジスト塗布装置
現像装置

貼り合わせプロセスモジュール
貼り合わせ装置
ﾌﾘｯﾌﾟﾁｯﾌﾟボンダ

メタル配線形成プロセスモジュール
微細パターン形成プロセスモジュール
メタルエッチャー
洗浄装置

目標デバイス構造



VCSEL:面発行レーザダイオード
PD:フォトダイオード
SPRAM:Spin-transfer torque RAM-- 高速不揮発性RAM

ネットワークの検討方針・目標と初期検討結果

■検討方針・目標

- ■対象アプリの通信要求に対し、近接通信を重視しつつも、大域通信も低レイテンシなNW
- ✓Fatノード＆高基数SWによる小直径NWを中心に、性能・コスト・電力の点で有望な構成を検討
- ■RAS等の技術的指針・研究開発ロードマップの提示

■前提とするデバイス技術

- ■2018年頃に実現の高基数SW
- ■電気・光インターコネクト

NWの全体構成

■NWトポロジの初期検討結果

	直径	リンク数
Fat tree 汎用性・運用性重視	◎	○
	△ケーブル遅延大	
低次元Torus コスト・拡張性重視	×	○
高次元Torus コスト・拡張性重視	○	×
Dragonfly 階層構成による 疑似高基数NW	◎	×
FTT-Hybrid Fat treeとTorusの 複合型	○	○
	◎ケーブル遅延・コスト低減化	

実装を考慮した詳細評価の実施へ

ストレージシステム

基本的要求要件

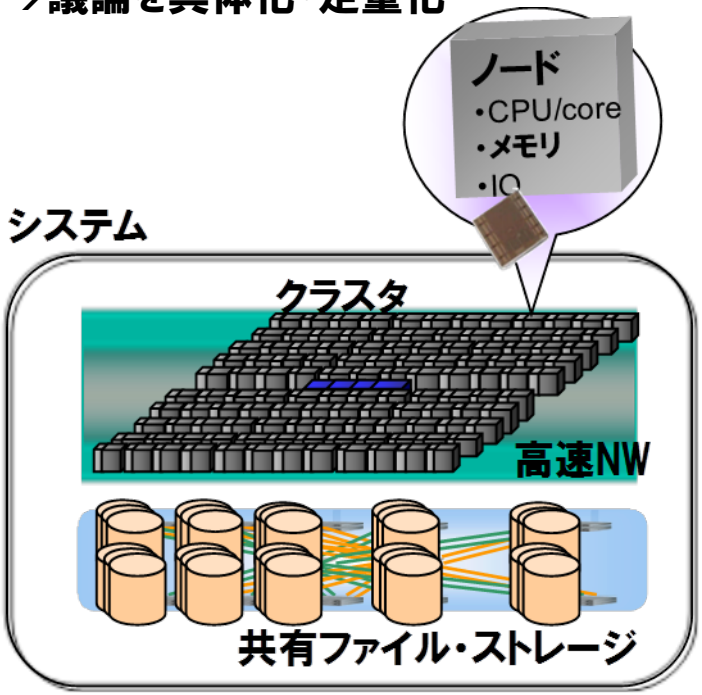
- スケーラビリティ:　数万～数十万クライアントでも動作
  - I/O競合、メタデータ競合の低減/削減
  - I/Oネットワークにおけるアクセス優先度制御(メタ、データ)
  - 帯域保障制御による性能確保
- コンシステンシ:　データ・メタデータの一貫性を保証
- 耐障害性:　障害遭遇を前提にしたファイルシステムの設計，単一障害点の排除
  - ストレージやI/Oサーバをまたぐソフトウェア RAIDによるデータ保証
  - 複数サーバ障害対応
  - パーティション分割等の高度化など
- EtoEの整合性保証　:ソフトウェアバグ等の論理的不整合を検出
- チェックポインティング機能
- ストレージ容量
  - ローカル領域容量　:メモリ容量の10倍程度を検討
  - グローバル領域容量 :最大でピーク性能(Flop/s値)の10倍までを検討

システムソフトウェア

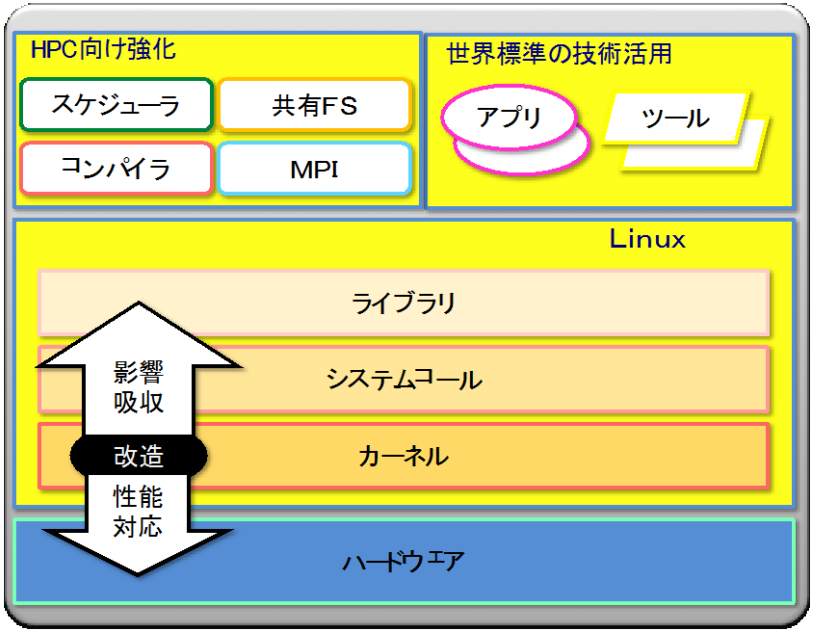
- **特にFS機で検討すべき課題を選定**
  - 課題に優先順位をつけて調査
  - 定量的な評価に基づき他WGと議論
- **文献調査に基づく現状分析**
  - 既存の技術では解決できない問題の明確化
  - 問題解決のためにHWに求める機能を明確化
  - アプリ開発者をどのように支援すべきかを議論

  →議論を具体化・定量化

- **特に優先順位の高い課題**
  1. 標準的な環境との互換性
     - Linuxを次世代HPC機の新アーキテクチャに対応
     - コンパイラ等を含めた開発環境の互換性向上
  2. スケーラビリティ
     - システムのリソースを有効活用するジョブ制御
     - 性能解析ツールとメモリ階層制御
  3. 耐故障性
     - チェックポイント・リスタートの高速化
     - 冗長リソース割り当てによるFTジョブ制御



FSベースシステムのアーキテクチャ



講演内容

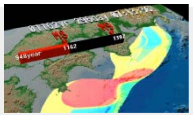
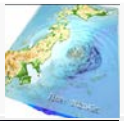
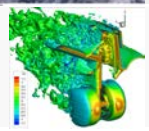
- 研究の概要と目的
- 対象するアプリケーション
  - 総合防災
  - ものづくり
- 要素技術・アーキテクチャ検討
  - デバイス
  - 高速・低消費電力チップ間配線
  - ノード間ネットワーク
  - ストレージシステム
  - システムソフトウェア
- 性能推定・実現性
  - アプリケーションの要求性能
  - システム諸元
  - コモディティクラスタとの性能比較
- まとめ

性能推定結果(1/2)

■ 性能要求要件と評価システムでの性能評価のまとめ

■VLSBあり＋2B/F構成時

【超高解像度 and/or 長時間単独計算】

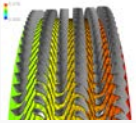
アプリケーション	B/F	必要倍精度浮動小数点演算数（×10^{18}）	メモリ容量	要求実行時間	推定結果	推定CPU数 (Peak Pflop/s)
RSGDX 	8.00	520	14 TB	24 時間	23 時間	100,000 (100 Pflop/s)
Seism3D 	2.14	1,000	2,900 TB	8 時間	7.4 時間	100,000 (100 Pflop/s)
MSSG	4.00	720	175 TB	6 時間	20.9 時間	102,400 (102 Pflop/s)
BCM 	5.47	1	13.6 TB	0.5 時間	0.38 時間	65,536 (66 Pflop/s)

性能推定結果（2/2）

■ 性能要求要件と評価システムでの性能評価のまとめ

■VLSBあり＋2B/F構成時

【高解像度アンサンブル計算】

アプリケーション	B/F	必要倍精度浮動小数点演算数（×10^{18}）	メモリ容量	要求実行時間	推定結果	推定CPU数 (Peak Pflop/s)
総合防災の連携シミュレーション（1000ケース）	2.1～8.0	25×1000	98 TB	3時間×1000＝約4ヶ月	3.1 時間×1000	100,000 (100Pflop/s)
数値タービン（20ケース同時実行） 	2.33	140	163.5 TB	20時間	17.1 時間	56,260 (56Pflop/s)

システム諸元（例えば，100PF構成での見積）

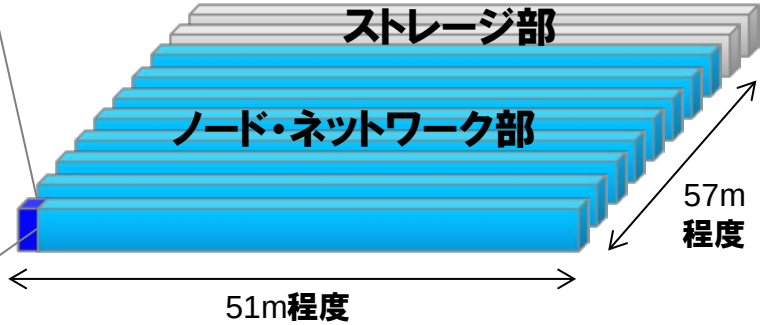
ピーク演算性能	100PFlop/s
総ノード数	25,000
総CPU数（コア数）	100,000 (400,000)
総メモリ帯域	100~200PB/s（2B/F）
総メモリ容量	3.2~12.8PB

筐体数	400(ノード)+300(NW)程度
総I/O帯域	10~30TB/s
総ディスク容量	300PB~1EB
システム設置面積	2,900m^{2}程度
本体系消費電力	25~MW程度

*A~C案における見積もり値

ラック

システム



ノード数： 64（256CPUs）
演算性能： 256TFlop/s
メモリ帯域： 256~512TB/s(1~2B/F)
メモリ容量： 8~32TB

耐故障性・信頼性

現在のHPCIシステム資源提供に用いられている大規模システムと同等以上の信頼性確保を目指す

- HWモニタリングによる障害回避機構
- HWによるデータ保護、エラー救済機構
- HW/SW連携による障害極小化機構

まとめ

■研究の概要と目的

- ■高計算性能と高メモリバンド幅で超並列に頼ることをできるだけおさえ、広汎なアプリで高い実効性能を実現することを目指す。

■対象するアプリケーション

- ■「総合防災（地震・津波・気象）」と「ものづくりにおける革新的設計」のアプリケーションの重要性、将来のシステムの利用形態、性能要求を明らかにした。
- ■アプリケーションとシステム・デバイスの研究者が協力してシステム設計を行なっている。

■要素技術・アーキテクチャ検討

- ■プロセス加工技術と3次元メモリ実装技術の発展が性能と電力要件を満たす鍵となる。
- ■ノード間ネットワークやストレージシステムを含めたトータルシステムのアーキテクチャ検討を行った。更に、プログラムの互換性・移植性や利用を支援するシステム・ソフトウェアも課題を明らかにしている。

■性能推定・実現性

- ■特に総合防災のアプリケーションの要求性能を詳細に検討し、それに必要で実現可能性のあるシステム諸元を示した。
- ■コモディティクラスタとの性能比較を行った結果、高メモリバンド幅のシステムが優位性を持つことを示した。