資料１－１

将来のHPCIシステムのあり方の調査研究

「高メモリバンド幅アプリケーションに適した将来のHPCIシステムに関する調査研究」

小林　広明
東北大学

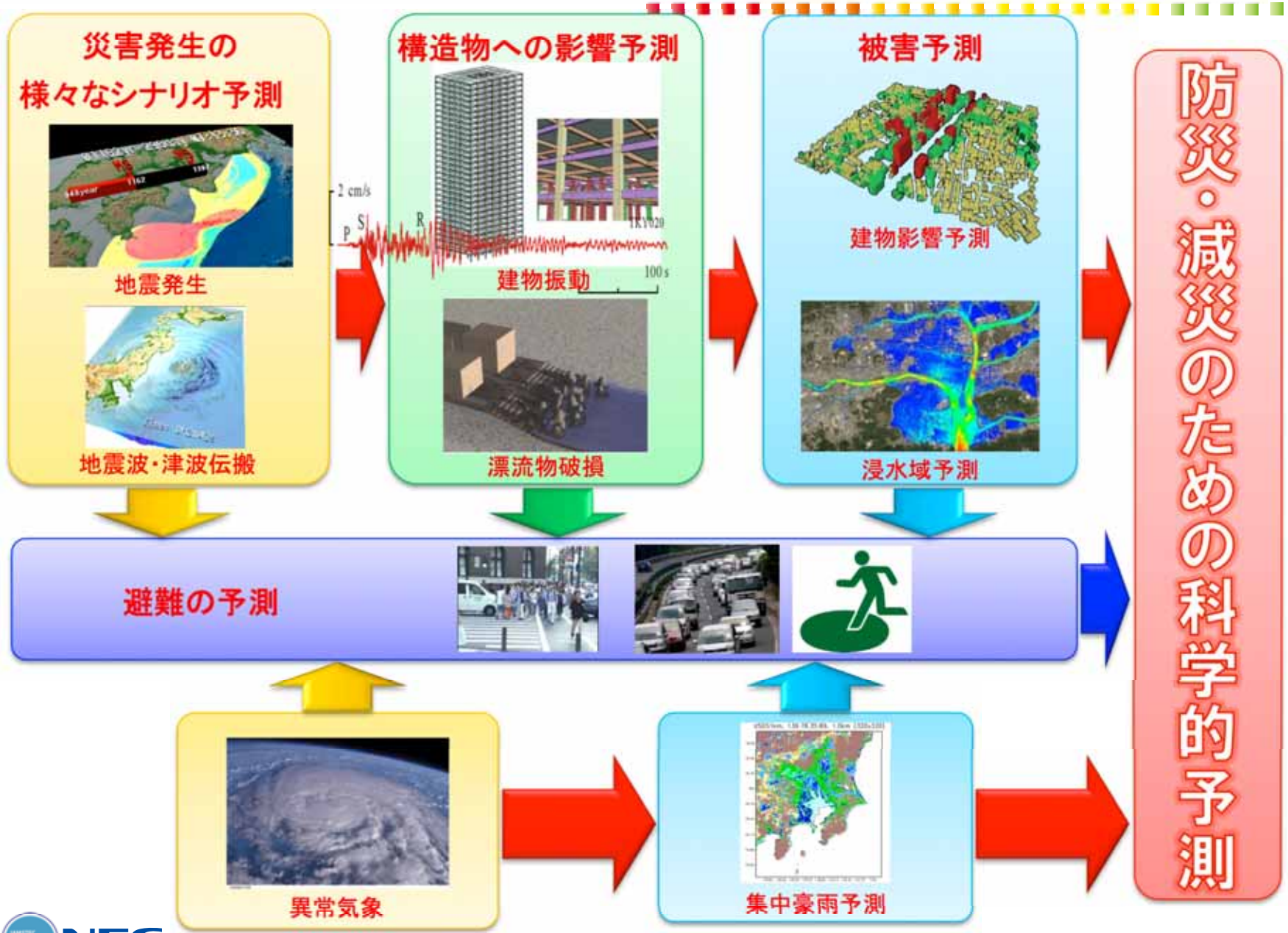
平成25年3月27日

東北大学

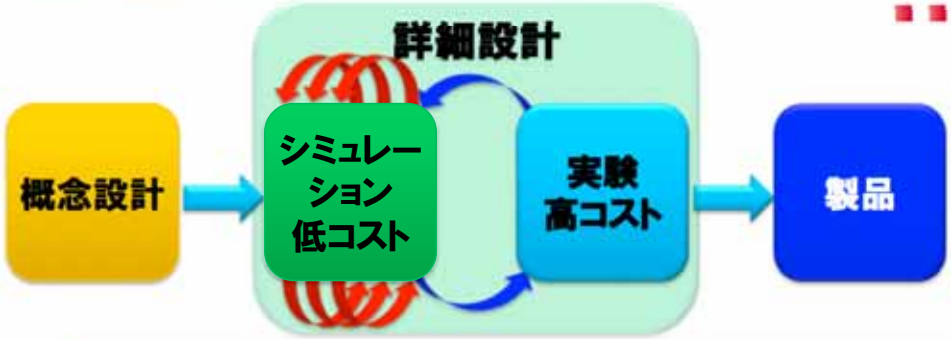
海洋研究開発機構

日本電気株式会社

社会的課題1：地震・津波・気象による災害の科学的な予測



社会的課題2：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

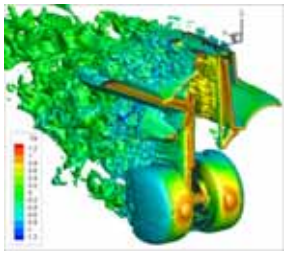
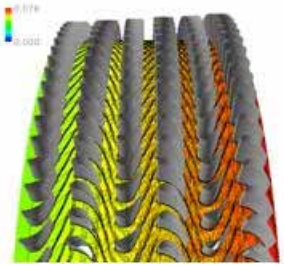
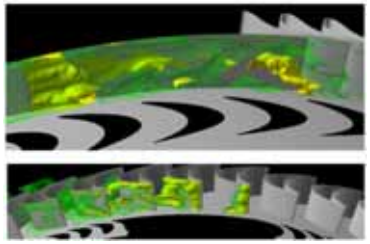
産業界へのフィードバック

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）

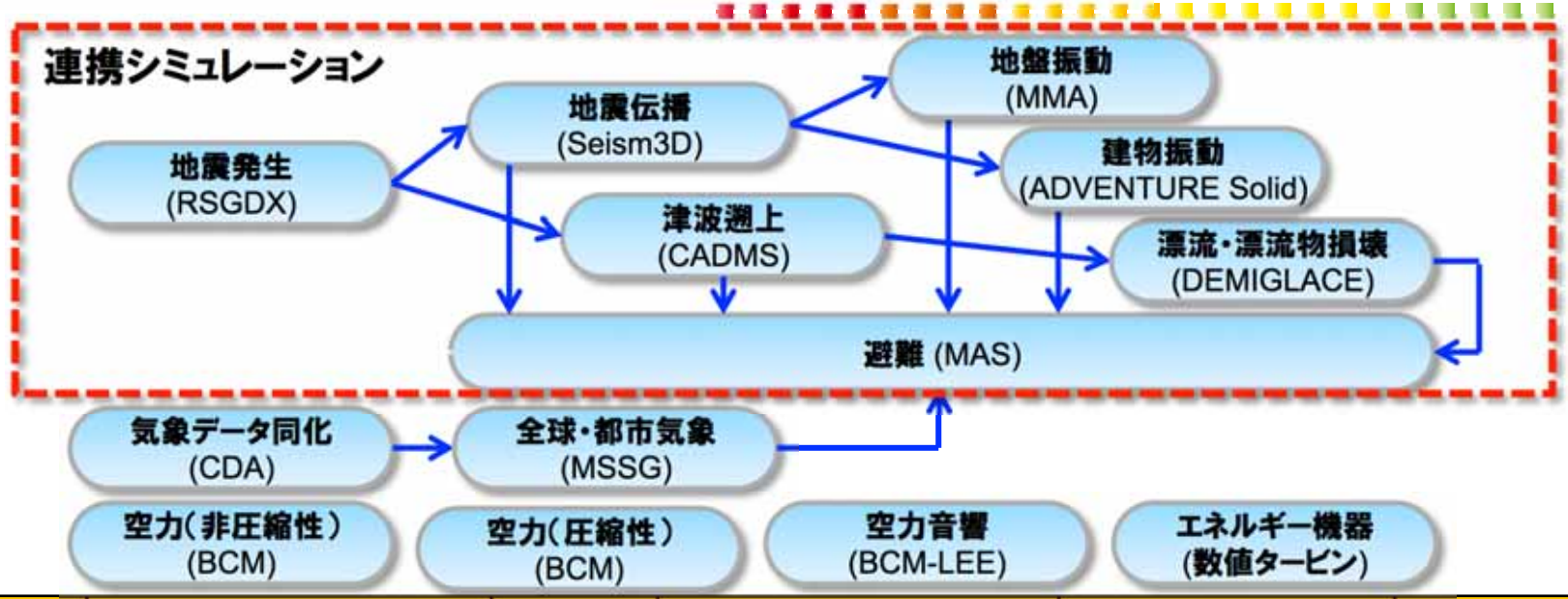
① 高効率タービンの実現
（タービンをまるごと熱流動解析）
② マルチフィジックスCFDの実現
（相変化，腐食，破壊の実現）

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

ターゲットアプリケーションとその計算要求



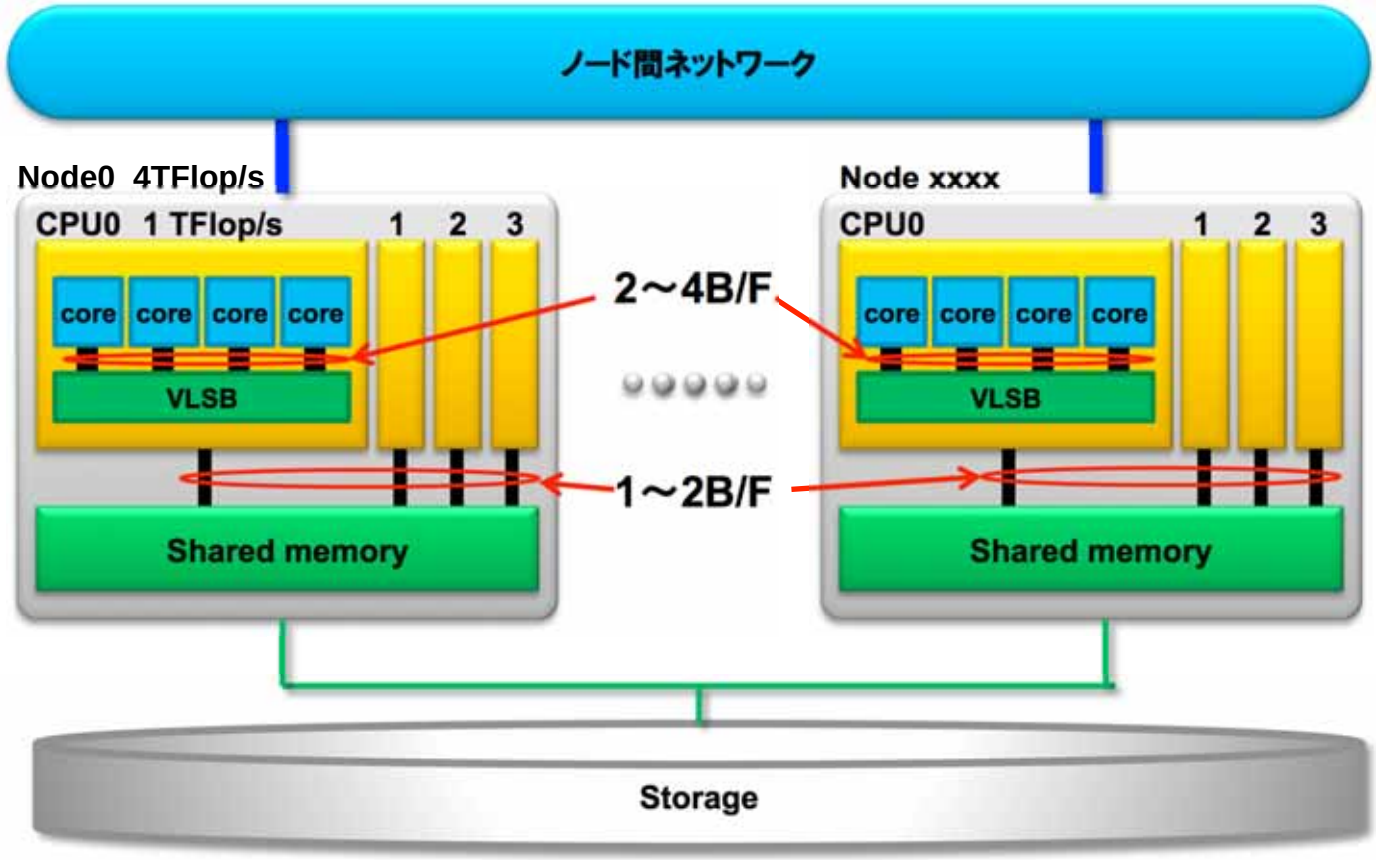
目的	アプリケーション	B/F	必要浮動小数点演算数(×10^{18})	総メモリサイズ	要求実行時間(H)
超高解像度単独計算	RSGDX	4.50	230	1.5 PB	10
	Seism3D	2.14	160	9.6 PB	2
	MSSG	4.00	720	175 TB	6
	BCM	5.47	1	13.6 TB	0.5
高解像度アンサンブル計算	総合防災の連携シミュレーション	2〜5	100	4 PB	2~3
	数値タービン	2.33	140	163.5 TB	20

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、帯域通信も低レイテンシ	**Hi-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



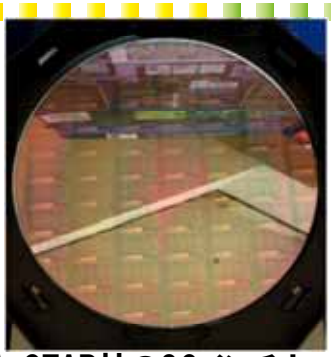
システム規模，メモリ性能，容量，ネットワーク構成はアプリケーション評価と併せて検討

デバイス技術動向

■ 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の加工技術が主流に
    - – Global Foundry 2014年，TSMCは2015年を目標に 14nmプロセスのサービスを提供予定 (@IEDM2012)

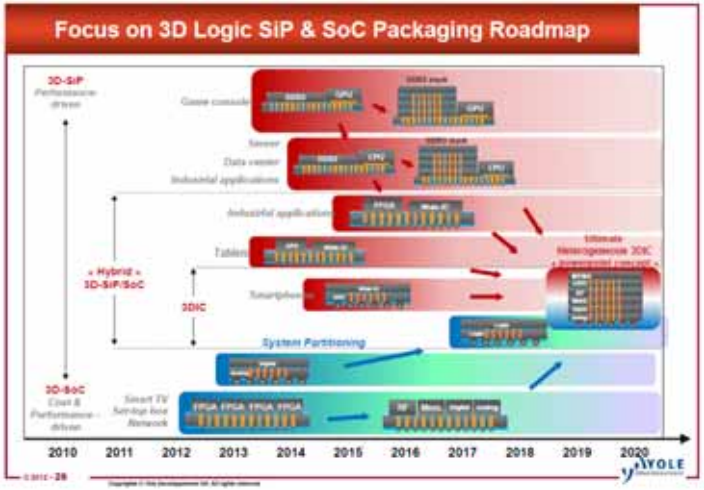
10 – 16nmのプロセスを想定



A・STAR社の30インチウエア
25x40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 mem 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50mm×50mm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

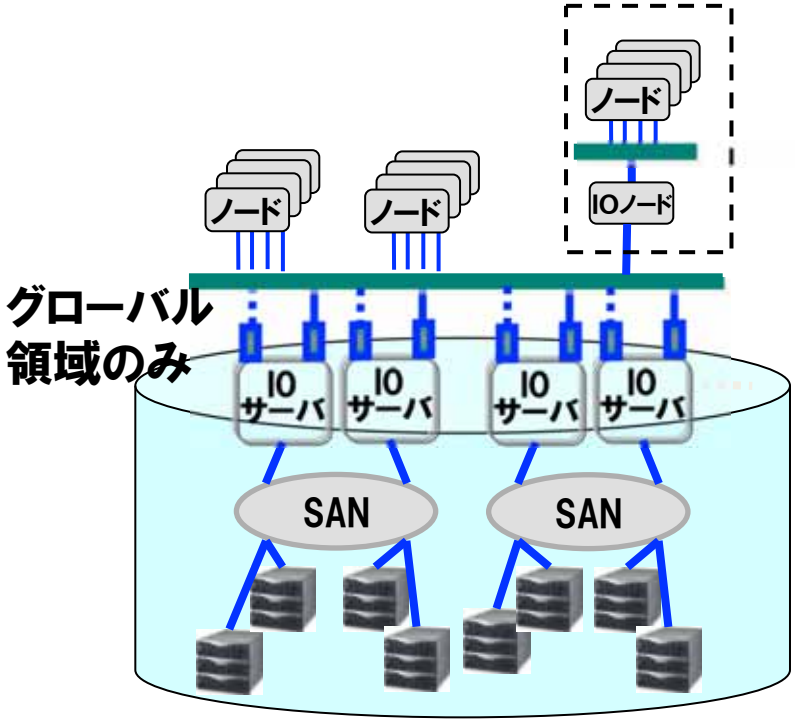
ノード間ネットワーク

	特徴	直径	リンク数	その他
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		○	×	–
Dragonfly	階層構成による疑似High-radix NW	◎	×	–
FTT-Hybrid	Fat treeとTorusの複合型	○	○	ケーブル遅延削減・コスト低減可

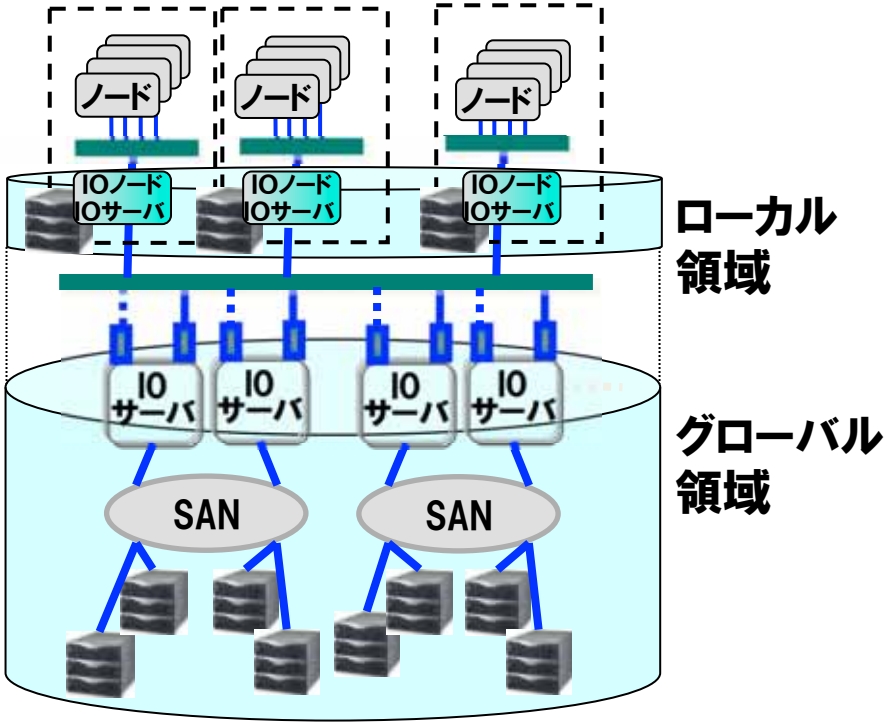
ストレージシステム

- 分散・並列型ファイルシステムは必須（メタデータ，ファイルデータの両方を分散）
- SANストレージのみのフラット型，またはローカル領域とグローバル領域のSANストレージの階層型

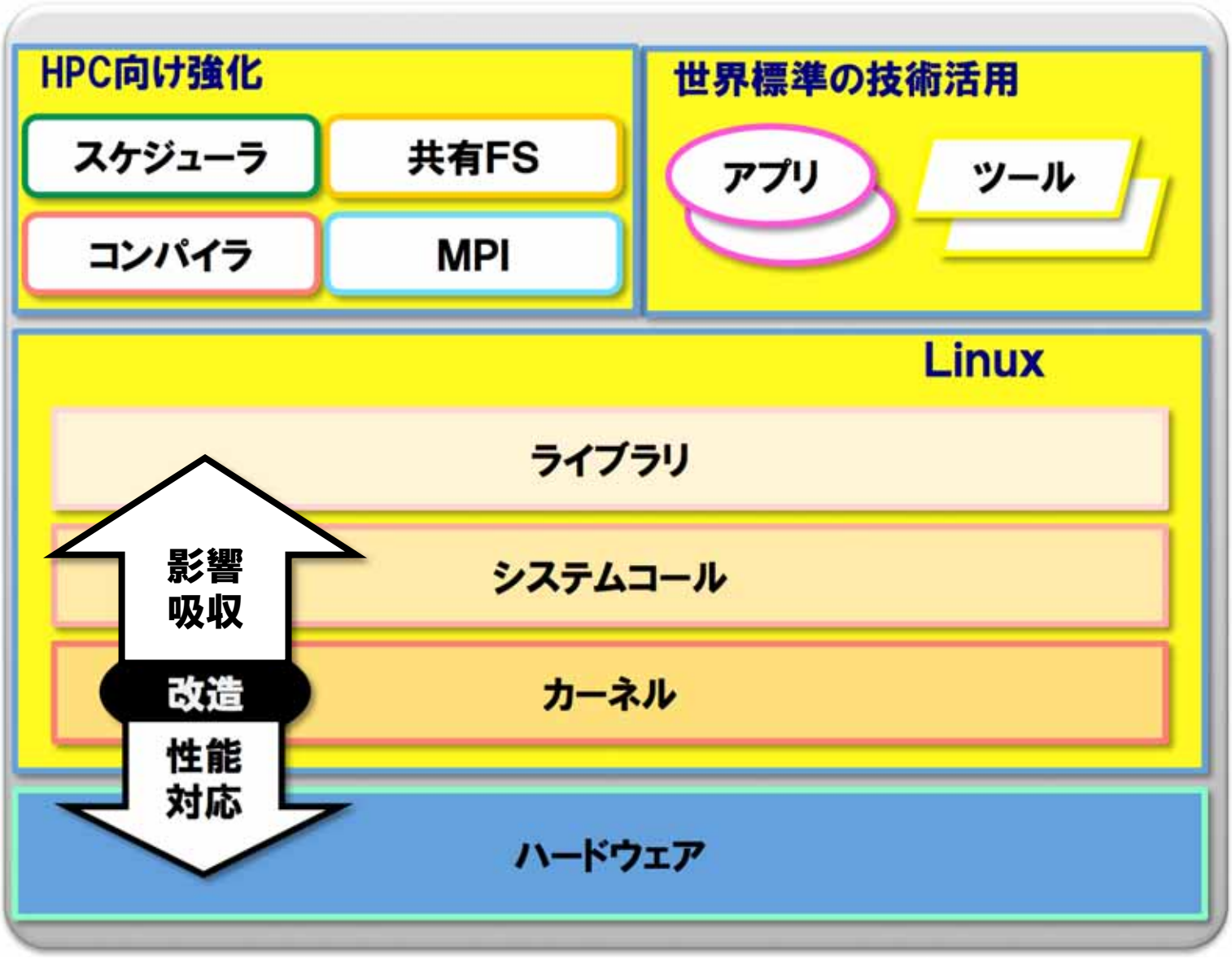
■ フラット型



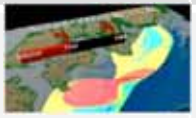
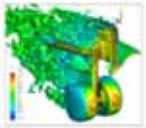
■ 階層型



システムソフトウェア


HPC向け強化
スケジューラ
共有FS
コンパイラ
MPI
世界標準の技術活用
アプリ
ツール
Linux
ライブラリ
システムコール
カーネル
ハードウェア
影響
吸収
改造
性能
対応

性能要求要件とFSシステムでの性能推定結果

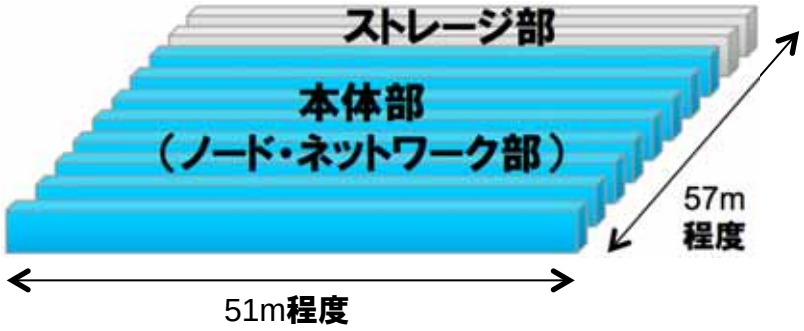
アプリケーション	B/F	必要倍精度浮動小数点演算数（×10^{18}）	要求実行時間	推定結果 ＊	推定CPU数 (Peak Pflop/s)
RSGDX 	4.5	230	10時間	9.1時間	100,000 (100Pflop/s)
Seism3D 	2.14	160	2時間	1.6時間	100,000 (100Pflop/s)
MSSG 	4.00	720	6時間	20.9時間	102,400 (102Pflop/s)
BCM 	5.47	1	0.5時間	0.38時間	65,536 (66Pflop/s)
数値タービン（20ケース同時実行） 	2.33	140	20時間	17.1時間	56,260 (56Pflop/s)

＊ いずれも2B/F, VLSBあり

システム諸元(100PF構成見積)

ピーク演算性能	100PFlop/s	本体系筐体数	400(ノード)+300(NW)程度
総ノード数	25,000	総I/O帯域	10~30TB/s
総CPU数(コア数)	100,000 (400,000)	総ディスク容量	300PB~1EB
総メモリ帯域	100~200PB/s(1~2B/F)	システム設置面積	2,900m²程度
総メモリ容量	3.2~12.8PB	本体系消費電力	25~40MW程度

システム・レイアウト(イメージ)

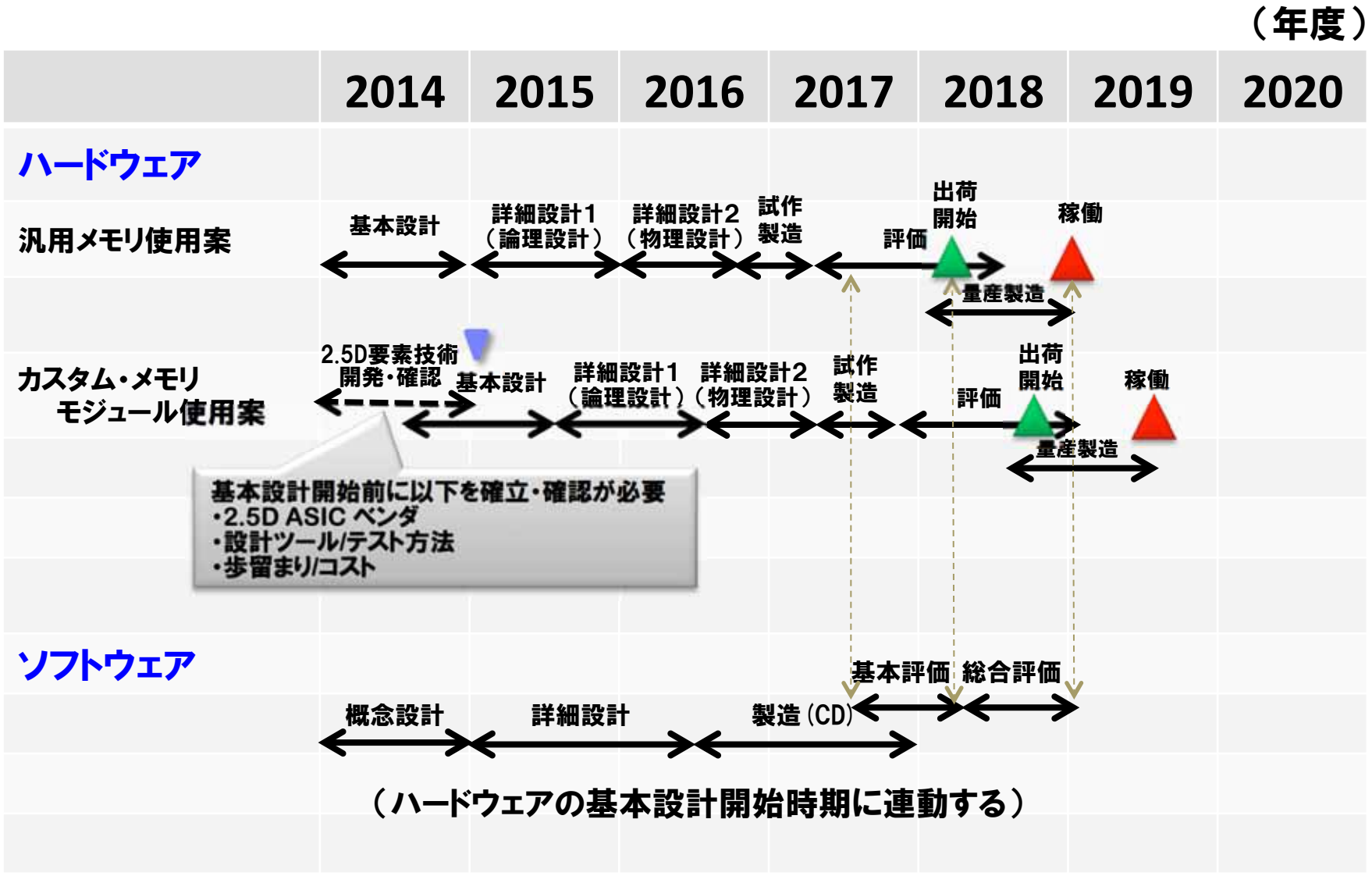


耐故障性・信頼性

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

開発計画


（年度）
2014
2015
2016
2017
2018
2019
2020
ハードウェア
汎用メモリ使用案
基本設計
詳細設計1（論理設計）
詳細設計2（物理設計）
試作製造
評価
出荷開始
稼働
量産製造
カスタム・メモリ
モジュール使用案
2.5D要素技術開発・確認
基本設計
詳細設計1（論理設計）
詳細設計2（物理設計）
試作製造
評価
出荷開始
稼働
量産製造
基本設計開始前に以下を確立・確認が必要
・2.5D ASIC ベンダ
・設計ツール/テスト方法
・歩留まり/コスト
ソフトウェア
概念設計
詳細設計
製造（CD）
基本評価
総合評価
（ハードウェアの基本設計開始時期に連動する）