FUJITSU

将来（1ペタフロップス超・2010年代前半頃を想定）の超高速計算機に必要な要素技術の研究開発について

平成16年8月20日

富士通株式会社

1. 将来の超高速計算機システムについて

FUJITSU

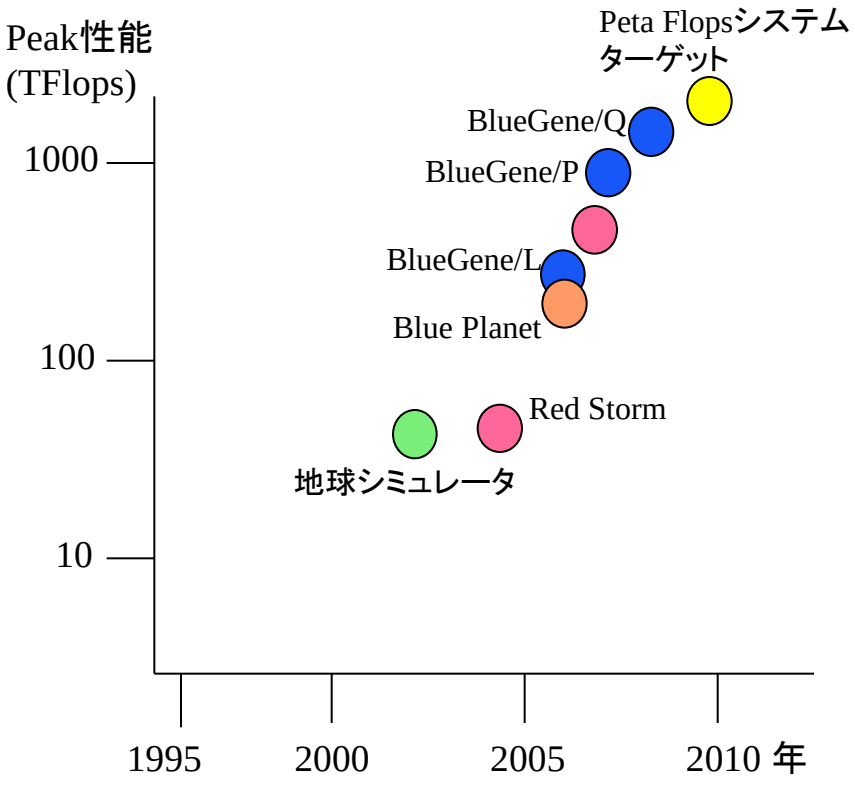
1PFlops超高速計算機（ペタフロップス）システム実現へ向けての挑戦

- 実効1PFlopsを引き出すアーキテクチャの開発
  - 性能＝プロセッサ＋インターコネクト＋ソフトウエア開発環境
  - 実効1PFlopsを実現するには，ピーク性能で3PFlops以上が必要
  - 数千ノード以上の並列実行には，高バンド幅，低遅延のインターコネクトが必要
  - アプリケーションから並列性を引き出すコンパイラ，ソフトウエア開発環境が必要
- 大規模システムの実現を可能とする技術の開発
  - 高密度化：コンパクトなシステム
  - 省電力：プロセッサ/システム全体としての省電力化
  - 半導体技術：微細加工技術，低消費電力技術
  - 高信頼：障害検知，切り離し，交換，自動復旧（ホットスワップ）
  - 運用管理システム：超高速計算機システムを効率よく活用，運用する技術

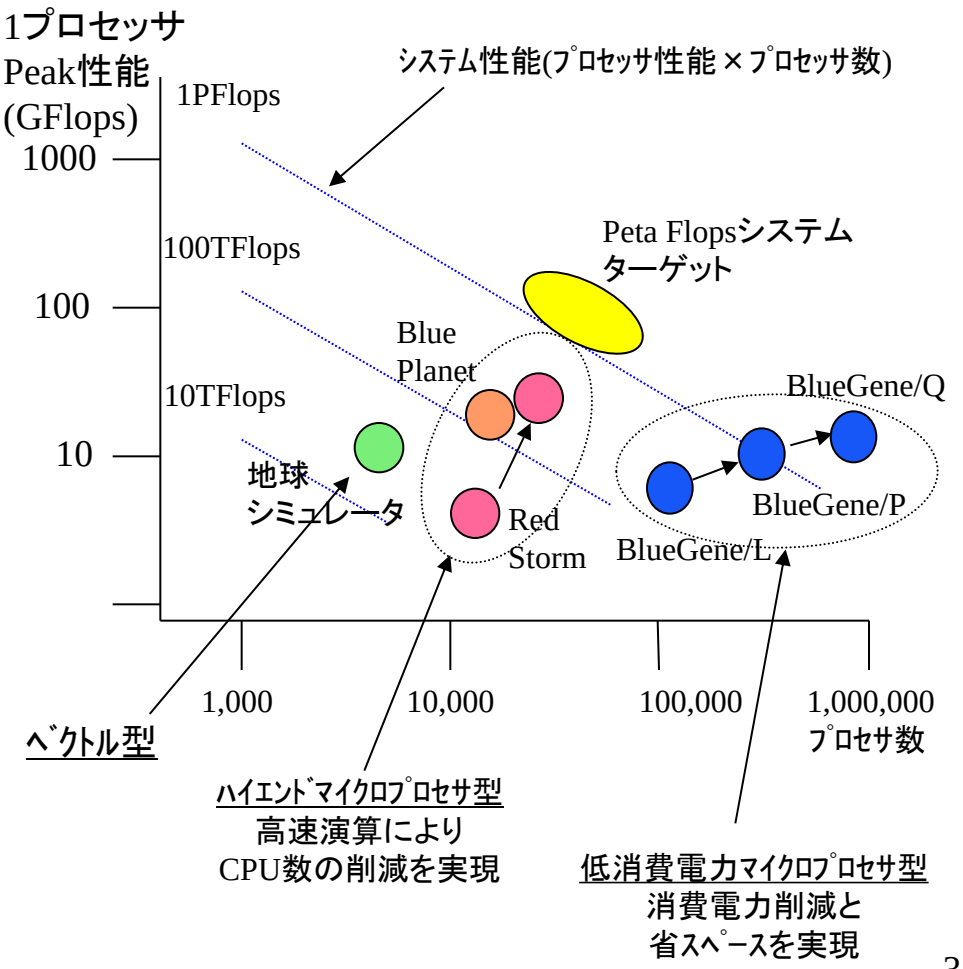
HPCシステムのトレンドと開発ターゲット

FUJITSU

＜HPCシステムの性能トレンド＞



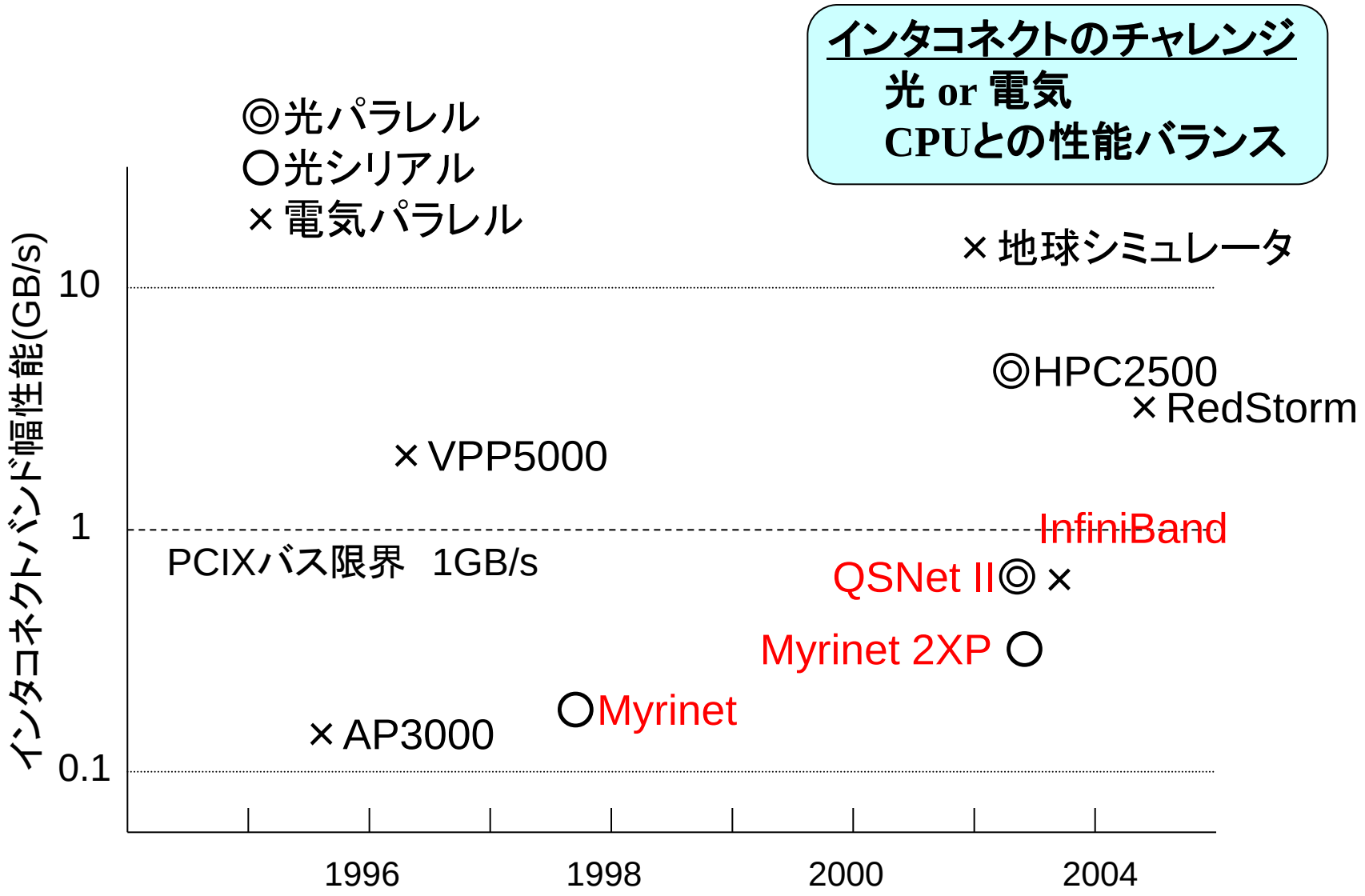
＜プロセサ数 vs. 単体性能＞



2010年のターゲット
ピーク性能3PFlops，実効1PFlops

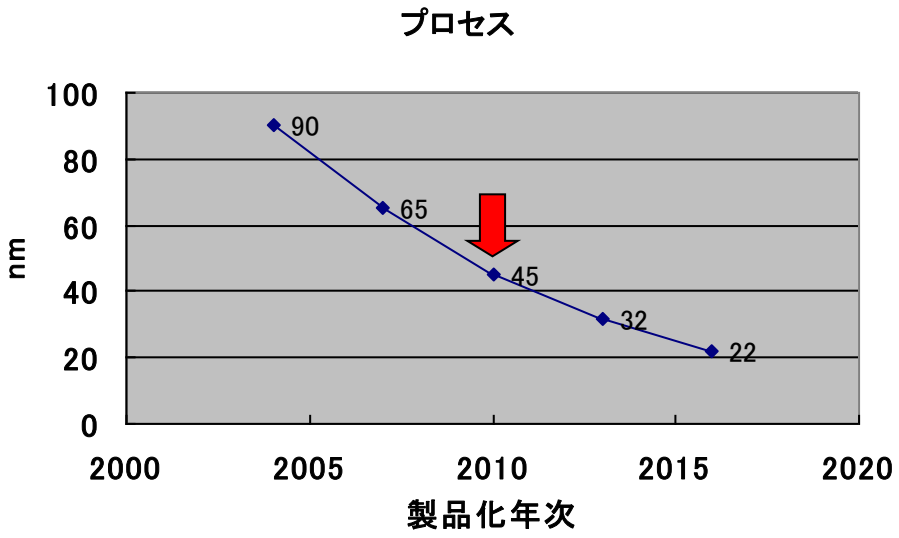
HPC向けインタコネクトの性能トレンド

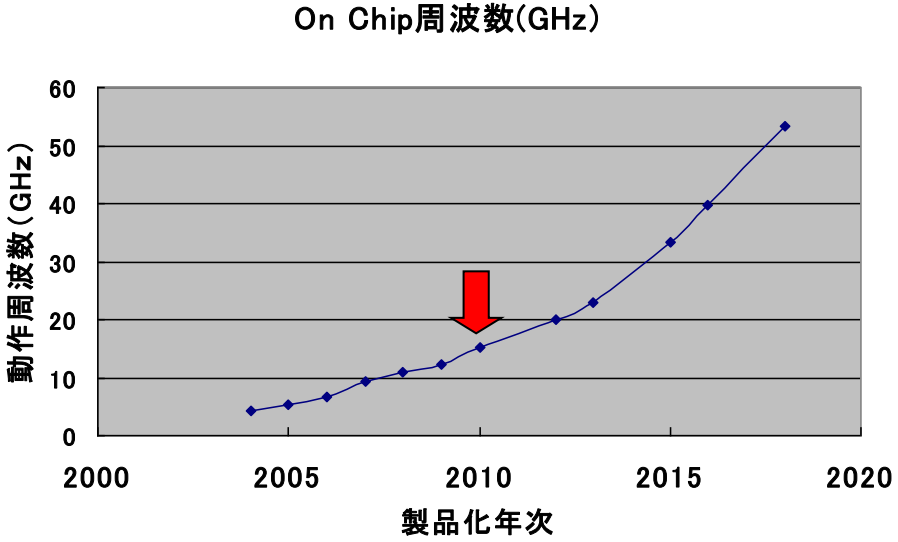
FUJITSU

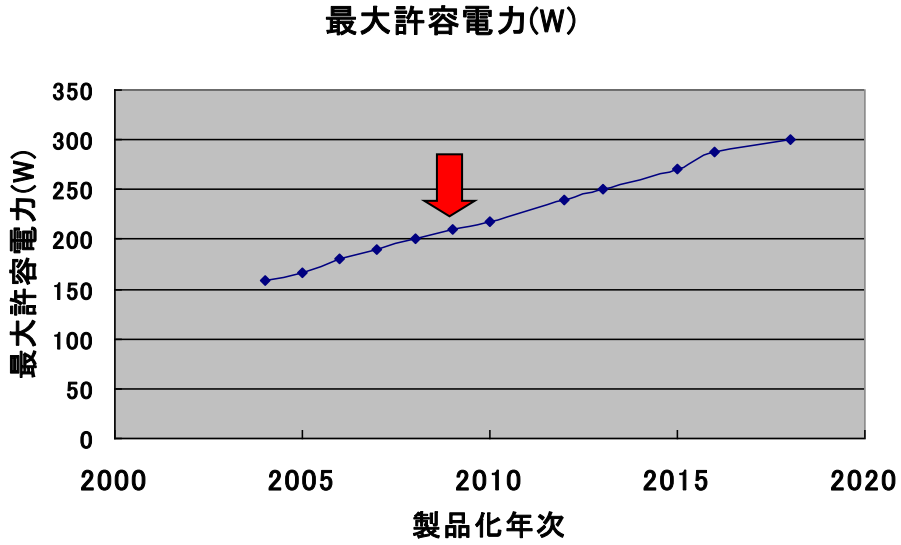


半導体技術のトレンド(ITRS2003資料から)

FUJITSU





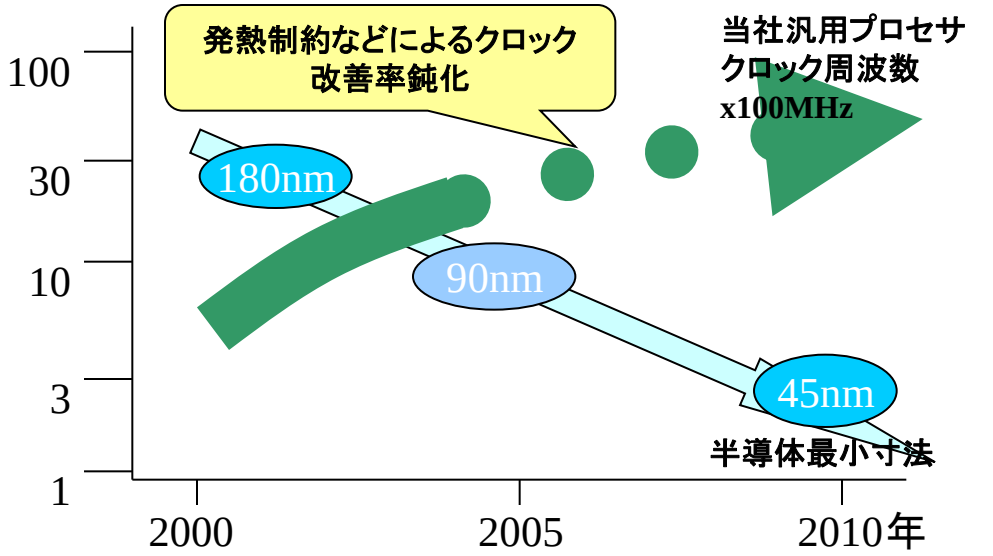


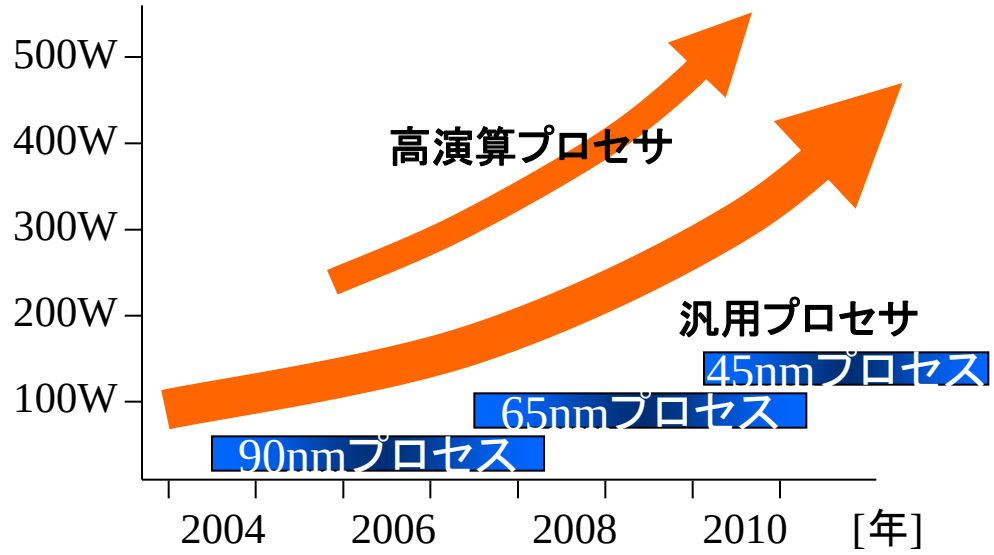
2010年頃
プロセス:45nm
オンチップ周波数:15GHz
最大許容電力:210W

半導体技術と高速化のトレンド

- 微細化は進展するが，以下の問題が顕著化
  - 製造バラツキの増大
  - リーク電流の増大
  - 消費電力の増大

- 要素技術上のチャレンジ
  - 半導体製造技術
    - 低リークゲート絶縁膜
    - 移動度改善（歪シリコン等）
  - 回路設計
    - 製造ばらつき，雑音に強い回路
    - 低電力回路





2. 主なアプリケーションソフトウェアについて

FUJITSU

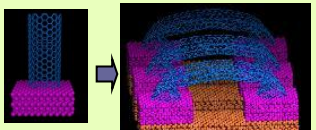
対象分野	アプリケーション	概要	目的	期待されるブレークスルー	経済的波及効果
ナノテクノロジ	Car-Parrinello法第一原理分子動力学	• MOSデバイス絶縁層・半導体・金属層界面の10,000原子超の系の，量子論に基づいた原子拡散，欠陥生成予測を行う。 • CNTデバイス10,000原子超の系の，量子論に基づいたCNT成長，電極接合成長の予測を行う。	• シリコンデバイスの極限微細化設計 • CNTデバイス構造の自己組織化制御	超微細デバイス、CNTデバイスの実現	我が国の半導体技術、しいてはIT技術が国際的に先行でき、高い競争力につながる。
バイオ・インフォマティクス	分子軌道法第一原理分子動力学（or 半古典MD，溶媒ﾎﾟﾃﾝｼｬﾙﾓﾃﾞﾙとの連成解析）	• タンパク質や水，他分子の数千～数万原子の系の量子論に基づいた構造解析や他分子との相互作用の予測を行う。	• タンパク質構造 • 動的機能予測	創薬候補の計算機ｽｸﾘｰﾆﾝｸﾞ	我が国の創薬・医療技術の高度化と米国寡占からの脱却
気候予測	大気・海洋・海氷結合モデル、他	• 水平解像度1KM（現在～100KM）で大気，海洋，海氷の振舞いを100年レンジでシミュレーション。 • 大気中化学反応、海中生物化学過程を含む。	• 大気変動、水循環、大気汚染の影響の予測	長期気候変動の予測と社会経済動向の予測	エネルギー需要、環境コストの長期的予測につながる
CAE	衝突解析、流体解析、及びこれらの連成解析	• 数千万ﾒｯｼｭﾚﾍﾞﾙ大規模構造解析、数百万ﾒｯｼｭﾚﾍﾞﾙの衝突解析の大量計算 • 現行ﾒｯｼｭｻｲｽﾞの1000倍ｸﾗｽの大規模ﾒｯｼｭによる精密又は大規模計算 • 高度な流体、構造連成計算	• 自動車、航空機、電子機器等の設計	最適化設計による開発期間短縮、ｺｽﾄ削減	基幹の産業競争力確保

ペタスケールコンピューティングの拓く世界

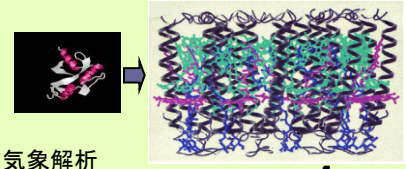
FUJITSU

大規模処理 計算規模の拡大・計算精度の向上

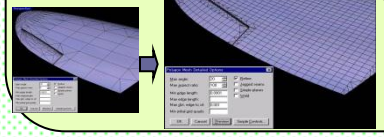
材料解析
1000倍以上の計算規模拡大による実際の設計に直結するシミュレーションへ



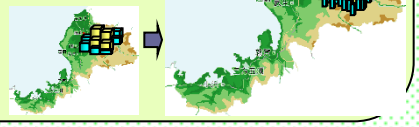
タンパク解析
厳密解法による精度向上と解析規模拡大により現実に近い複雑な系の挙動解析を実現。



構造・衝突解析
1000倍以上の計算メッシュ拡大による解析精度向上。より正確かつ安全な設計へ



気象解析
解析メッシュの微細化により、予想精度の向上・より局地的な現象の予測が可能に。

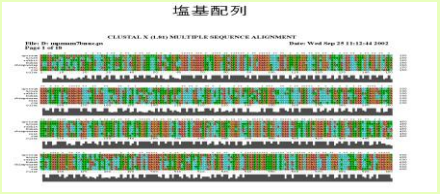


大規模処理 Scale up
複雑処理 Super-Scale
大量処理 Scale out

次元を越える
視点が変わる
見えないものが見える
価値観が変わる
：

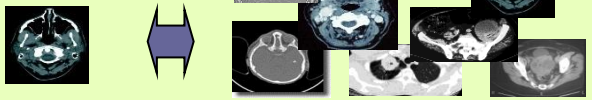
大量処理 大量データの検索

遺伝情報検索
大量の遺伝子情報の中から目的遺伝子との類似情報を超高速に検索。未知ウイルスの探索、疾病診断などに活用。



暗号解析
コンピュータウイルス検出
大量のインターネット情報からのコンピュータウイルス検索、暗号解析を高速に行い、インターネットのセキュリティを確保

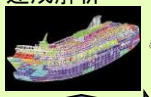
多次元類似度検索
大量の4次元画像からの類似画像を高速に検索。医療診断等に活用

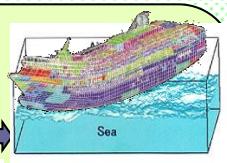


複雑処理 マルチスケールコンピューティング・マルチフィジクスコンピューティング

DNAから人体までの一貫した解析
DNAから発現するタンパク、細胞内のタンパクの挙動、それらが構成する器官、さらに器官から構成される人体までをモデル化し薬剤の効果、遺伝疾患を遺伝レベルで対応可能

構造流体連成解析
船の構造解析

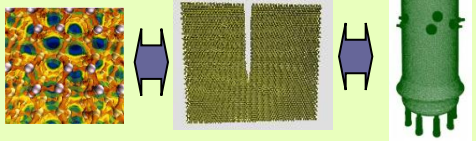




波の解析

関連する全ての要素を計算する動的解析によるより安全な、より効率的な製品作りへ

原子レベルからの破壊の一貫した解析
原子レベルの結合破断から目に見える破壊までを解析し、原理の理解、材料の改良へ。



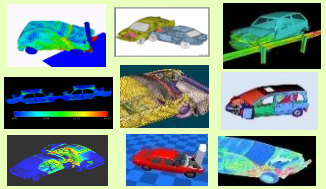
生体の動的解析
血管の構造解析

流れ解析

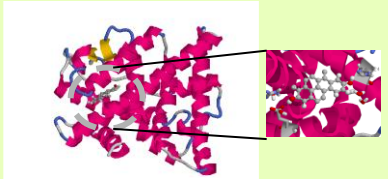
流れによる影響を考慮した血管の動的な解析により、より安全な診断・手術へ

大量処理 処理回数の増大

衝突解析
パラメータを変化させた大量解析による、計算モデルの検証、コード検証統計解析を行い、より安全かつ効率的な製品作りへ



タンパク解析
目的タンパクと多種類の薬候補化合物とのドッキング計算を高速に実行し、候補薬剤を絞り込み、疾患対策を迅速に。



安全・安心な社会の実現に向けて

医療・健康	防災	セキュリティ	物作りの革新	
創薬支援 診断・治療支援 未知疾病対策支援 ：	災害シミュレーション 災害予知・予測 災害避難支援 ：	暗号処理 コンピュータウィルス検出 金融工学 ：	新材料設計・開発支援 製品設計・開発支援 ：	・・・・・

３－１．必要な要素技術について(ハードウエア) FUJITSU

- プロセサアーキテクチャ
  - 実効1PFlopsを実現する演算器，キャッ
    - 数百GFlops/チップの実現方式
  - 演算エネルギー消費の低減
    - 現状の延長で性能を追求すると500W/chip → 200W以下
  - 高信頼度の実現（演算エラー等の検出）
    - 従来の汎用マイクロプロセサのチェッ
    - 20,000CPUシステムでは，500時間に1回
    - 演算器二重化は電力，チップ面積の負担大 →エネルギー効率の高い検出手法の開発
- システムアーキテクチャ
  - 対象アプリケーション群に対して実効1PFlopsを実現できるシステムアーキテクチャ
    - 現状では数百～数千プロセサが限界 → 数万～数十万プロセサを効率的に動かす技術
    - 分散メモリ＋インタコネクト，階層メモリ構造（バンド幅，レーテンシ，メモリ量など）の最適化
  - 高可用性（エラーリカバリ）の実現方式
    - ホットスワップの実現

SIMD型CPU vs Scalar型CPU

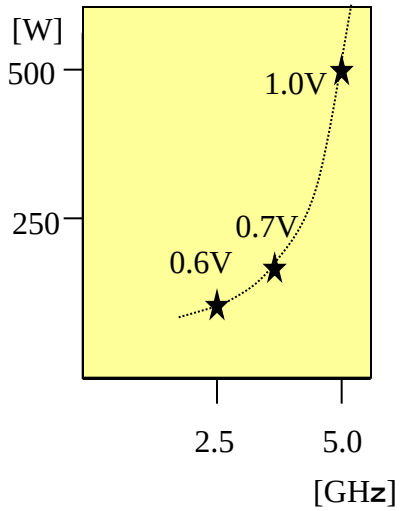
高信頼化技術

両方式の比較

FUJITSU

	汎用Scalar型CPU	SIMD型CPU
ノード	2CPU+Memory	1CPU+L3$+Memory
CPUチップ	5GHz 通常電圧 160GFlops ～200W	3GHz 低電圧駆動 768GFlops ～200W
メモリBW	128GB/s/CPU	400GB/s/CPU L3$ 102GB/s/CPU
システム	10240ノード ～8MW	4096ノード ～1.6MW
アプリケーション	スカラ並列の高度化	スカラ並列の高度化 ＋SIMD型CPU向きアルゴリズム

低電圧駆動の効果
(SIMD型CPUの例)



青字：チャレンジ項目

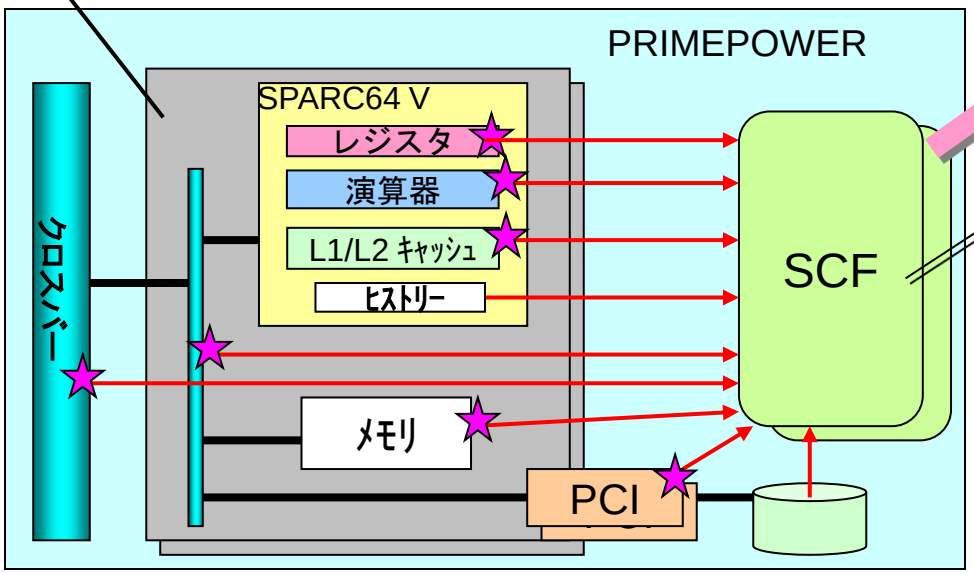
高信頼化技術：高精度な予兆監視・エラー原因特定

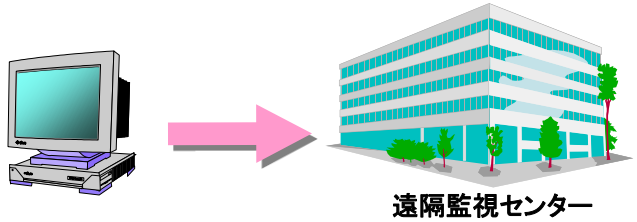
FUJITSU

～ 万一のエラー発生時にも迅速に対処・復旧 ～

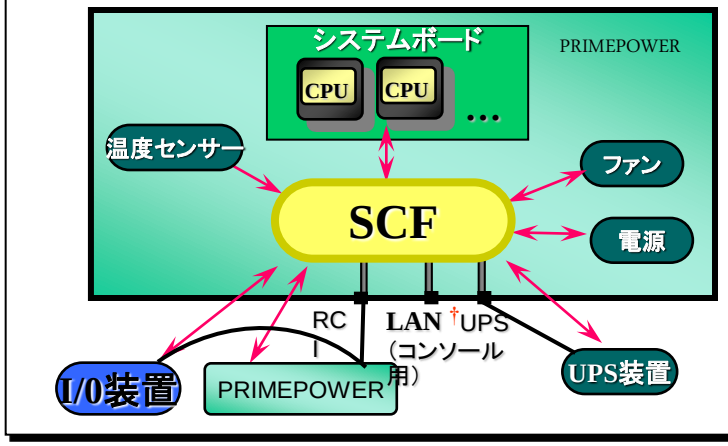
- エラー検出回路（チェッカー）を随所に内蔵
- 障害部品の特定、ヒストリデータ等による解析
- e-mailにて遠隔監視センターへ通知

➢チェッカー
- SPARC64™ V プロセッサ当たり *679* 個
- システム当たり *90,000* 個
（PRIMEPOWER2500/32CPUの場合）





SCF (System Control Facility) :
専用のプロセッサを搭載し、サーバ・ハードウェアの初期診断／監視／異常通知、および電連投入の制御などを行なうH/W。



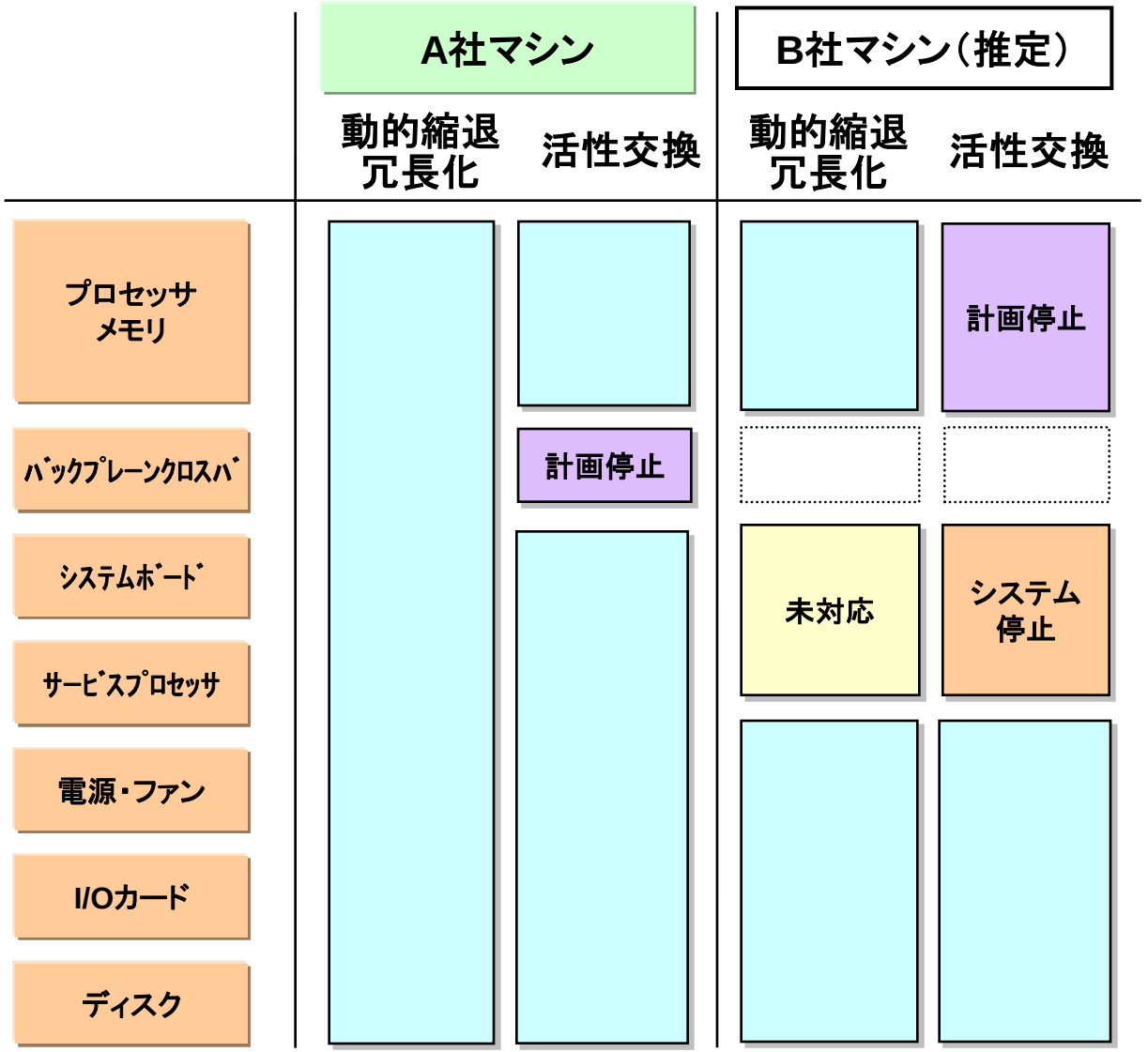
RCI (Remote Cabinet Interface): ETERNUS等との電源連動、クラスタでのノード監視
UPS: 無停電電源装置との信号ケーブル接続
†: PRIMEPOWER250/450 に標準装備するXSCF (eXtended System Control Facility)でサポート

高信頼化技術：連続運転へのこだわり

FUJITSU

当社の連続運転技術例

- 自己診断による予兆監視機能
- 局所的電源切断機能
- 故障箇所の自動切離し機能



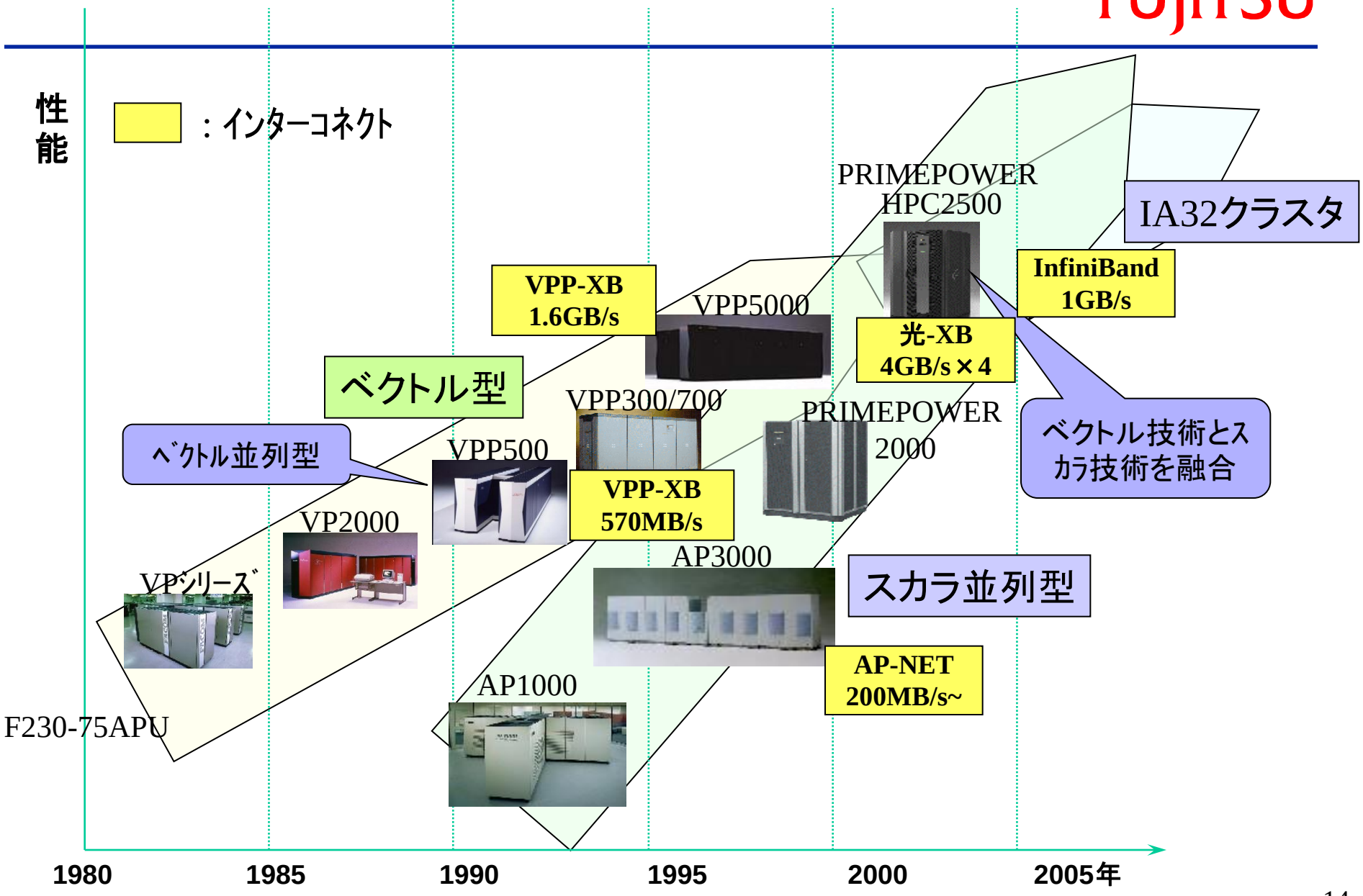
必要な要素技術について(ハードウエア) (2)

FUJITSU

- 低電力論理回路
  - 200W/チップで実現可能なロジックの開発と検討
    - 低電圧動作と動作マージンを確保する回路設計技術
- 高速伝送回路
  - プロセサチップとメモリ，インタコネクト間の高速信号伝送
  - 現状3～6Gbps，20～100mW/Gbit/s → 6～12Gbit/s，10mW/Gbit/s以下の消費電力
- 高速ノード間インタコネクト
  - 最適トポロジー，必要性能の検討
  - アプリケーションが要求するデータ転送パタンを低オーバヘッドで実現するホストインタフェース
  - スイッチチップ（あるいはプロセサ内蔵のスイッチ機能）の検討
    - 要求性能を実現するアーキテクチャ，ハードウェア規模，消費電力等の検討
    - 現状，10Gbps x 12port → 20Gbps x 16port のインテリジェントスイッチ
  - インタコネクトを実現する素材の検討

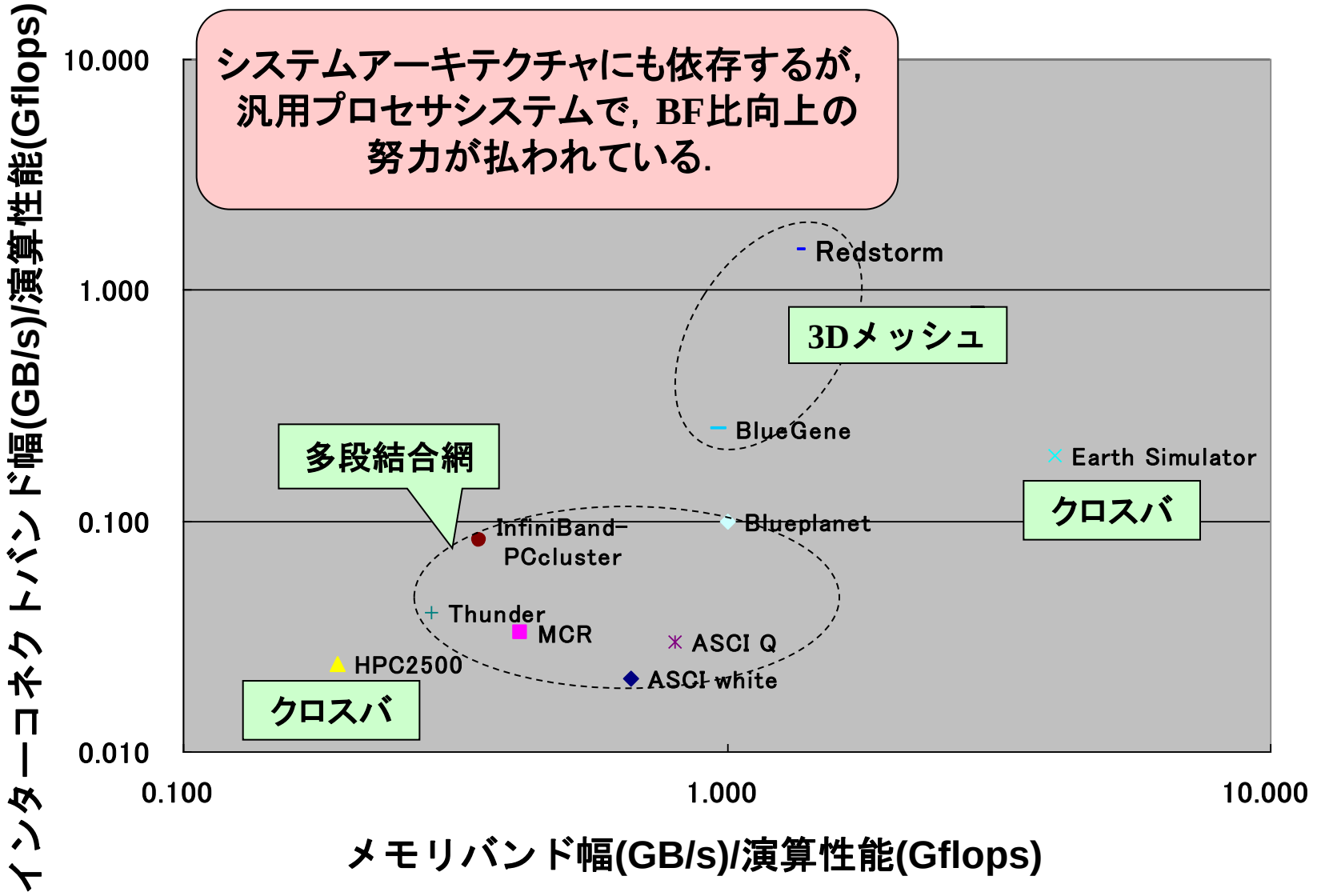
富士通のHPCプラットフォームとインターコネクト

FUJITSU



高速インタコネクト性能バランス

FUJITSU



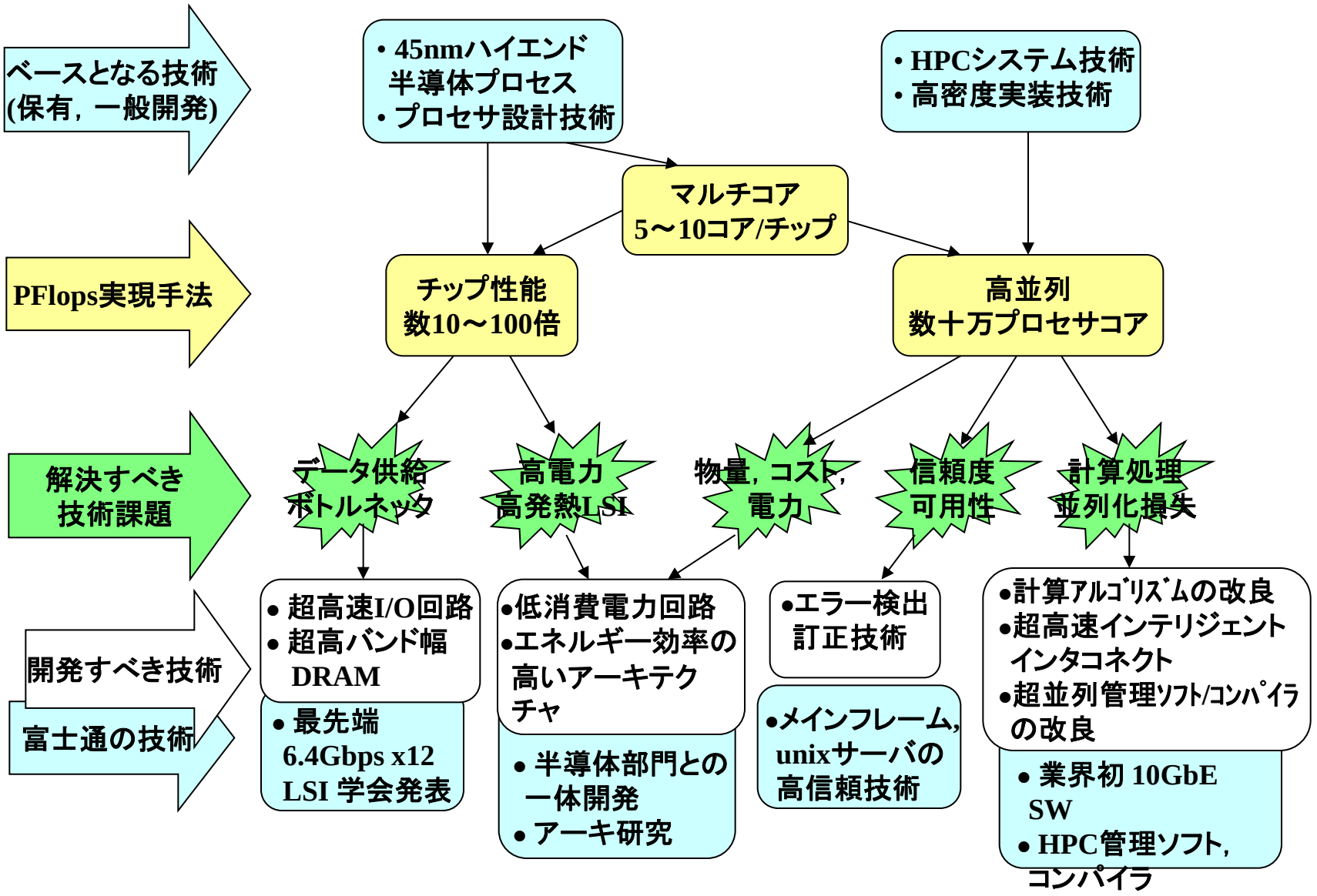
３－２ 必要な要素技術について(基本ソフトウェア)

FUJITSU

- OS・ミドルウェア
  - ～数万ノードの効率的運用管理機能の検討
    資源管理，ジョブスケジューリング，操作・監視，稼動情報など
  - 大容量高速分散並列ファイルの検討
    ローカルファイル，ネットワークファイルなど
  - 超並列向けカーネル拡張の検討
- コンパイラ
  - ベクトル並みの実効性能を実現する自動並列コンパイラとハードアシスト機能の検討
  - ～数万ノードの並列最適化と開発環境の検討
  - 超並列ライブラリ（MPI、数学ライブラリなど）の検討
- 効率的な超並列実行の実現
  - 15万～50万演算器に処理を分割する方法は自明ではない.
  - バッファ，キャッシュ容量，メモリバンド幅，インタコネクトバンド幅等を考慮して，処理の並列化を支援するツールが必要
  - 実システムでのチューニングを可能にする豊富な動作状態情報の提供

必要な要素技術のまとめ

FUJITSU



３－３．必要な要素技術について（アプリケーション開発）
― 利用技術の研究開発と普及 ―

FUJITSU

超PFlops・シミュレーションの実現には，ハードウェア・ミドルウェアの研究開発のみならず，既存手法の延長線上にない新しいシミュレーション手法・アルゴリズム等の研究開発が重要．

- 超大規模シミュレーションを実現するアルゴリズム，最適化設計問題・連成解析などの先端シミュレーション手法の研究開発
- 超高並列処理技術・アルゴリズムの研究開発

シミュレーション技術の発展には、理論研究者，利用技術開発者等が連携して，統合的にシミュレーション技術に関する研究開発を目的として，これを実施する環境・仕組みの整備が必要．

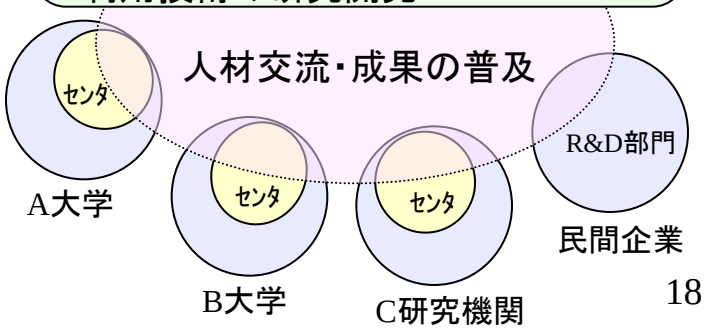
(米国では，ソフト開発者，利用者，理論研究者が連携してｼﾐｭﾚｰｼｮﾝの研究開発を実施)

シミュレーション技術センターの設立

- 産学官連携体制による計算科学に関するCOEを設立し，人材育成，アプリケーションに関する研究開発を実施．成果を全国に普及．
- 必要に応じて，各機関に人材を派遣．

シミュレーション技術センター

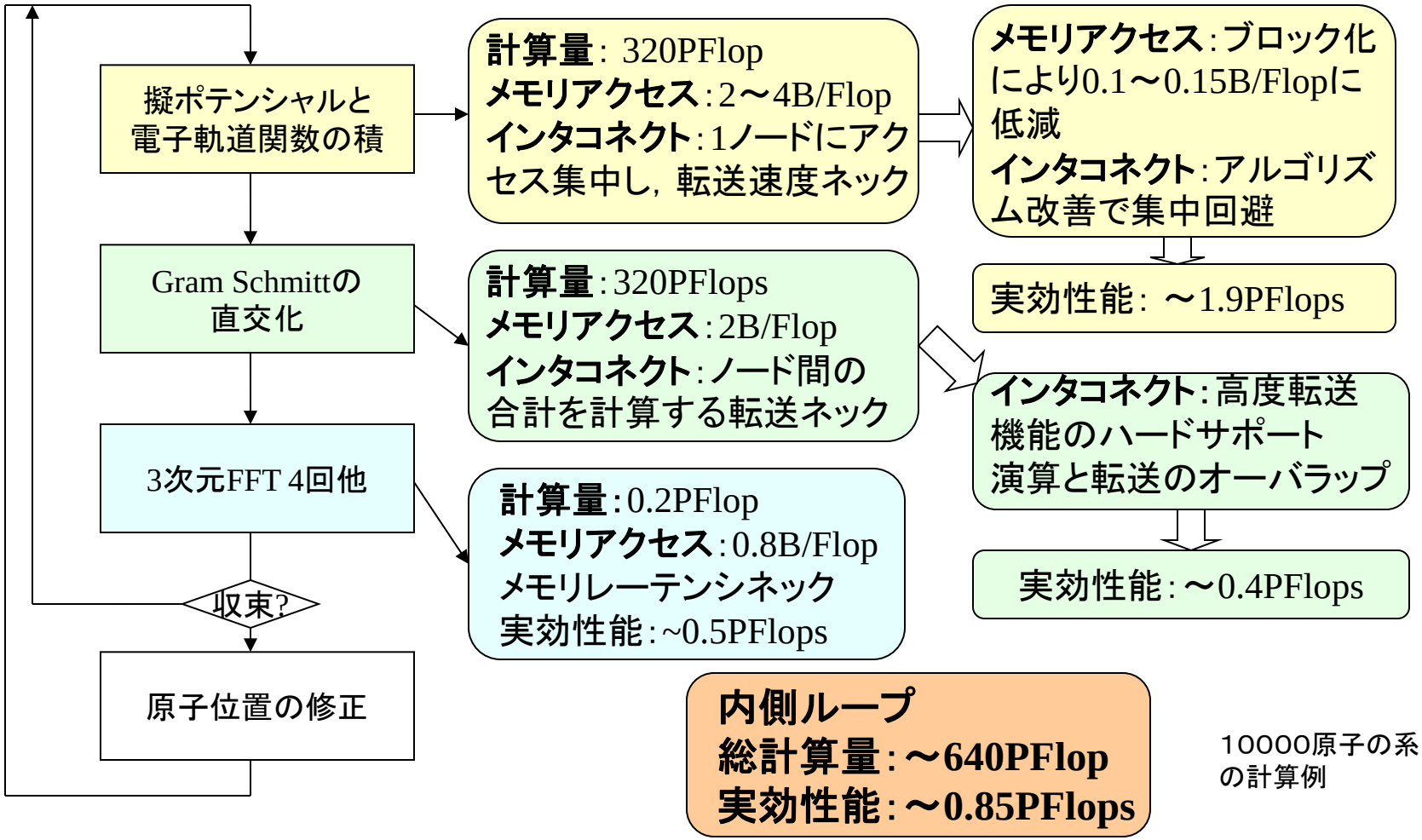
・理論研究者を交えたｼﾐｭﾚｰｼｮﾝ技術の研究開発
・利用技術の研究開発



SIMD高演算システムによるCar-Parrinello法の処理分析例

FUJITSU

シリコン等の固体の原子レベルの第一原理シミュレーション法

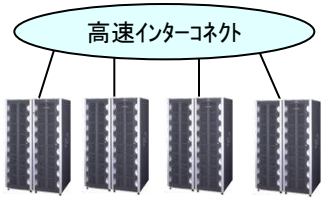


10000原子の系の計算例

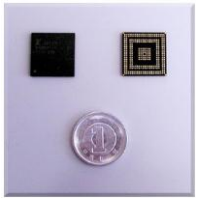
実現する技術の波及分野

FUJITSU

- クラスタ
  - PFlopsシステムより小規模であるが，PFlopsシステムで開発したシステムアーキテクチャ，インタコネクト，並列実行技術が有効に利用できる．



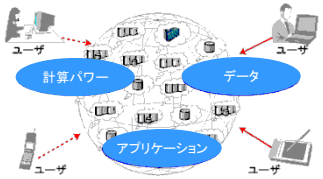
- メディア処理LSI等
  - 画像，ビデオ処理，音声認識等の分野では大量の演算を低電力で実行する必要があり，演算エネルギーの低減技術は汎用的に利用可能である．また，メモリとの間の高速伝送技術も適用可能である．



- LAN，機器間接続（次世代インタコネクト等）
  - 現状は2.5~3Gbpsの伝送であるが，低電力の6～12Gbps伝送技術は，これらの性能向上を可能にする．

- 組込機器等
  - 半導体の低消費電力技術は組込機器への応用が可能である．

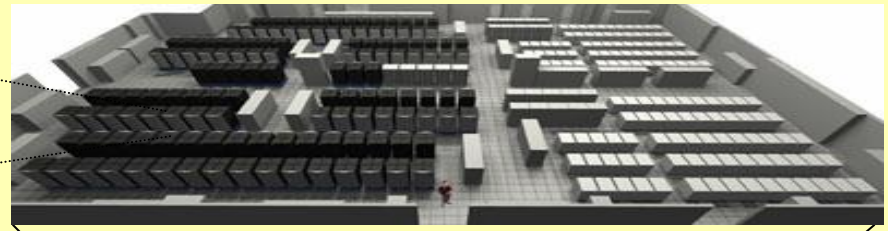
- グリッドコンピューティング
  - 超高速計算機の運用管理システムは,グリッドコンピューティングへの展開が可能．



ターゲットシステムの実現イメージ

FUJITSU

現存のスーパーコンピュータシステム



現存スーパーコンピュータの
一つのロッカー並の性能

現存スーパーコンピュータ
システム全体並の性能

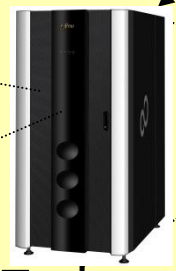
Peta Flopsシステム

プロセサチップ
数百GFlops

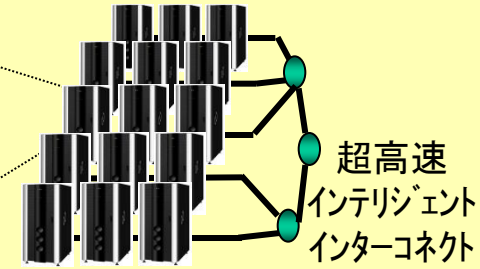
単体プロセサ
8個程度を
1チップに集積

ボード
1~4プロセサ
0.5～1TFlops

DRAM



ロッカー
数十ボード
10～数10TFlops



システム
100～500ロッカー
数PFlops

4－1．要素技術(ハードウェア)の研究開発について

FUJITSU

○研究開発実施体制

大学、研究機関

○スケジュールおよび費用

要素技術の実現：20億円

超高速計算機の実現：460億円

要素技術	初年度	第2年度	第3年度	第4年度	第5年度	第6年度
システムアーキテクチャ，およびプロセサアーキテクチャ	方式検討：1億円	シミュレータ開発：2億円	シミュレーションによる実証：1億円 商用システム方式設計：15億円	商用システム設計：120億円		チップ，システム製造，試験：250億円
低電力論理回路，および高速伝送回路	回路設計：2億円	チップ試作(65nm)：6億円	評価：1億円			
高速ノード間インタコネクト	方式検討：1億円	論理設計：1.5億円 FPGA実装：1.5億円	商用方式設計：5億円 実証：3億円	商用インタコネクト設計：40億円		製造，試験：80億円

４－２．要素技術（ソフトウェア）の研究開発について

FUJITSU

○研究開発実施体制
　　大学、研究機関
○スケジュールおよび費用

要素技術の実現：10億円

超高速計算機の実現：84億円

要素技術	初年度	第２年度	第３年度	第４年度	第５年度	第６年度
OS・ﾐﾄﾞﾙｳｪｱ ・運用管理機能 ・高速分散並列 ファイル	方式検討 ：1.5億円	プロトタイプ開発 ：3億円	プロトタイプ実証 ：3億円 商用方式設計 ：8億円	商用システム開発：32億円		商用システム実証 ：16億円
コンパイラ ・自動並列 ・最適化/ 開発環境 ・ライブラリ	方式検討 ：0.5億円	プロトタイプ開発 ：1億円	プロトタイプ実証 ：1億円 商用方式設計 ：4億円	商用システム開発：16億円		商用システム実証 ：8億円

4－3. 要素技術（アプリケーション）の研究開発について

FUJITSU

○研究開発実施体制

大学、研究所、民間企業が積極的に連携

○スケジュールおよび費用

要素技術の実現：9億円

超高速計算の実現：27億円

研究基盤整備と運営：150億円

要素技術	初年度	第2年度	第3年度	第4年度	第5年度	第6年度
超大規模・超高並列シミュレーション技術開発	アルゴリズム・手法研究開発：1億円/年					
		超並列アプリプロトタイプ開発：3億円/年				
				実用シミュレーション・ソフトの開発：9億円/年		
研究基盤整備技術の普及	基盤整備 10億円		基盤整備 40億円			
		産官学連携シミュレーション技術センターの運営　：20億円/年				

まとめ

FUJITSU

- プロセッサ
  - コスト性能比の良いアーキテクチャの採用
    - SIMD型 vs 汎用Scalar型
- インタコネクト
  - 光 vs 電気
  - 性能バランス（メモリスループット：BF比）の改善
- 半導体
  - 微細化に伴う課題の解決：製造ばらつき，リーク電流，消費電力増大
- ソフトウエア
  - 開発環境の整備：コンパイラ，性能モニタ，etc
- 運用管理
  - 高信頼の追及
  - 使いやすさの追求
- アプリケーション
  - 適用分野の拡大：Compute-intensive から探索，マイニング等へ