秘密情報，回収資料

資料1

ベクトル部の成果利用及びシステム構成変更による影響について

2009年6月11日

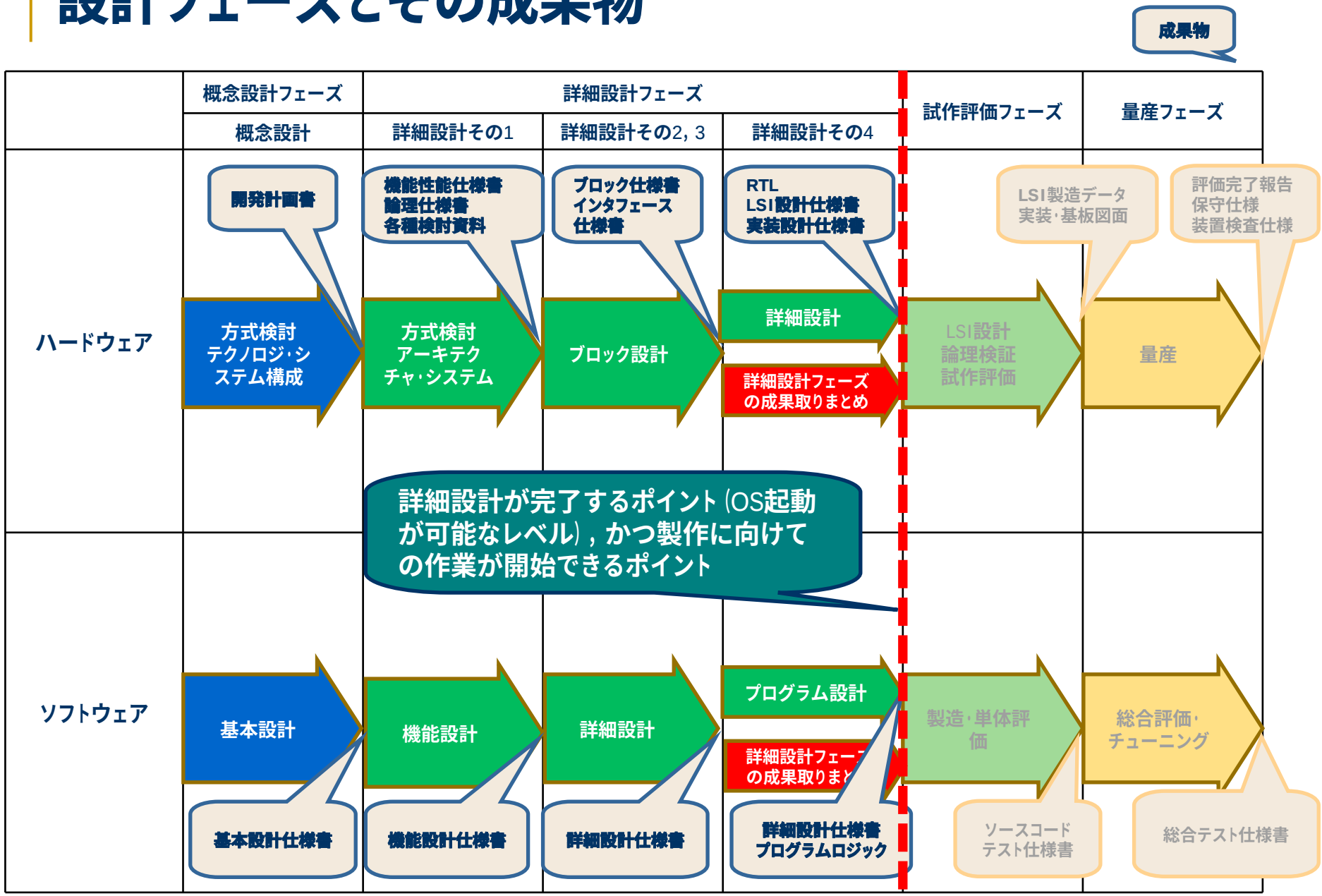
理化学研究所
次世代スーパーコンピュータ開発実施本部

説明内容

- ベクトル部詳細設計成果の取扱いについて
- スカラ部単独システムへの構成変更に伴うベクトル機ユーザへの影響について

ベクトル部詳細設計成果の取扱いについて

設計フェーズとその成果物



成果物の例

ソフトウェア仕様

OS

メモリ管理

(1) Int ptmemall32

形式	void ptmemall32(reg_t *rp, pas_t *base, int offset, int size, int validate, int nosleep, int listsz, int plmnum);	
引数	*rp*	Region 構造体へのポインタ
	base	ページテーブルへのポインタ
	offset	ページテーブル内オフセット
	size	要求ページ数
	validate	ページを有効にするかどうかのフラグ
	nosleep	空きページがない場合に sleep して(swap して)ページを確保するかどうかのフラグ
	listsz	ページテーブルのエントリ数
	plmnum	LM 番号(物理)
処理概要	スモールページを size で指定されたページ数だけ割り当てます。 ● plmnum で指定された LM のページを割り当てます。	

ファイルシステム

I/Ogateway

関数概要	関数名	ソースファイル名
	main (I/Ogateway)	iogw_thread.c

関数形式

[形式] int main(int argc, char *argv[])

[引数] 使用しない

機能

本関数は、I/Ogateway の入り口となる main 関数であり、各種起動処理関数の呼び出しと各種資源の初期化を実行します。

捕捉するシグナルのシグナルハンドラを設定し、捕捉しないシグナルすべてをブロックします。

SIGTERM シグナルを受け付けると、シグナルハンドラである I/Ogateway 停止処理関数(iogw_stop_iogateway())を起動します。

下記の関数を呼び出します。

- config ファイル読み込み関数(iogw_get_config())
- I/Ogateway 起動処理関数(iogw_startup_iogateway())
- verb インタフェース起動処理関数(iogw_startup_verb())
- 実行中処理管理リストとオープンファイル管理リストを初期化する関数(iogw_init_hash_list())
- レコードロックファイル管理リストとブロック済みレコードロック要求リストを初期化する関数(iogw_init_lkhash_list())
- 受信済みキューと動的な SEP フリーキューを初期化する関数(iogw_init_queue())

ハードウェア RTL

```
//# ADD Exception Report ID V IN
 always @(posedge XXCLK2G)
  if (VIEEXPCPRST_0)
   VI2EEXCADDVINR <= 1'd0;
  else
   VI2EEXCADDVINR <= VIEEXCINSTIDADDV;

 always @(posedge XXCLK2G)
  if (VIEEXPCPRST_0)
   VI2EEXCADDVINRP <= 1'd0;
  else
   VI2EEXCADDVINRP <= VIEEXCINSTIDADDV;

 //# ADD Exception Report ID IN
 always @(posedge XXCLK2G)
  if (VIEEXPCPRST_0)
   VI2EEXCADDIDINR[0:7] <= 8'd0;
  else
   VI2EEXCADDIDINR[0:7] <= VIEEXCINSTIDADD[0:7];

 always @(posedge XXCLK2G)
  if (VIEEXPCPRST_0)
   VI2EEXCADDIDINRP <= 1'd0;
  else
   VI2EEXCADDIDINRP <= (^ VIEEXCINSTIDADD[0:7]);

 //*********************************************************
 //# ADD Report End ID number buffers 00-11

 //# ID buffer V 00 (n=00)
 always @(*)
  casez ({VIGENDRPYADD,VI2EEXCADDIDV00R,VI2EEXCADDIDV01R})
   3'b00? : VIEEXCADD0IDV00SEL = VI2EEXCADDVINR;     //#Get New V
   3'b01? : VIEEXCADD0IDV00SEL = VI2EEXCADDIDV00R;   //#NOC
   3'b10? : VIEEXCADD0IDV00SEL = VI2EEXCADDVINR;     //#Get New V
   3'b110 : VIEEXCADD0IDV00SEL = VI2EEXCADDVINR;     //#Get New
        V
   3'b111 : VIEEXCADD0IDV00SEL = VI2EEXCADDIDV01R;   //#Get n+1
   default : VIEEXCADD0IDV00SEL = VI2EEXCADDIDV00R;  //#NOC
  endcase //#
        casez({VIGENDRPYADD,VI2EEXCADDIDV00R,VI2EEXCADDIDV0
        1R})

 always @(posedge XXCLK2G)
  if (VIEEXPCPRST_0)
   VI2EEXCADDIDV00R <= 1'd0;
  else
   VI2EEXCADDIDV00R <= VIEEXCADD0IDV00SEL;
```

ハードウェア設計仕様一覧

- CPU設計
  - CPUチップ仕様書
  - SPU仕様書
  - VPU仕様書
  - ACU仕様書
  - PNU仕様書
  - SCU仕様書
  - RCU仕様書
  - 共通診断論理仕様書
  - インタフェース仕様書
- ネットワークスイッチ論理設計
  - RTRユニット仕様書
  - RTR-RCU-HCA インタフェース仕様書
  - RTRチップ仕様書
- メモリ制御論理設計
  - 主記憶装置 (MMC) ユニット仕様書
  - MMCチップ仕様書
- HCA詳細設計
- IOノード詳細設計
  - IOノードサーバ仕様
  - IOノード間インタフェース設計
- 診断システム設計
  - DGP (Diagnosis Processor)
  - SVP (Service Processor)
- 実装設計
  - 実装構造設計
  - パッケージ基板設計
  - 設置環境仕様の設計
- フロントエンドシステム・システム管理系ネットワーク・ファイルシステム構成
  - ユニットB全体構成
  - フロントエンドシステム
  - システム管理系ネットワーク
  - ファイルシステム構成
  - その他ハードウェア
- 論理仕様書

詳細設計の成果利用について

- 詳細設計の成果は，今後理研が行うハイパフォーマンス・コンピューティング技術の研究・開発に広く利用可能
  - 東北大学，海洋研究開発機構等のベクトル機ユーザとの研究開発
    - 将来のアーキテクチャに関する研究開発
    - アプリケーションによる性能評価

- 知財の取扱い（以下の方向で現在協議中）
  - 理研は，本プロジェクト及び国の事業に関係する研究・開発を行う場合において，知財の利用が可能
  - 第3者が国の事業に関係する研究・開発を行う場合において，理研から第3者への利用及び実施許諾が可能

スカラ部単独システムへの構成変更に伴う ベクトル機ユーザへの影響について

スカラ部とベクトル部の比較

	スカラ部	ベクトル部
CPU性能	128GFLOPS（8コア）	256GFLOPS（4コア）
演算器	SIMD演算器（積和演算2個）×2本 ベクトル長：2	ベクトル演算器（積和演算256要素）×16本 ベクトル長：256
キャッシュ	L2キャッシュ：5MB （8コア共有）	L2キャッシュ：8MB （4コア共有）
B/F値	0.5	1.0
ネットワーク	6次元メッシュ/トーラス （プログラムビューは3次元トーラス）	2段ファットツリー

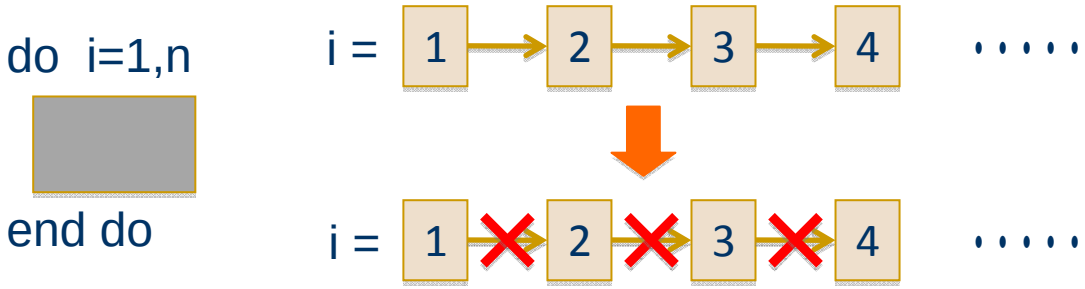
【ポイント】

- スカラ部/ベクトル部共にベクトル化の有効利用が必須
- スカラ部/ベクトル部共にキャッシュの有効利用が必須

- スカラ部/ベクトル部共に大きなB/F値ではない（Cf. ESのB/F値：4）

ベクトル化の有効利用について

- ベクトル化とは. . .
  → doループのインデックス方向の依存関係がなくなるようにプログラムする.



- 次世代スパコンのスカラ部では

  依存関係を取り除いたループに対し

  → ・マルチスレッド並列に使う
  ・SIMD化（ベクトル化）に使う

  IMPACT(Integrated Multicore Parallel ArChiTecture)

結論

ベクトルプログラミングされたコードは次世代スパコンのスカラ部でベクトル化の有効利用が可能

キャッシュの有効利用について

- 現在，6本のコードを重点化アプリケーションとして選択し，高性能化作業を実施中
  1) 超高並列に向けた並列化
  2) キャッシュアーキテクチャでの単体高性能化

重点化アプリケーション (6本)

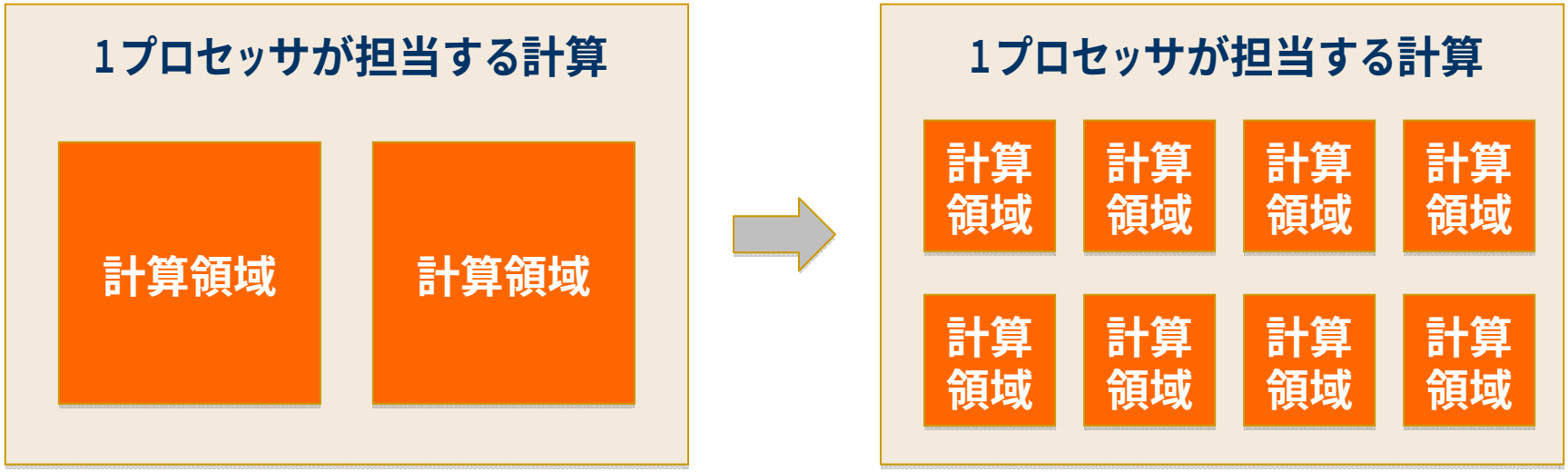
プログラム名	分野	アプリケーション概要	期待される成果	手法
NICAM	地球科学	全球雲解像大気大循環モデル	大気大循環のエンジンとなる熱帯積雲対流活動を精緻に表現することでシミュレーションを飛躍的に進化させ，現時点では再現が難しい大気現象の解明が可能となる．	FDM (大気)
Seism3D	地球科学	地震波伝播･強震動シミュレーション	既存の計算機では不可能な短い周期の地震波動の解析･予測が可能となり，木造建築およびコンクリート構造物の耐震評価などに応用できる．	FDM (波動)
PHASE	ナノ	平面波展開第一原理分子動力学解析	第一原理計算により，ポスト35nm世代ナノデバイス，非シリコン系デバイスの探索を行う．	平面波 DFT
FrontFlow/Blue (FFB)	工学	Large Eddy Simulation (LES)に基づく非定常流体解析	LES解析により，エンジニアリング上重要な乱流境界層の挙動予測を含めた高精度な流れの予測が実現できる．	FEM (流体)
RSDFT	ナノ	実空間第一原理分子動力学計算	大規模第一原理計算により，10nm以下の基本ナノ素子(量子細線，分子，電極，ゲート，基盤など)の特性解析およびデバイス開発を行う．	実空間 DFT
LatticeQCD	物理	格子ＱＣＤシミュレーションによる素粒子･原子核研究	モンテカルロ法およびCG法により，物質と宇宙の起源を解明する．	QCD

- 従来のベクトルマシンで高性能を達成していたコードのうちNICAMを例として，スカラ部上での単体性能チューニングの検討結果について示す．

NICAMの単体性能向上について (1/2)

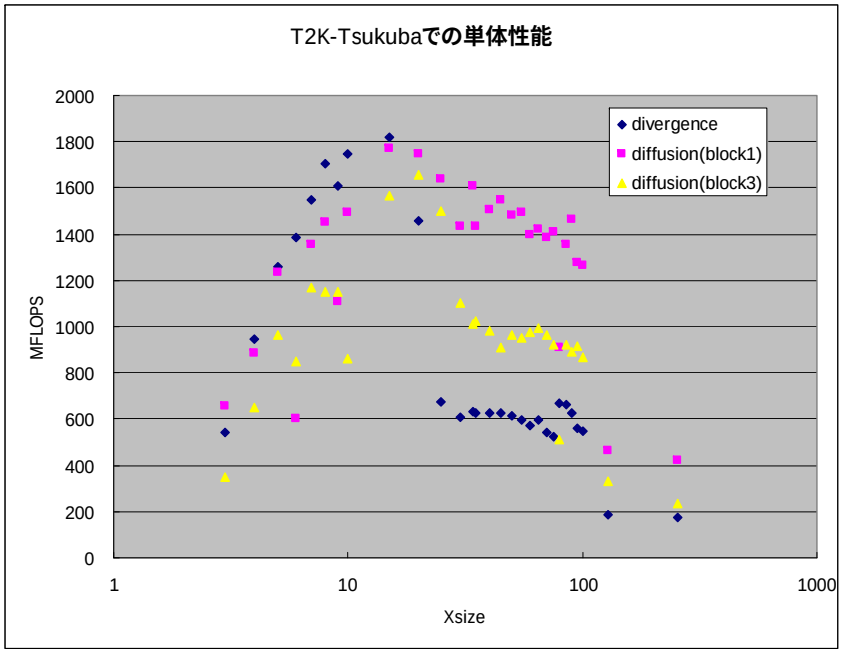
- **実コードNICAMからのカーネルを洗い出し.**
  - divergence/diffusion-1/diffusion-3
  - **特徴は右表のとおり**
- **1プロセッサが担当する計算領域全体を変えずに，計算領域を小さく，その数を多くするチューニング (キャッシュブロッキング) を試行中**

	divergence	diffusion-1 (block1)	diffusion-3 (block3)
FLOP値 (F)	42	44	18
Load + Store数 (L+S)	43	20	20
F/ (L+S)	0.98	2.20	0.90
B/F値	8.19	3.64	8.89



NICAMの単体性能向上について (2/2)

- 計算領域のサイズを小さくすることで，キャッシュに乗り，使い回せる配列が多くなり大幅に性能が向上している.
- さらにSIMD化による性能向上の余地あり.
- さらにループ分割によりキャッシュの利用効率が上がる予測.



- T2K-Tsukubaによる現段階の評価で計算領域の調整によりピーク性能比:20%程度の性能が得られる見通し
- この性能はNICAMの要求B/F値が3.5-8.5程度，T2KのB/F値が約0.3である事を考えると十分高性能と言える
- さらにSIMD化やキャッシュを考えたループ分割によりコードをスカラ部へ適合させる事で更なる性能向上を目指す

NICAMのESでのピーク性能比:40%程度とここに示したピーク性能比:20%以上を比較するとESとスカラ部とのB/F値の比以上の性能は実現可能と予測

まとめ

- 従来のベクトル機向けに作成されたアプリケーションは次世代スパコンのスカラ部で

★ベクトル化の有効利用が可能
★キャッシュを有効利用する事で高い性能を実現可能
（NICAMと同様の検討を他のコードについても実施・検証中）

従ってベクトル部がなくなった影響は限定的と考える．

【参考資料】重点化アプリケーションの評価状況

コード名	高並列性	キャッシュアーキテクチャ上での単体性能
NICAM	640プロセスまでの実測でスケーラビリティは良好(理研実施)．隣接通信がほとんどであり，実行時間に占める通信時間の割合は1割以下である．さらなる高並列が可能と判断している．	カーネル部分によるベンチマークコードを用いた評価では，数万プロセッサで14%程度の性能が得られている（スカラ部での予測評価）．
Seism3D	2048プロセスまでの実測でスケーラビリティは良好(理研実施)．隣接通信が主体である．集団通信も一部あるが，参加するプロセスは一部であり，高並列時ほど影響は小さくなる．65000プロセスまで十分にスケールすると見込んでいる（スカラ部での予測評価）．	現在，単体性能チューニング作業中であるが，高性能が得られる感触を得ている．
PHASE	地球シミュレータで3072並列，クラスタ等にて数百〜2000並列程度の計算実績あり．演算量O(N3)，通信O(N2)であり，大規模系の計算では演算が支配的．現在のバンド並列に加え，波数空間方向も並列化することで更なる超並列に対応可能であり現在作業中．現状のバンド並列のみでも6万原子系であれば数万並列までスケールする評価を得ている（革新プロジェクト）．	Gram Schmidt法による直交化計算等の主要部分は行列-行列積の形で書き換えが可能であり，BLAS Level3のDGEMMを使うことで，90%〜95%の実行効率が期待できる．
FrontFlow/Blue	1024プロセスまでの実測でスケーラビリティは良好(理研実施)．単純問題かつESの通信性能を前提とすれば，数万プロセッサまでのスケーラビリティがある（革新プロジェクト）．複雑問題，現実的な通信性能での評価を実施中．	革新プロジェクトにおいてキャッシュアーキテクチャでのチューニングを実施済み．2-3倍の性能向上を達成し，ピーク比15%程度の性能が得られている．現在更なるチューニングを実施中．
RSDFT	・空間方向とバンド方向の2軸で並列処理が可能であり，数万規模の並列処理が可能である．2軸の並列処理により，集団通信の対象プロセス数を少なくでき，通信コストの大幅削減が可能． ・詳細設計において，カーネル部分によるベンチマークコードを用いて，2軸並列に対する万オーダーの並列化効率を評価し，良好な結果が得られている(スカラ部/ベクトル部)．PACS-CS上で空間方向のみ並列化した． ・実アプリケーションに対して，1,000並列以上で30%程度の実効性能の実績あり．	演算コストの中心は行列-行列積となり全体の約90%を占める．行列-行列積はBLAS Level 3のZGEMMを使っており，90-95%の実効効率が期待できることから，単体性能が十分得られると判断している．
LatticeQCD	T2K-TSUKUBAにて1000プロセス規模の実績あり．隣接通信主体のため高並列性は高い．昨年度の評価ではスカラ部で7万プロセス程度，ベクトル部で1万プロセス程度まで良好にスケールするとの結果が出ている．さらに，アルゴリズムの改善によりノード間通信を劇的に削減できることがわかっており，今年度の評価では当該アルゴリズムを用いたコードで推定を行っている．	現状のアルゴリズムではデータの再利用性が限定的で，3B/FLOPS程度のメモリバンド幅が必要であることが分かっているが，現在再利用性を高めるアルゴリズムの開発を行っており，これを適用できれば，キャッシュの利用効率を格段に向上できると見込んでいる．