秘密情報，取扱注意

資料 6

平成 21 年 4 月 9 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

中間評価に関する質問への回答（理化学研究所）

質問１） 全体構成について
資料８では，複合構成の次世代スーパーコンピュータ全体のハードウェア構成，ネットワーク構成を把握することができない．特にシステムインターコネクト，ファイルサービス，インターネット接続の構成，性能が分かるような，１枚の図（出来れば A3 サイズ）を作成して下さい．

全体システムの構成（概念図）を図 1 に示す．スカラ部及びベクトル部については，第 1 回資料 8-1 に述べたとおりである．各演算部のローカルファイルシステム及び共有ファイルシステムについては，添付資料に示す．

統合フロントエンド用のサーバ等についてはフェイルオーバー機能を持つように二重化構成とするが，平成 23 年度に整備することとしており，それらの詳細な機器構成については今年度以降に詳細を確定する予定である．外部ネットワーク（SINET 等）への本格的な接続は，平成 24 年 4 月を予定しており，NII の次期 SINET の整備状況を踏まえながら，建屋内ネットワーク機器及び外部ネットワーク機器等については今年度以降に設計を行う．

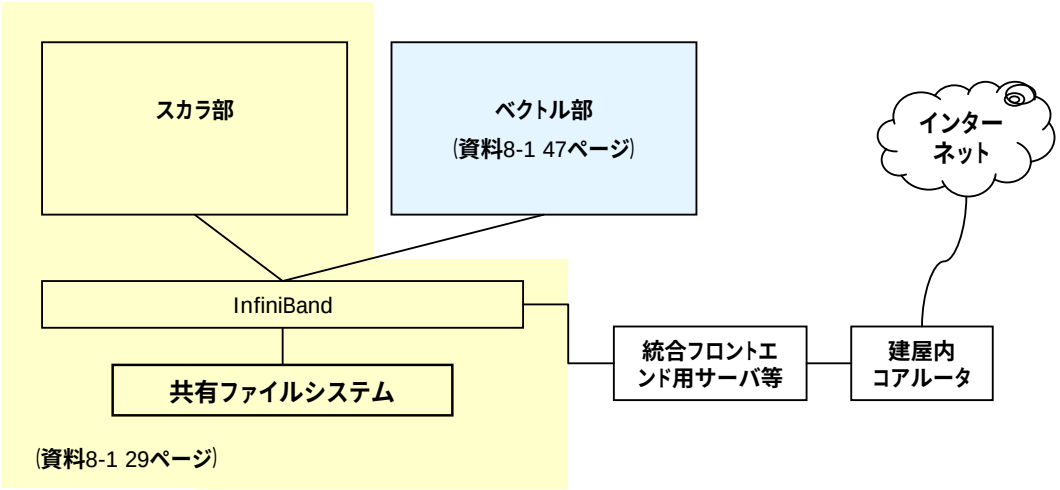


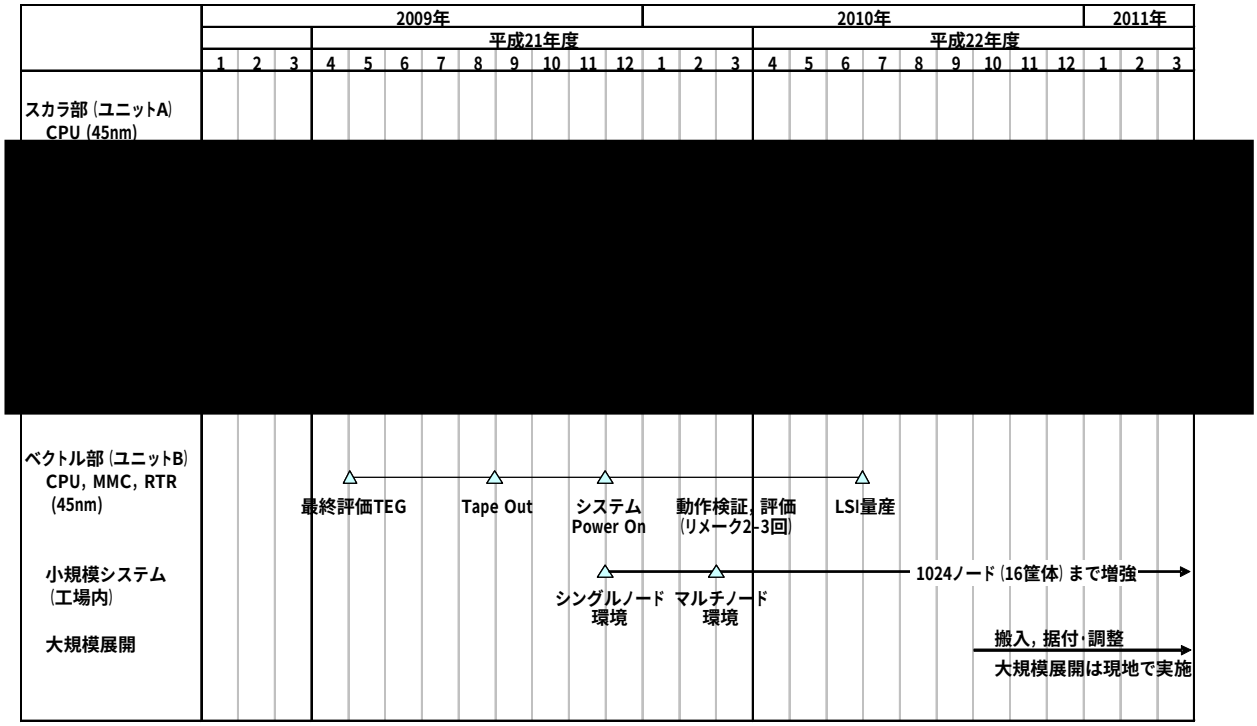
図１ システム構成

質問２）スケジュールについて

２－３）資料８－１，３ページにおいて，システムは H22 年 4 月から製造・搬入・据え付けであり，かなりの規模（スカラ 5PF）が H23 年 4 月に全体動作を開始するように記載されている．現在，設計段階で TAPE OUT が今年（H21）夏から秋と予想される開発で，このスケジュールは非常に厳しいものがある．完成までの道程が可能であるか判断できる中間段階の目標時期を示して欲しい．ただし，中間段階とは，プロセッサチップ，ネットワークチップ，光インターフェースチップなどについてプログラム実行が可能なものの TAPE OUT，リメーク回数予測と，リメークに要する期間，小規模システム（たとえば数筐体）の動作開始時期，システムのデバグが終了し，大規模展開が可能となる時点を含む．

スカラ部及びベクトル部の整備スケジュールは表 1 のとおりである．

表１ 整備スケジュール



質問３）複合構成の最適性について
３－１）資料８－１，６，７ページにおいて，複合構成が必要な JOB のイメージがあるが，必要とされる性能見積りに欠けている．具体的に想定している連成ジョブ（タイプ２，タイプ３各々について）アプリケーション内容，必要とされるスカラ部・ベクトル部間接続バンド幅とレイテンシ，なぜ，すべてスカラ部で実行するのではなく，連成ジョブにすることが必要かを示して下さい．

統合汎用システムの意義は，アプリケーション・ソフトウェア資産のより有効な利用，共用施設としての効率的なユーザ対応，マルチフィジックス・マルチスケールのシミュレーションに新しい実行環境を提供することであると考えている．また，実際の運用に際しても，システムパーティションによる運用を想定しており，複合構成によりパーティションに分けられた各演算部で効率的なジョブタイプを割当てることにより，システム全体のターンアラウンドを高める運用ができると考えている．

したがって，一つのジョブ実行形態として，ハイブリッドシステムによる連携アプリケーション実行についてユーザに指針を与えるために，今後の評価により適切な利用及び運用方法を検討することとしている．また，設計・製造計画評価検討部会での評価により，具体的な連携アプリケーション，それらの連携アプリケーションによるコネクト部構成の評価をするべきとの指摘を受けており，次の質問への回答のとおり 2 つの連携アプリケーションに対して，評価を実施することにしている．

３－２） 資料８－２，19 ページ．連携アプリケーションとして
RISM－OpenFMO（九州大学との共同研究）
MSSG－放射モデル（JAMSTEC との共同研究）
について，Job の効率的実行に必要なコネクト部のバンド幅，連成する場合のベクトル部・スカラ部の負荷分散，両方をスカラ部で実行する場合に比較しての高速化見積りを示してください．

上記 2 つの連携アプリケーションを評価を行い，ハイブリッドシステムとしての評価を実施することにしている．

RISM－OpenFMO については，FFT を使う RSIM をベクトル部で実行し，スカラ部ではフラグメント間の並列化を行うことが有効であると考えている．また，MSSG－放射モデルでは，地球シミュレータで開発された MSSG コードをベクトル部で実行し，モンテカルロ法による放射モデルをスカラ部で実行する方法を評価する．いずれの場合も，スカラ部－ベクトル部間のデータ転送量は時間方向のシミュレーションパラメータに依存し，現状のシステム構成に最適な連携方法を検討できると考えている．

質問３） コネクト部について
資料８－１には，コネクト部および３種のファイルシステムの性能および構成の詳細が記載されていない．ネットワーク接続構成，バンド幅，実効バンド幅の記載されている図を作成して下さい．
共有ファイルシステムの構成，実効性能（Read，Write）を示してください．また，実行性能が 80+22=102GB/s であると仮定し，ファイルサーバが 30 台である場合（資料８－１，30 ページ），サーバ 1 台あたり 3GB/s 程度の性能が要求される．この性能が妥当であるという根拠を示してください．

コネクト部および３種のファイルシステムについては別添資料 1 に示したとおりである．資料から，グローバル IO サーバ（OSS:Object Storage Server）あたりの実行帯域 3GB/s は十分確保できる．

質問４） スカラ部
４－１）資料８－１，２０ページにおいて，バイセクションバンド幅が 49TB/s（双方向）と記載されているが，このバンド幅の計算式を示してください．

24(X) x 17(Z) x 12(link/group) x 5GB/s x 2（双方向） = 48,960GB/s

４－２）ORNL の Jaguar と比較して，ノードあたり性能が約 10 倍であるが，3D トーラス性能がかえって小さくなっている．リンクあたり 5GB/s が適切である根拠を示して欲しい．

Jaguar（Cray XT5）のノードあたりの性能は，チップ性能が 36.8GFLOPS（AMD Opteron 2.3GHz Quad-Core），ノード当たり Dual Socket で 73.6GFLOPS である．したがって，スカラ部のノードあたりの性能は 1.7 倍程度である．
また，Jaguar は CPU - SeaStar2+ 間の Hypertransport の帯域が 6.4GB/s(双方向)で 0.087 B/F，スカラ部は CPU - ICC 間の帯域が 20GB/s（双方向）で 0.313B/F である．Jaguar のノード間帯域は 3.2GB/s(双方向)で 0.087B/F,スカラ部のノード間帯域は 5.0GB/s(双方向)で 0.078B/F である．性能あたりのスイッチ帯域は，Jaguar が 0.78B/FLOP（=4.8x2x6/73.6），スカラ部が 0.78B/FLOPS（=5x2x10/128）であるが，ノードあたりの同時通信数は Jaguar が 1 であるのに対しスカラ部は 4 であり，現時点としてはこのリンク性能は適切であると考える..

質問５） 電力消費量について

全体の消費電力から概算すると，スカラ部は電源部を除いて 1 プロセッサチップあたり 144W，ベクトル部は 453W 以下で動作することが必要である．しかしながら，ベクトル部の RTR，光モジュールおよび Fat Tree スイッチの消費電力は示されていない．ネットワークを含め，この消費電力が可能である見積りを示して下さい．

ベクトル部の電力バジェットは以下の通り．電源部，冷却部，I/O ノード，ローカルディスク装置を除いて１プロセッサあたり 366W である．

表 2 ベクトル部の電力バジェット

ベクトル部全体	単体 (W)	数量	電力 (kW)
Nノード	18225	384	6998
スイッチ筐体	4400	80	352
合　計			7350

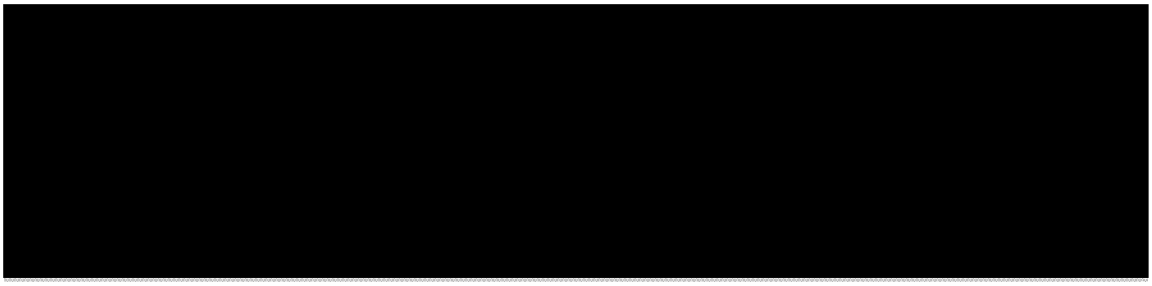
内　訳		単体 (W)	数量	電力 (W)
Nノード				
	CPU	200	32	6400
	メモリ	150	32	4800
	RTR (ノード内スイッチ)	230	1	230
	光接続部	300	1	300
	冷却部	750	1	750
	DDコンバータ他			3245
	I/Oノード	1000	1	1000
	DISK	1500	1	1500
小　計				18225
スイッチ筐体				
	RTR	230	8	1840
	光接続部	150	8	1200
	冷却部	60	8	480
	DDコンバータ他			880
小　計				4400

質問６） 性能目標であった，HPCC Award 4 項目の実現可能性について，
資料８－１，63 ページの表によると
(1) Global HPL は，Top500 とほぼ同じであるため，最高値の達成は困難である，
(2) Global Random Access は，目標値が現在の最大値の約 2 倍であり，2012 年には US システム性能が 20 倍近くに上昇することを勘案すると最高値の達成は非常に困難である，
(3) Global FFT は，ベクトル部のネットワークが SX-9 より弱いため，BlueGene など FFT に向いた構成と比較すると不利であり，最高値の達成は非常に困難である，
(4) EP Stream は，メモリバンド幅の総和とほぼ等しいため，20PF を達成するシステムを上回ることは非常に困難である．
ため，4 項目すべての達成が困難と判断されるが，この判断で正しいか．

なお，Global FFT の推定値が，システムのバイセクションバンド幅と比較して著しく高いが，その根拠，または，180 および 226TFlops を得た計算式を示してください．

(1) Global HPL については，TOP500 の世界一を奪取することと同程度に困難であると認識している．
(2) Global RandomAccess については，ピーク性能よりもバイセクションバンド幅が重要であると認識している．その意味で，Sequoia（BlueGene/Q）よりも BlueWaters の方がより広いバイセクションバンド幅になると思われ，これを上回るのは困難であると認識している．
(3) Global RandomAccess と同様，バイセクションバンド幅が重要となるため，BlueWaters を上回るのは困難であると認識している．
(4) EP stream(triad) per system については，BlueWaters のメモリバンド幅は 0.5B/FLOP と予想され（http://www.channelregister.co.uk/2008/07/11/ibm_power7_ncsa/より），メモリ帯域の実行効率が同程度と仮定すれば，スカラ部の最大メモリバンド幅の方が大きいので，スカラ部が有利である．また，Sequoia（BlueGene/Q）については，メモリバンド幅が 0.2B/FLOP（Sequoia RFP より）の場合，スカラ部が若干有利と考えている．

○ Global FFT の推定値について
いずれの演算部についても，演算時間（推定値）と通信時間（推定値）を求め，問題規模から求められる演算量を用いて推定 Flops 値を求めた．
【スカラ部】



【ベクトル部】

- 問題規模とシステム構成： プロセス数 49,152，サイズ 5,566,277,615,616，演算量 5NlogN = 1178TFlop，
- 演算時間 1.8 秒
  １プロセス分のカーネルに対し，実測環境（SX-8，8CPU 構成）にて CPU 時間とメモリアクセス時間を性能情報として取得（実測時間 3.5 秒）．演算器構成及びメモリ性能の差を考慮して，机上にて推定時間を算出．なお，6 ステップ FFT アルゴリズムを用いたときの 1 プロセス分のカーネルは，48 組の FFT（サイズ 2,359,296）である．
- 通信時間 3.4 秒
  All-to-All の通信性能の基礎データ（通信時間）を，ベクトル部のファットトリーを模擬した環境により，プロセス数およびデータサイズのテーブル関数として推定．
  データサイズは，16 バイト（複素） x 2,359,296 x 48（組）/ 49,152 = 36864 (Byte) なので，プロセス数 49152，データサイズ 36,864（バイト）の時の All-to-all の通信時間は 1.26（秒）と算出．3 回の転置転送があるので
  転送時間 = 3 x 1.13 秒 ＝3.78 秒 ※平成 24 年 6 月公開時点の注意書き「3.39 秒」に修正
- FLOPS 値 ＝ 1178TFlop/(1.8+3.4) = 226 TFLOPS

質問７） 重点化アプリケーションの予測性能について
各々のアプリケーションについて，ベクトル部およびスカラ部での予測される性能値を示してください．ベクトル部が有利である重点化アプリケーションはどれかを示してください．

RSDFT，NICAM，QCD の 3 本についてはベンチマークコード（実アプリケーションのカーネルを抽出）による評価を実施している．各ユニットの現時点の評価結果は，以下の通りである．ただし，ターゲットとした問題で使用できる最大プロセッサ数がコード毎に異なるためコード毎にピーク性能が異なっている．

【スカラ部】

RSDFT	5.58 PFLOPS	(ピーク性能 11.28Pflops に対しピーク比 49.45%)
NICAM	1.42 PFLOPS	(ピーク性能 10.49Pflops に対しピーク比 13.57%)
QCD	1.12 PFLOPS	(ピーク性能 8.39Pflops に対しピーク比 13.36%)

【ベクトル部】

RSDFT	1.72 PFLOPS	(ピーク性能 2.99Pflops に対しピーク比 57.6%)

NICAM　0.45 PFLOPS　(ピーク性能 2.62Pflops に対しピーク比 17.2%)
QCD　0.30 PFLOPS　(ピーク性能 3.15Pflops に対しピーク比　9.6%)

現在は実アプリケーションの性能最適化に着手しており（資料 8-1　67 ページ），それらの状況を踏まえた実アプリケーションの性能に対する見通しを以下に示す．

<RSDFT>　コストの中心は行列-行列積であり，90-95%の 実効効率が期待できること，また空間方向とバンド方向の 2 軸並列の試作まで終了している事から，ピーク性能比で 35% 程度は可能と考えている．

<NICAM>　高並列性が確認できている．現在，実コードに対しスカラ部及びベクトル部共にベンチマーク結果以上の性能向上を目指して性能最適化作業中である．特にベクトル部に対しては効率 30%を目指して性能の最適化中である．

<QCD>　通信を減らすアルゴリズム，キャッシュを有効利用可能なアルゴリズムを開発中であり，最終的にはベンチマーク結果以上の性能が出る可能性がある．

Seism3D，PHASE，FrontFlow/Blue（FFB）については，ベンチマークコードによる評価結果はないが現状の性能見通しを以下に示す．

<Seism3D>　高並列性は確認できており，また計算の中心部分は NICAM より B/F 値を要求しない事が分かっているため，NICAM 以上の性能が期待できる．

<PHASE>　RSDFT と同様に，コストの中心が行列-行列積であり，またバン ド方向と波数方向の２軸並列化の見通しも立っているため， RSDFT と同程度の性能が得られると考えている．

<FFB>　高並列性は確認できており，単体性能は現状のスカラ機で 15%程度出ている．スカラ部においても 15%程度の性能を得るために開発元と性能向上の方策について検討している．

<アプリケーションから見たベクトル部の意義>

NICAM については，3.5Km 格子の 10 年間積分を 1 年に 10 本程度走らせる要望がある.その計算時間を見積もる.現状 ES 上で 20Tflops 分(ES の 1/2)を使用し 3.5Km 格子の 1 日分のシミュレーションに 6 時間かかる.ES での実行効率は約 40%であるため,8Tflops の性能があれば１日分のシミュレーションに 6 時間必要.従って 8Tflops の性能であれば 10 年間のシミュレーションには 2.5 年の計算時間が必要.1Pflops の実行性能が得られれば 8Tflops の 125 倍の性能であるため,365 日×2.5 年/125=7.3 日で実行可能となる.従って 10 本の 10 年積分を実行する時間を見積もると全体で 73 日の計算時間となる.実際には,10 年積分では鉛直方向メッシュを２倍にする要望もあり,その場合ベクトルの運用を考えると 5 ヶ月占有する見積りとなる.5 ヶ月の計算時間は,一つのプロジェクトの占有期間としては最大かと考える.もし 10 年積分に２週間掛かると,この倍の時間の占有期間となり運用上非常に難しいと考える.これらの理由によりピーク性能（※平成 24 年 6 月公開時点の注意書き「実行性能」に修正）1Pflops のベクトル部は必要と考える．

またベンチマーク結果から，スカラ部とベクトル部で大きな性能差は出ていないが，NICAM や Seism3D のようにベクトル向けにコーディングされているコードについては，性能を出すためのコードの書き換えがベクトル部において少ないと考えており，ユーザの負担を少なく速やかにアプリケーションの成果が得られるものと考えている．

コメント） Tofu インターコネクトの VC（Virtual Channel）数，及びパケット転送方式なども知りたいところである．

VC（Virtual Channel）数は 4 チャネル，パケット転送方式はカットスルー方式である．

質問） ベクトルユニットが間に合わない

ベクトル部の製造スケジュールは，平木委員の質問 2−3)への回答のとおりである．リメーク 2−3 回程度で過去の同種の CPU 設計期間と比較してほぼ同等であり，計画通り製造できると考えている．

以上