Vol.2014-HPC-145 No.36
2014/7/30

HPC アプリケーション I/O 特性を考慮した 省電力指向階層化ストレージシステムの検討

鵜飼敏之[†1]　清水正明[†1]　飯田恒雄[†1]　岡嶌正道[†1]　守屋卓[†1]　石川裕[†2 †3]

我々は 2018 年頃のポストペタスケールシステムを視野に入れた，次世代大規模 HPC システムにおけるストレージシステムのあるべき姿の明確化に取り組んだ．次世代大規模 HPC システムで問題となる消費電力を削減するためにはオンライン及びオフラインストレージを組み合わせた階層化ストレージシステムの適用が有効である．階層化ストレージシステムの適用実現に向けた課題は，(1)容量および性能最適化，(2)ストレージ階層のトランスペアレント化である．このうち，本報告では(1)容量および性能最適化に取り組み，アプリケーションの I/O 特性を考慮した構成案策定方法を提案する．提案方法を適用して，次世代大規模 HPC システムにおいて実行が想定されるアプリケーションの I/O 特性に基づく，省電力指向および性能指向の二つの構成案を策定した．これら構成案では，消費電力を，1EB フル RAID 構成の 2,967KW と比較して，省電力指向構成で約 66%削減する 1,006KW，性能指向構成で約 50%削減する 1,476KW に低減可能な見通しである．この結果，次世代大規模 HPC システム全体の消費電力を 20MW とした場合に，システム全体の消費電力に対するストレージシステムの消費電力の割合を，省電力指向構成で約 5%，性能指向構成で約 7.4%にすることが可能となることを確認し，目標の 10%以内に抑える見込みを得た．

1. はじめに

近年，スーパーコンピュータや PC クラスタを利用した HPC (High Performance Computing)は，気象予報，地球環境予測，防災など，社会基盤の一部を形成するだけでなく，物質・材料，ライフサイエンス，製造などの分野の企業や研究機関においても重要性を増している．これらの数値シミュレーションなどの科学技術計算，大量データ処理においては，より大規模で詳細な解析を実現するため，演算量は今後も拡大する．

日本では，スーパーコンピュータを，科学技術の発展，産業競争力強化，安全安心な国づくりに不可欠な国家基幹技術として，継続的に整備・共用を推進してきている．最近では，Top500 の 2012 年 6 月と 11 月のランキングで首位となったスーパーコンピュータ「京®」[1-2](以下，京コンピュータ)がこの取り組みの成果である．この京コンピュータ後継も視野に入れた，次世代大規模 HPC システムの検討は，まず，文部科学省に設置された「HPCI (High Performance Computing Infrastructure)計画推進委員会」のもとで，大学，研究機関，企業から有識者を集めた「今後の HPC 技術の研究開発のあり方を検討するワーキンググループ(WG)」発足から始まった．同 WG は 2012 年 3 月に「今後の HPCI 技術開発に関する報告書」「計算科学ロードマップ白書」「HPCI 技術ロードマップ白書」をまとめている[3-5]．次に，2012 年度より 2 ヵ年計画で，京コンピュータの後継についての調査検討を行う，「将来の HPCI システムのあり方の調査研究」が開始された(文部科学省委託研究，2012 年 7 月から 2014 年 3 月まで実施)．これは，2018 年から 2020 年頃のスーパーコンピュータなどの計算インフラに求められる機能・性能等の要件を明らかにし，それに対する技術的知見を獲得することを目的とした調査研究プロジェクト（Feasibility Study，以下 FS）である．この FS では，東京大学，筑波大学，東北大学の各大学を中心とする 3 チームが「システム設計分野」について FS を実施し，さらに「アプリケーションソフトウェア分野」について，理化学研究所と東京工業大学が共同で実施した．

我々はこの FS のうち，東京大学中心に検討を進める研究題目「レイテンシコアの高度化・高効率化による将来の HPCI システムに関する調査研究」[6]において 2018 年頃の実現を目標にしている次世代大規模 HPC システムにおけるストレージシステムのあるべき姿の検討に取り組んだ．

次世代大規模 HPC システムでは消費電力がその設計や実効性能を制約する最大の要因の一つと考えられており，ストレージシステムに限ってもそれは同様である．演算量の飛躍的な拡大により，入出力するデータ量は増大し，それと共にストレージ容量も増大を続ける．例えば，京コンピュータにおけるストレージ容量は総計 41PB (Peta Byte)以上であり，次世代大規模 HPC システムではこれが 1EB (Exa Byte)に近くなることが想定される．容量の増大に伴い，ストレージシステムの消費電力は大きくなる．多くの大規模 HPC システムでは，ストレージシステム(I/O ノードとストレージ)の消費電力は非公表だが，例えばストレージのラックあたりの消費電力を 10KW と仮定した場合，システム全体の消費電力に対して，10%前後を占めると考えられる．本報告では，次世代大規模 HPC システムにおけるストレージシステムにおいて，消費電力の低減を目指し，階層化ストレージシステムの適用を提案する．階層化ストレージシステムでは，RAID 装置などいわゆるオンラインストレージに加えて，テープライブラリなど消費電力当たりの容量効率がよいオフラインストレージを積極的に活用する．さらに，階層化ストレージシステム適用に際し，実行するアプリケーションプログラムの I/O のふるまいに基づいて，

†1 (株)日立製作所
Hitachi Ltd.
†2 東京大学　情報基盤センター
Information Technology Center, The University of Tokyo
†3 独立行政法人理化学研究所　計算科学研究機構
RIKEN AICS

容量および性能を最適化する構成案策定方法を提案し，消費電力低減の有効性について検討した結果を述べる．

2. 大規模 HPC システム

ここでは，現状の大規模 HPC システムの構成を示し，それをベースに，2018 年頃に想定される次世代大規模 HPC システムのラフイメージと，ストレージシステムの構成について述べる．

2.1 京コンピュータシステムの概要

現状の大規模 HPC システムとして，京コンピュータシステム[1-2]の構成について述べる．京コンピュータシステムは，大きく，実際にジョブを実行する本体システムと，ユーザが京コンピュータを利用するためのログインサーバやジョブ管理サーバ，演算のプレ処理やポスト処理を行う周辺システムから成る(図 1).

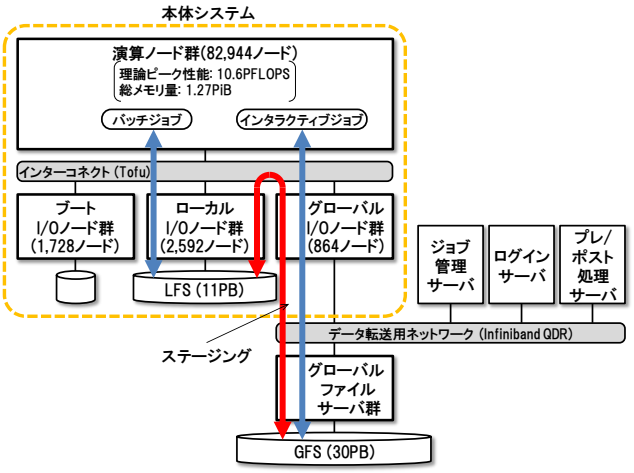

図 1　京コンピュータシステムの構成概要

本体システムは，演算ノード群と I/O ノード群，およびノード間を接続するインターコネクトから成る．I/O ノードはその用途により，ブート I/O ノード，ローカル I/O ノード，グローバル I/O ノードに分類される．ブート I/O ノードはその名前の通り，システムブート用のデータを保持する．ローカル I/O ノードは，演算ノードにおける演算で利用するデータを保持する．グローバル I/O ノードは，データ転送用ネットワークを介して，本体システムの外部に存在するグローバルファイルサーバとローカル I/O ノードの間のデータの授受を中継する．

京コンピュータのストレージシステムは，ローカル I/O ノードに接続されたストレージ(11PB～)とグローバルファイルサーバに接続されたストレージ(30PB～)の二種から成る．それぞれのストレージでは，FEFS® (Fujitsu Exabyte File System)でファイルシステムが構築されており，ローカルストレージ側は LFS (Local File System)，グローバルストレージ側は GFS (Global File System)と呼ばれている．LFS は演算ノードで実行するジョブ専用の高速一時保存領域として利用され，実行中または実行待ちのジョブのファイルが一時的に置かれる．GFS はユーザファイルを格納する大容量の共用領域として利用される．GFS はログインサーバからのアクセスに加えて，計算ノードで実行されるインタラクティブジョブ利用で直接アクセスすることができる．

バッチジョブで利用されるデータは GFS から LFS へ転送されて用いられ，演算終了後の結果は LFS から GFS へ転送されて，ユーザが利用できるようになる．このデータ転送をステージングと呼ぶ. ステージングは, GFS から LFS へ適切なファイルを配置により，大量のジョブが同時実行される環境において，ジョブ間の I/O の競合を防ぐことができる反面，使い勝手の面などで制約もある．

2.2 次世代大規模 HPC システムのラフイメージ

次世代大規模 HPC システムにおけるストレージシステム検討に当たり，最初にシステムのラフイメージを次のように仮定した．

◆演算ノード
- 総演算性能：1EFLOPS
- メインメモリ：10PB
- ノード数：100,000 ノード（性能 10TFLOPS/ノード，メモリ 100GB/ノード）

◆I/O ノード(ファイルシステム)
- LFS 相当: 100PB (メインメモリ 10PB の 10 倍)
- GFS 相当: 900PB (総ストレージ容量を 1EB と仮定)

上記のラフイメージを元に，2018 年頃のストレージシステムの消費電力を算出する．

表 1 に，記憶メディアやインタフェースの技術動向[7-10]，および，現状 RAID 装置の構成などから，2018 年頃に想定されるストレージ装置セット(接続するサーバ，FC スイッチ等も含む)の諸元を示す．テープライブラリおよび光ディスクライブラリにはフロントエンドのキャッシュとしてのディスク装置を含む仮定である．

表 1　2018 年頃に想定されるストレージ装置セット諸元

比較項目	RAID		テープライブラリ	光ディスクライブラリ
	容量重視	容量非重視		
容量	10.7PB	5.3PB	8.4PB	4PB
スループット	16GB/s	←	2GB/s	←
定格消費電力	23.7KW	15.7KW	5KW	4KW

この諸元に基づき，上記容量合計 1EB のファイルシステムを全て RAID 装置で賄う場合の消費電力は 2,967KW となった(表 12 「(参考)」欄).

「HPCI 技術ロードマップ白書」[5]では，次世代大規模 HPC システムは全体の消費電力を 20～30MW に抑えることを想定している．システムの消費電力は可能な限り低く抑えたいため，全体を 20MW と考えると，ストレージシステムの消費電力 2,967KW は, 全体に対して 15%で過大となる恐れがある．従って，我々の検討では，ストレージシステムの消費電力を 2MW 以内に抑えることを目標とする．

このようなストレージシステムの消費電力の削減には，アーカイブ装置など，いわゆるオフラインストレージの活用が考えられる．

現状考えられるアーカイブ装置は, テープ, 光メディア, MAID (Massive Array of Idle Disks)などである．これらアーカイブ装置には一長一短あり,使い分けが必要である．テープは，かつて磁気ディスクに比較して安価であったが，昨今では磁気ディスクとの価格(ビットコスト)差が狭まっている．一部のケースでは使い勝手の面で，テープ装置が敬遠され，磁気ディスクのみでシステムを構築する場合も見られている．しかし，電力面においては，常時稼働させておく部品が少ないテープが優位である．このため，電力消費が大きくなる大規模なシステムではテープが必要になると考えられる．

光メディアは, Blu-ray の次世代規格でビットコストが大きく低下することが期待できる．ただし，性能面，特に書込み性能で磁気ディスク，テープと同等になる可能性は低いため，WORM (Write Once Read Many)での用途に利用が限定される．

磁気ディスクでは，RAID を電源制御可能にした MAID 技術があり，これを活用した省電力システムの研究が進んでいる．しかし，データが複数の磁気ディスク上に一様にストライピング格納されているような場合は，省電力効果が薄れるなど使用上の課題がある．

低消費電力化に向けては，オンラインストレージである RAID と各オフラインストレージを適材適所で使い分ける構成が必要である．ただし，現段階で各オフラインストレージの使い分けについて言及することは難しい. このため，以降本報告ではオンラインストレージとして RAID を，オフラインストレージとしてテープライブラリを代表として扱うことで構成を検討する．

検討に先立ち，オンラインストレージとオフラインストレージの消費電力について効率を押さえておく．表 1 より，2018 年頃に想定される RAID 装置とテープライブラリに関して，必要スループットを無視した場合の容量/消費電力比は次のように算出できる．

◆RAID (容量重視):

・容量/消費電力 比 = 10.7PB/23.7KW = 0.45PB/KW

◆テープライブラリ:

・容量/消費電力 比 = 8.4PB/5KW = 1.68PB/KW

消費電力当たりの容量で 3.7 倍(= 1.68/0.45)テープライブラリの方が効率がよい．従って，消費電力の削減にはアーカイブ装置など，オフラインストレージ装置の活用が有効であることがわかる．

ただし，ここではスループットの要件を無視しているため，上記試算になるが，アプリケーションプログラムのスループット要件を満たしつつ，オンラインストレージとオフラインストレージの容量の最適化を図る必要がある．

3. 階層化ストレージシステムとその構成策定方法の提案

ここでは，消費電力低減に有効な階層化ストレージシステムの提案と, 実現に向けた課題を示す. この課題のうち，階層化ストレージシステム構成の容量および性能に関して，アプリケーション I/O のふるまいに基づいて最適化する構成案策定方法を提案する．

3.1 基本方針

2018 年頃に実現すべき階層化ストレージシステムの概念図を図 2 に示す.

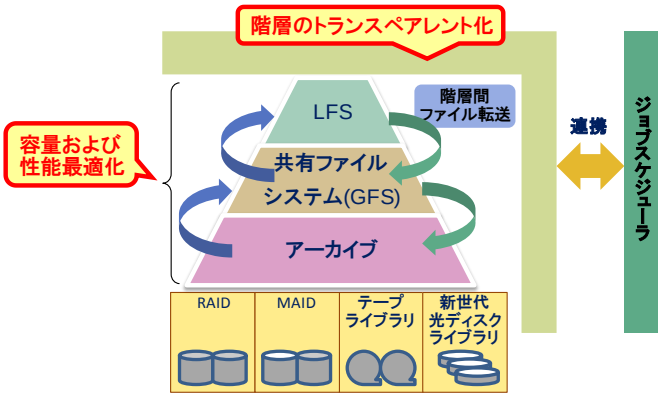

図 2 階層化ストレージシステムの概念図

以降の検討ではこの具体化に取り組む．大きな課題は，(1) 容量および性能最適化と, (2) ストレージ階層のトランスペアレント化である. これら課題について簡単に述べる．

(1) 容量および性能最適化

階層化ストレージの各階層において，必要な容量および性能(スループット)を明確化して，従来型 RAID やテープの他，新型光ディスクや省電力 RAID (MAID)を適材適所で組み合わせて，余剰をなくす必要があるが，指針が明確でない．

(2) ストレージ階層のトランスペアレント化

現状の京コンピュータではユーザが明示的にステージン

グを行っている．ステージングは，GFS から LFS へ適切にファイルを配置することにより，大量のジョブが同時実行される環境において，ジョブ間の I/O の競合を防ぐことが可能になり，演算資源の有効活用が可能になる．しかし，その反面，次のような弊害がある．

・TAT (Turn Around Time)の短いジョブ実行が要求される場合はステージングにかかる時間が相対的に大きくなる．例えば，プログラムの開発/チューニング段階などで，試行を繰り返すケースでは，ステージング時間が無視できない．また，現在はジョブ実行中のステージアウトができないため，演算の途中状況を確認しながらジョブを実行するようなことも不可能である．

・基本的に LFS が保持するデータは GFS のサブセットなるため，ストレージ容量が余計に必要になる．

これらを解消するため，ストレージの階層構造は維持しつつも，データの物理的な格納場所をユーザに意識させない，透過的なファイルアクセスを実現する必要がある．さらに，ジョブ実行時においては，必要に応じて自動的にファイルの階層間転送を行う必要がある．このためには，透過的ファイルアクセス機能と，ジョブの実行と終了を管理するジョブスケジューラが連携する機能が必要になる．

以下，本報告では，「(1)容量および性能最適化」にフォーカスして検討した結果を述べる．具体的には，階層化ストレージの各階層の容量および性能最適化に向け，その指針を明らかにする必要がある．本報告では，アプリケーションの I/O のふるまいから要件を抽出して，それに基づいて，階層化ストレージシステム構成(容量およびスループット)を策定する方法を提案する．

3.2 データアクセスモデル

アプリケーション I/O のふるまい把握に向け，アプリケーションの動作とアクセスするファイルを用途ごとに定義し，アプリケーションのデータアクセスモデルとして定める．これを図 3 に示す．

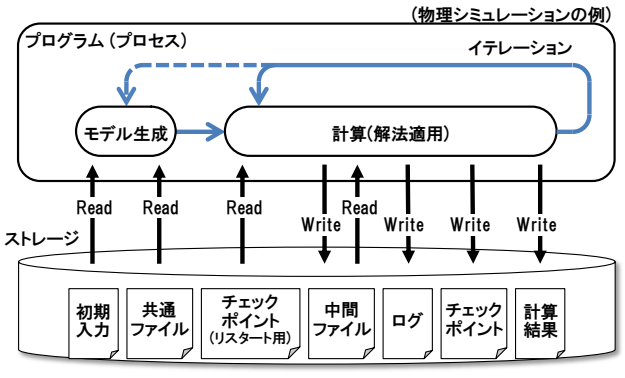

図 3　データアクセスモデル

また，このデータアクセスモデルで定義したファイル種別を表 2 に示す．

表 2　ファイル種別の定義

#	ファイル種別	説明
1	初期入力	プログラム開始時に Read する初期データ. 先行のプログラムの計算結果を引き継ぐ場合も含む.
2	共通ファイル	辞書などを想定．一つのプログラムが繰り返し Read, または，複数のプログラムで共有 Read.
3	中間ファイル	作業用データ．プログラム内のみで利用．プログラム終了後は不要なデータ.
4	ログ	プログラムの実行状況，実行結果などの確認用．追記型.
5	計算結果	計算の結果，出力される解．プログラム終了後も必要なデータ.
6	チェックポイント	運用上の制限(使用時間制限を越える実行時間)，対障害，アプリ都合(一定ステップごとに出力など)で，ジョブ実行を再開できる情報．システムの信頼性が十分で, 運用上の制限(24 時間以上のマシン割り当て不可など)がない場合，I/O する必要がないデータ.

ここで定義したデータアクセスモデルに，実際のアプリケーションの I/O を当てはめ，容量，I/O 回数などの情報からふるまいを特定する．

3.3 I/O に着目したアプリケーションの分類

ここでは，アプリケーション毎に I/O 基本特性を押さえた上で，後の構成案策定に活用するため，アプリケーション I/O について俯瞰される特性から，アプリケーションを分類する．

「計算科学ロードマップ白書」[4]に記載されている，次世代大規模 HPC システムでの実行が想定されるアプリケーションプログラムについて俯瞰的に調査した結果，I/O 特性として次の点を見出した．

(A) データ量およびスループットの観点で，支配的な I/O は，計算結果，および，チェックポイントの出力.

(B) (A)の出力に関しては，プログラムのイテレーションごとに，最新 1～2 世代が永続化されていれば上書き可なデータ(イテレーションごとに最適化された状態を求める場合やチェックポイント)と，上書き不可なデータ(時間発展する状態を時系列に蓄積)が存在.

(C) 入力(初期入力，共通ファイル)に関しては，データ量およびスループットの観点で支配的になるアプリケーションは少ないが，ノード間で共用利用されるデータや，設定ファイルなどイテレーションごとに読みださ

れるファイルなど，再利用性の高いデータが多く存在.

(D) 個人ゲノムの同定を典型例として，多数の入力データと各種リファレンスデータとのマッチングなど検索的な用途に伴うアプリケーションの重要性が増大.

これら俯瞰される I/O 特性に基づき，初期入力，計算結果出力，チェックポイント(出力)に着目した，アプリケーションの分類案を表 3 に示す.

表 3　I/O のふるまいに基づくアプリケーションの分類案

#	分類	観点	トータルの I/O 量		
			アプリで必要な I/O		チェックポイント
			初期入力	計算結果	
i	時系列シミュレーション型	計算結果を時系列に蓄積	小～大	大 (上書不可)	大 (上書可)
ii	最適化問題型	最終の計算結果のみ必要	小～大	大 (上書可)	大 (上書可)
iii	検索型	入力ファイル数多	中～大 (ファイル数多)	小～中	無

本報告では，表 3 に示す通り，アプリケーションを (i) 時系列シミュレーション型，(ii) 最適化問題型，(iii) 検索型 に分類した．このうち(i)と(ii)は従来の HPC アプリケーションが当てはまる．(iii)は生命科学分野の特にゲノム解析処理のアプリケーションが当てはまる.

ゲノム解析処理などのアプリケーションは，従来 HPC アプリケーション(基本的に大規模なシーケンシャルアクセス)と異なる I/O のふるまいを示す．今後，次世代大規模 HPC システムをビッグデータ処理基盤にも活用していく潮流が考えられるが，その際，重要性が増す．(iii) 検索型に分類するゲノム処理系のアプリケーションは，その I/O ふるまいを別報する.

3.4 基本構成モデル

階層化ストレージシステムの構成案策定を具体化するため，図 2 で示した概念図と，前節で示した I/O 特性(A)～(C)に基づき，階層化ストレージシステムの基本構成モデルを定める(図 4).

この基本構成モデルの狙いは次の通りである.

・消費電力削減のため，基本はオンライン(第 2 階層)とオフライン(第 3 階層)の二階層構成.
・データ量に比較して負荷の高い I/O を受けとめるための，キャッシュとしての第 1 階層を配置．ただし，演算ノード直結ストレージなど，共有範囲は限定.

以降，アプリケーション I/O 要件を反映し，各階層のストレージ容量(S1～S3)と，各層間の必要スループット(Ba～Bd)を明確化することで階層化ストレージシステムの構成案を策定する.

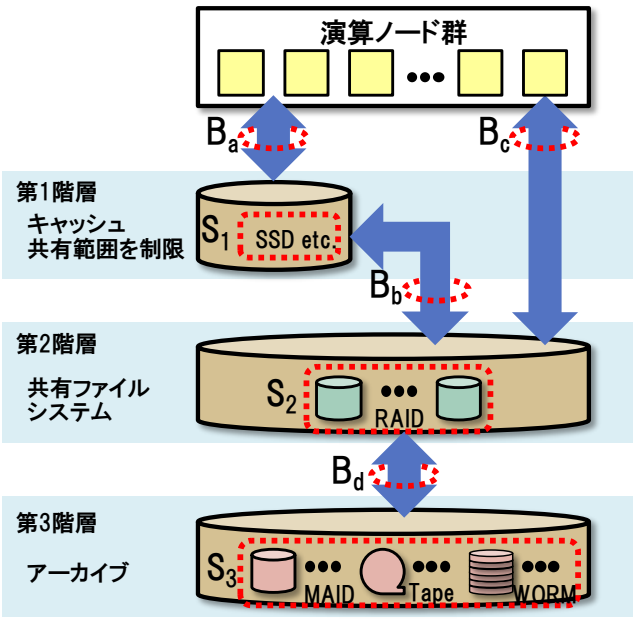

図 4　階層化ストレージシステムの基本構成モデル

3.5 容量およびスループット算出の考え方

ここでは，図 4 で示した基本構成モデルのストレージ容量(S1～S3)と，各層間の必要スループット(Ba～Bd)の明確化に向け，ストレージ容量およびスループット算出方法の考え方を述べておく.

(1) ストレージ容量

ストレージ容量最適化としては，各階層で同時期に存在する可能性のあるファイルを定義し，それらのファイルを格納可能な容量を算出することで，各階層で最低限必要なストレージ容量として定義する.

ここでは，実行中のジョブのデータ，および実行後のジョブのデータ，実行準備中のジョブのデータが存在すると仮定し，次の七つのデータを各階層で同時期に存在する可能性のあるファイルとした.

①実行中のジョブの出力データ
②実行中のジョブのチェックポイントデータ
③実行中のジョブの一つ前のチェックポイントデータ
④直前に終了したジョブの出力データ
⑤直前に終了したジョブの最終チェックポイントデータ
⑥次に実行予定のジョブの入力データ
⑦解析中(または，アーカイブとの間で転送中)の過去の出力データ

これら①～⑦のデータについて，第 1 階層と第 2 階層を利用する場合と，第 2 階層のみを利用する場合で，配置をどのようにするかを図 5 に示す.

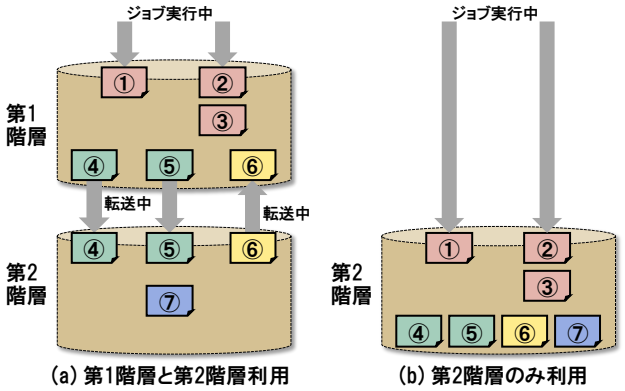

図 5　各階層で同時期に存在する可能性のあるファイル

各階層で最低限必要なストレージ容量を，図で示したデータ配置に基づき，次のように算出する．

◆第 1 階層と第 2 階層を利用する場合
・第 1 階層：①+②+③+④+⑤+⑥
・第 2 階層：④+⑤+⑥+⑦
◆第 2 階層のみ利用する場合
・第 2 階層：①+②+③+④+⑤+⑥+⑦

(2)スループット

(a) I/O 時間の仮定

前述の通り，I/O 時間としては演算時間の 10%が許容されると仮定する．つまり，ジョブの実行時間は CPU 時間の 1.1 倍の時間となる．これは，ジョブ実行時間において，演算時間と I/O 時間のオーバーラップは考えないことを意味する．実際には OS やライブラリ内にバッファがあり，オーバーラップを期待できる場合もあるが，HPC 用途ではメモリを使いきるアプリケーションが多く，この効果に過剰に期待できないと判断している．

なお，ここで仮定した「10%」については，アプリケーションによって要求が様々であり，今後必要に応じて適宜見直す必要がある．

(b) 上位階層がある場合の下位階層のスループット

(2)に書いた通り，I/O 時間としては，演算時間の 10%分が許容されると仮定している．このとき，上位階層(例えば第 1 階層)がある場合に，下位階層(例えば第 2 階層)に必要なスループットは次のように考えることができる．

10%の I/O 時間で上位階層に出力されたデータは，残りの演算時間のうちに下位階層へ転送し終えればよい．つまり，上位階層-下位階層間に必要なスループットは，上位階層のスループットの 1/10 でよい．

さらに，上書き可能なデータ(チェックポイントデータ，および，最適化問題型の途中段階の結果出力)については，最終データのみ下位階層に転送すればよいので，上位階層-下位階層間のスループットはさらに低くて良い．図 3 のデータアクセスモデルでは，1/(10×イテレーション数)となる．

3.6 アプリケーション I/O 特性を考慮した構成案策定方法の提案

これまでに本章で述べてきたことを踏まえて，階層化ストレージシステムの容量およびスループット最適化するための，構成案策定方法を提案する．

アプリケーションの I/O 要件は，次世代大規模 HPC システムのノード構成やアプリケーションプログラムの実行シナリオにより変化する．そのため，本構成案策定方法では，それらが変化した場合に対応可能なように，最低限必要な容量，スループットの算出式を明確化する．

ここでチェックポイント出力についての考え方を示す．アプリケーション I/O の俯瞰調査の結果からは，チェックポイント出力がスループット要件に与える影響は無視できない．チェックポイント出力は，表 2 に示した通り，ジョブ実行を再開できる情報を取得する処理である．このジョブ実行再開のための情報は，基本的にシステムの信頼性が十分で，運用上の制限(24 時間以上のマシン割り当て不可など)がない場合，出力する必要がない．チェックポイントの取得の要否や頻度は運用と合わせて考える必要があり，現状の大規模 HPC システムにおける運用に合わせても必ずしも望ましい結果が得られるわけではない．従って，本報告ではチェックポイントを現状の大規模 HPC システムと同様の頻度で取得するケースと，チェックポイントを取得しないケースの二種を算出する．

構成案策定のための，容量とスループットの算出に利用するパラメータを表 4 に示す．

表 4　容量および性能構成案策定向けパラメータ

項目		初期入力	計算結果出力	チェックポイント出力	結果+チェックポイント出力
データ量[PB]	総和	A	B*m	C*n	-
	1 回分	↑	B	C	-
スループット[GB/s]		X	Y	Z	***Y+Z***

X = Y (∵B*m >> A とみなして Read 性能は考えない)
Y = (B*m)／(演算時間の 10%)
m: 結果出力を行うイテレーション回数
Z = (C*n)／(演算時間の 10%)
n: チェックポイント出力を行うイテレーション回数
Y+Z = (B*m+C*n)／(演算時間の 10%)

算出のために必要なパラメータとして，「初期入力」「計

算結果出力」「チェックポイント出力」の値に加え，計算結果とチェックポイントの出力を合わせた「結果+チェックポイント出力」の値を挙げている.「データ量」に関しては，表 3 の「上書き可」データに関しては「1 回分」の値を，「上書き不可」データに関しては「総和」の値を用いる.「結果+チェックポイント出力」のデータ量は，「計算結果出力」と「チェックポイント出力」の和をとればよいので，表 4 には値を入れない.

表 4 のパラメータと，3.5 節で示した考え方を利用して，階層化ストレージシステム各階層の必要容量および必要スループットを算出する．具体的には，各階層のストレージ容量(S1～S3，ただし S2 と S3 は合計値)と，各層間の必要スループット(Ba～Bd)を，次の 5 パターンで算出する.

◆パターン 1: 第 1 階層で全 I/O を処理.
◆パターン 2: パターン 1 と同様に，第 1 階層で全 I/O を処理．ただし，チェックポイント出力を含めない.
◆パターン 3: 第 2 階層で全 I/O を処理.
◆パターン 4: パターン 3 と同様に，第 2 階層で全 I/O を処理．ただし，チェックポイント出力を含めない.
◆パターン 5: ハイブリッド型，上書き可な出力を第 1 階層，上書き不可で蓄積すべき出力を第 2 階層で処理．各階層では次を I/O 対象とする.
・第 1 階層…最適化問題型アプリの計算結果出力，および，チェックポイント出力
・第 2 階層…時系列シミュレーション型アプリの計算結果出力

3.3 節のアプリケーション分類で示した通り，時系列シミュレーション型と最適化問題型では，計算結果出力の上書き可否により，必要容量などが変わる．時系列シミュレーション型アプリケーションと最適化問題型アプリケーションの必要容量および性能の算出式を，各々表 5 と表 6 に示す.

表 5　時系列シミュレーション型アプリケーションの必要容量および性能の算出式

パターン	1	2	3	4	5
チェックポイント	有	無 (C=0, Z=0)	有	無 (C=0, Z=0)	有
Ba	***Y+Z***		-		Z
S1	A+B*m*2+C*3		-		C*3
Bb	(Y+Z/n)/10		-		(Z/n)/10
Bc	-		***Y+Z***		Y
S2+S3	A+B*m*2+C		A+B*m*3+C*3		A+B*m*3+C
Bd	Y/10		Y/10		Y/10

表 6　最適化問題型アプリケーションの必要容量および性能の算出式

パターン	1	2	3	4	5
チェックポイント	有	無 (C=0, Z=0)	有	無 (C=0, Z=0)	有
Ba	***Y+Z***		-		***Y+Z***
S1	A+B*2+C*3		-		B*2+C*3
Bb	(Y+Z/n)/10		-		(Y/m+Z/n)/10
Bc	-		***Y+Z***		-
S2+S3	A+B*2+C		A+B*3+C*3		A+B*2+C
Bd	Y/10		Y/10		Y/10

これら算出式に基づいて，アプリケーション I/O のふるまい，および，チェックポイント有無など運用条件に合わせて，個々のアプリケーションごとに値を算出する．アプリケーションごとに算出した，各階層のストレージ容量(S1, S2+S3)と，各層間のスループット(Ba～Bd)の最大値を組み合わせることで，算出に用いたアプリケーションの I/O 要件を反映した構成案の策定を可能にする.

4. 評価

ここでは，次世代大規模 HPC システムにおいての実行が想定されるアプリケーションについて，I/O のふるまいの調査を行い，得られた結果に提案方法を適用して求めた階層化ストレージシステム構成案を示す．その構成案の消費電力量を見積もることで，提案方法の評価を行う.

4.1 アプリケーション I/O の調査

対象とするアプリケーションとしては，現在京コンピュータで実行されており，FS の研究題目「レイテンシコアの高度化・高効率化による将来の HPCI システムに関する調査研究」[6]において評価対象として用いられるアプリケーション(強相関量子格子モデル計算 ALPS[11-12]，実空間第一原理分子動力学計算 RSDFT[13-14], 全球雲解像大気大循環モデル計算 NICAM[15]，超高解像度海洋大循環モデル COCO[16])を選択した.

これらアプリケーションに加えて，同 FS の研究課題「将来 HPCI のあり方調査研究 アプリケーション分野」において評価対象となっている種々のアプリケーションについてもヒアリングシートによる調査依頼も実施しているが，ここではその値を用いていない.

このアプリケーション I/O 要件の一部を表 7 に示す.

表 7　アプリケーション I/O 要件

アプリケーション		COCO 大規模	ALPS 大規模	NICAM (アンサンブル)
	分類	時系列シミュレーション型	最適化問題型	時系列シミュレーション型
容量	初期入力	0.1PB	0.0001PB	0.27PB
	計算結果	172.3PB	7.9PB	0.98PB
	チェックポイント	0.04PB	0.79PB	0.25PB
スループット	計算結果	629GB/s	11GB/s	814GB/s
	チェックポイント	573GB/s	2,185GB/s	205GB/s
	計算結果+チェックポイント	1,203GB/s	2,195GB/s	1,018GB/s

4.2 階層化ストレージシステム構成案と評価

アプリケーション I/O 要件(表 7)を，アプリケーション分類の時系列シミュレーション型および最適化問題型について，表 4 に示した形で，データ量およびスループット要件をまとめた結果を，各々表 8 と表 9 に示す.

表 8　時系列シミュレーション型アプリケーションの I/O 要件

項目		初期入力	結果出力	チェックポイント出力	結果+チェックポイント
データ量[PB]	総和	0.27	172.3	-	-
	1 回分	↑	0.047	0.25	-
スループット [GB/s]		→	814	573	1,203

表 9　最適化問題型アプリケーションの I/O 要件

項目		初期入力	結果出力	チェックポイント出力	結果+チェックポイント
データ量[PB]	総和	0.011	-	-	-
	1 回分	↑	0.89	0.79	-
スループット [GB/s]		→	206	2,185	2,195

さらにこれら結果から，表 5 および表 6 に示した形で，容量および性能最適化構成を算出した結果を，各々表 10 と表 11 に示す.

表 10　時系列シミュレーション型アプリケーションの必要容量およびスループットの算出結果

パターン	1	2	3	4	5
チェックポイント	有	無	有	無	有
Ba [GB/s]	1,203	814	-	-	573
S1 [PB]	346	345	-	-	1
Bb [GB/s]	120	81	-	-	57
Bc [GB/s]	-	-	1,203	814	814
S2+S3 [PB]	345	345	518	517	517
Bd [GB/s]	81	81	81	81	81

表 11　最適化問題型アプリケーションの必要容量およびスループットの算出結果

パターン	1	2	3	4	5
チェックポイント	有	無	有	無	有
Ba [GB/s]	2,195	206	-	-	2,185
S1 [PB]	4	2	-	-	4
Bb [GB/s]	220	21	-	-	219
Bc [GB/s]	-	-	2,195	206	206
S2+S3 [PB]	3	2	5	3	3
Bd [GB/s]	21	21	21	21	21

これら算出した必要容量およびスループットの値を集約し，図 4 の基本構成モデルに当てはめた結果を図 6 に示す.

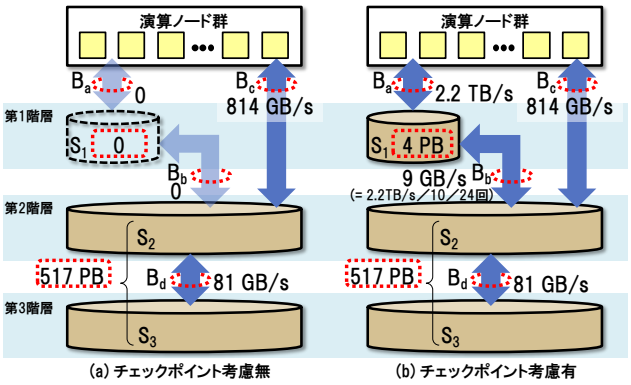

図 6　算出値の集約結果

この集約結果をベースに，表 1 に示した 2018 年頃に想定されるストレージ装置セットの諸元から，容量およびスループットの要件に基づき，必要なセット数(表 1 で示したストレージ装置セット)を求め，階層化ストレージシステム構成案として策定する．結果を図 7 に示す.

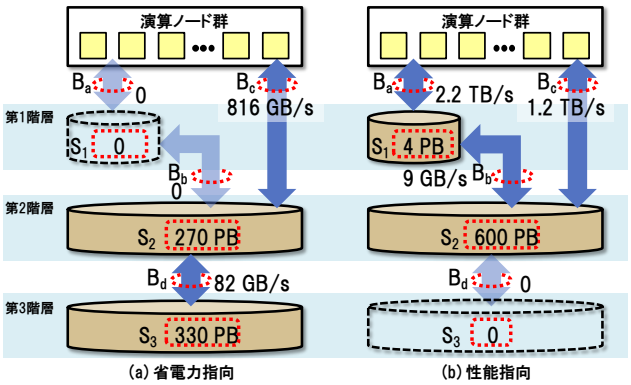

図 7　階層化ストレージシステムの構成案

構成案の算出根拠は次の通りである．

◆省電力指向構成

第 2 階層は図 6 の集約結果よりスループット 814GB/s が要件である．容量要件は第 2 階層と第 3 階層を合わせて 517PB だが，余裕を見て 600PB とする．本構成では第 2 階層の容量を可能な限り減らし，第 3 階層で補う方針とする．

・第 2 階層

スループットが構成決定要因となるため，RAID 装置は表 1 の「容量非重視」を選択．

セット数：814GB/s / 16GB/s = 51 セット

容量：5.3PB * 51 セット = 270PB

スループット：16GB/s * 51 セット = 816GB/s

・第 3 階層

第 2 階層の算出結果より，第 3 階層の容量は 330PB.

容量：600PB – 270PB = 330PB

セット数:81GB/s / 2GB/s = 41 セット(スループット基準)

330PB / 8.4PB = 40 セット(容量基準)

→41 セット(スループット基準で構成決定).

スループット：2GB/s * 41 セット = 82GB/s

◆性能指向構成

第 1 階層で上書き可なデータ(チェックポイントと最適化問題型アプリケーションの計算結果)を格納できる容量，スループットを確保する．また，第 2 階層では，表 7 のアプリケーションのうち，データ量が最大である COCO の出力(計算結果+チェックポイント)を格納可能にする方針とする．COCO の必要スループットは 1.2TB/s であるため，これを第 2 階層の要件とする．第 2 階層と第 3 階層を合わせた容量は省電力指向構成と同様に 600PB とする．

・第 2 階層

セット数：1.2TB/s / 16GB/s = 75 セット

容量： 容量重視 RAID 10.7PB * 75 セット = 802.5PB

容量非重視 RAID 5.3PB * 75 セット = 397.5PB

→600PB をカバーするように RAID の搭載ディスク数を調整し．第 3 階層は不要にする．

これら構成案について，容量とスループットの要件をそれぞれ横軸と縦軸にとり，アプリケーションの I/O 要件をプロットし，カバー範囲を示した結果が図 8 である．例えば，省電力指向構成では，チェックポイント出力を要件にいれなければ，今回評価対象にしたアプリケーションの I/O 要件を満たすことができることがわかる．また，チェックポイント出力を含めてすべてのアプリケーションの I/O 要件を満たすには性能指向構成が必要になることがわかる．

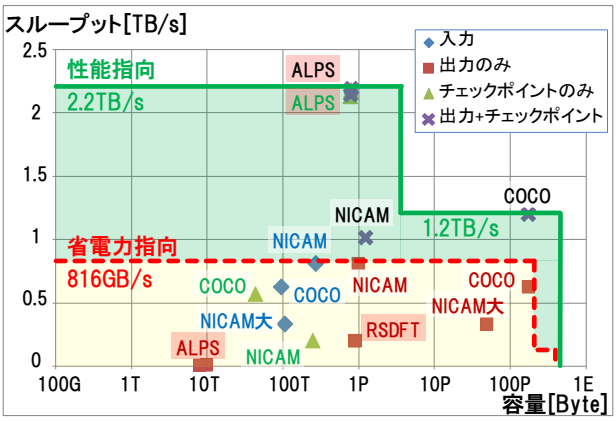

図 8　アプリケーションプログラムの I/O 要件と階層化ストレージシステム構成案のカバー範囲

最後に，図 7 で示した各構成案について，表 1 に示した 2018 年頃に想定されるストレージ装置セットの諸元から，消費電力を算出した結果を表 12 に示す．当初想定した 1EB の容量を全て RAID 装置で構成した場合の消費電力も比較のために示した．

表に示す通り，アプリケーションプログラムの I/O 要件分析と，それに基づく構成案を策定することにより，1EB フル RAID 構成に比較して，省電力指向構成では，消費電力を約 66%削減する．また，性能指向構成でも消費電力を約 50%削減する．システム全体の消費電力に対するストレージシステムの消費電力の割合は，システム全体の消費電力を 20MW とした場合でも，省電力指向構成で約 5%，性能指向構成で約 7.4%にすることが可能で，目標の 10%以内に抑える見込みを得た．

表 12　階層化ストレージシステム構成案の消費電力

階層	項目	構成案		(参考)
		省電力指向	性能指向	1EB フル RAID
第 1 階層	スループット	0	2.2TB/s	-
	容量	0	4PB	-
第 2 階層	スループット	816GB/s	1.2TB/s	3TB/s
	容量	270PB	600PB	1EB
第 3 階層	スループット	82GB/s	0	0
	容量	330PB	0	0
消費電力(第 2+3)		1,006KW	1,476KW	2,967KW

66%減　50%減

5. 関連研究

最初に，階層化ストレージシステムを制御する階層記憶管理(HSM: Hierarchical Storage Management)技術に関して述べる．

HSM については，X/Open で策定された仕様である DMAPI (Data Management API)[17]が代表的である．DMAPI ではフロントとなるファイルシステムと HSM 間のインタフェースを定義している．しかし，ファイルシステムイベントの取得が同期的であるため高負荷環境でイベント処理がネックとなる可能性が高い．

Hazen 等は, DMAPI 実装で課題になった点について, ILM (Information Lifecycle Management)機能を利用した並列メタデータ検索を採用して非同期化する等，スケーラビリティで改善を図っている[18]. HSM 技術は Lustre®ファイルシステムでも Ver.2.5 でも実装され[19]，イベント監視は，ファイルの変更ログを非同期に参照する方法を採っている．本報告では HSM 実装方法については，触れていないが，実装に当たっては同様の考慮が必要である．

次に, 階層化ストレージシステムの活用に関して述べる．

藤本等は，HPC のセンタ運用において，ニアラインストレージと位置付ける MAID (Massive Arrays of Idle Disks)を活用し，ジョブスケジューラと連携したステージングを行うことで，システムの消費電力を削減する方法を提案している[20]．ジョブ開始前にステージングを完了するための予測成功率の期待値に基づき，必要なオンラインストレージの容量を示すなど，本報告の構成策定方法と通ずる点がある．ステージングをボリューム単位で実施するなどが本報告との相違点である．

上村等は, 階層化ストレージシステムの活用技術として, Delegation 機能を持つマイクロカーネル上でのユーザレベル I/O や階層間データ転送を隠ぺいする方式を提案している[21]．本報告で示した構成策定方法を適用した階層化ストレージシステムを活用し, 次世代大規模 HPC システムの I/O 系において，性能，および，使い勝手を改善する研究である．

6. おわりに

2018 年頃のポストペタスケールシステムを視野に入れた, 次世代大規模 HPC システムの実現を目指した検討が進んでいる. 我々は次世代大規模 HPC システムで実現すべきストレージシステムの明確化に取り組んだ．

次世代大規模 HPC システムでは, 消費電力がその設計や実効性能を制約する最大の要因の一つであり，ストレージシステムに限っても同様に問題となる．次世代大規模 HPC システム全体の消費電力の想定は約 20～30MW である. 消費電力は可能な限り低く抑えたいため，システム全体の消費電力を 20MW とした場合に, ストレージシステムの消費電力をその 10%に当たる 2MW 以内に抑えることを目標とした．

ストレージシステムの消費電力低減には，RAID 装置などいわゆるオンラインストレージに加えて，テープライブラリなど消費電力当たりの容量効率がよいオフラインストレージを積極的に活用する階層化ストレージシステムが有効である. 次世代大規模 HPC システムにおける階層化ストレージシステム適用実現に向けて次の二点を課題とした．

(1) 容量および性能最適化
(2) ストレージ階層のトランスペアレント化

本報告では，「(1)容量および性能最適化」にフォーカスして検討した．具体的には，アプリケーションの I/O 特性に基づき, 階層化ストレージシステムの構成案(容量およびスループット性能)策定方法を提案した．

構成案策定に利用するため，アプリケーションを (i) 時系列シミュレーション型，(ii) 最適化問題型，(iii) 検索型に分類した．提案する構成案策定方法では，このうち(i)と(ii)に分類されるアプリケーションについて, 計算結果出力とチェックポイント出力の上書き可否に着目した出力先ストレージの使い分けを行うことが特長である．

提案方法を適用して, 次世代大規模 HPC システムにおいて実行が想定されるアプリケーションの I/O 特性に基づき, 省電力指向および性能指向の二つの構成案を策定した．これら構成案では，消費電力を，1EB フル RAID 構成の 2,967KW と比較して，省電力指向構成で約 66%削減した 1,006KW, 性能指向構成で約 50%削減した 1,476KW に低減可能な見通しである．

この結果, 次世代大規模 HPC システム全体の消費電力を 20MW とした場合に，システム全体の消費電力に対するストレージシステムの消費電力の割合を，省電力指向構成で約 5%，性能指向構成で約 7.4%にすることが可能となることを確認し，目標の 10%以内に抑える見込みを得た．

今後は，課題として挙げた「(2) ストレージ階層のトランスペアレント化」に関して，別途検討中の HSM 実装方式の詳細化を行う予定である．また，I/O に着目したアプリケーションの分類で「(iii) 検索型」に分類したアプリケーションについて，I/O ふるまいの詳細調査の結果に基づき, 次世代大規模 HPC システムにおける課題の明確化と解決方式の検討を実施する予定である．

謝辞 本研究の一部は，文部科学省「将来の HPCI システムのあり方の調査研究」の研究課題「レイテンシコアの高度化・高効率化による将来の HPCI システムに関する調査研究」による．研究にあたり検討に参加いただいたメンバに，謹んで感謝の意を表する．

参考文献

[1] 宮崎博行，草野義博，新庄直樹，庄司文由，横川三津夫，渡邊貞 : スーパーコンピュータ「京」の概要，FUJITSU，Vol.63，No.3，PP.237-246 (2012-5).
[2] 黒川原佳 : 京速コンピュータ「京」，理研シンポジウム(2011).
http://accc.riken.jp/secure/4721/kurokawa-riken.pdf
(2014/6/20 アクセス)
[3] HPCI 計画推進委員会 今後の HPC 技術の研究開発のあり方を検討する WG アプリケーション＆コンピュータアーキテクチャ・コンパイラ・システムソフトウェア合同作業部会 : 今後の HPCI 技術開発に関する報告書(2012-3).
http://open-supercomputer.org/wp-content/uploads/2012/03/FutureHPCI-Report.pdf (2014/6/20 アクセス)
[4] アプリケーション作業部会：計算科学ロードマップ白書 (2012-3).
http://open-supercomputer.org/wp-content/uploads/2012/03/science-roadmap.pdf (2014/6/20 アクセス)
[5] コンピュータアーキテクチャ・コンパイラ・システムソフトウェア作業部会：HPCI 技術ロードマップ白書(2012-3).
http://open-supercomputer.org/wp-content/uploads/2012/03/hpci-roadmap.pdf (2014/6/20 アクセス)
[6] 石川裕，平木敬，青柳睦，新庄直樹，飯田恒雄，中村祐一 :「レイテンシコアの 高度化・高効率化による将来の HPCI システ ムに関する調査研究」報告，SDHPC (2014).
http://www.open-supercomputer.org/wp-content/uploads/2014/01/sdhpc11-ishikawa.pdf (2014/6/20 アクセス)
[7] JEITA 情報システム標準化委員会 テープストレージ専門委員会：テープシステム技術資料，第 4 章 データ転送速度(2011-4).
http://home.jeita.or.jp/is/committee/tech-std/std/201104/tape_system_04.pdf (2014/6/20 アクセス)
[8] SCSI Trade Association: Serial Attached SCSI Master Roadmap, SAS Master Roadmaps, Jun 2011.
http://www.scsita.org/library/sas-master-roadmaps/
(2014/6/20 アクセス)
[9] FCIA - Fibre Channel Industry Association: Fibre Channel Roadmaps v1.8.
http://www.fibrechannel.org/fibre-channel-roadmaps.html
(2014/6/20 アクセス)
[10] IBTA - InfiniBand Trade Association: InfiniBand Roadmap.
http://www.infinibandta.org/content/pages.php?pg=technology_overview (2014/6/20 アクセス)
[11] S. Todo, K. Kato: Cluster Algorithms for General-S Quantum Spin Systems, Phys. Rev. Lett. 87, 047203 (2001).
[12] B. Bauer, L. D. Carr, A. Feiguin, J. Freire, S. Fuchs, L. Gamper, J. Gukelberger, E. Gull, S. Guertler, A. Hehn, R. Igarashi, S.V. Isakov, D. Koop, P.N. Ma, P. Mates, H. Matsuo, O. Parcollet, G. Pawlowski, J.D. Picon, L. Pollet, E. Santos, V.W. Scarola, U. Schollwoeck, C. Silva, B. Surer, S. Todo, S. Trebst, M. Troyer, M.L. Wall, P. Werner, S. Wessel: The ALPS project release 2.0: Open source software for strongly correlated systems, J. Stat. Mech. P05001 (2011).
[13] Y. Hasegawa, et al.: Performance evaluation of ultra-largescale first-principles electronic structure calculation code on the K computer, Published online before print October 17, 2013, doi:10.1177/1094342013508163 International Journal of High Performance Computing Applications October 17, 2013 1094342013508163.
[14] J.-I. Iwata, et al.: A massively-parallel electronic structure calculations based on real-space density functional theory, Journal of Computational Physics 229, 2339-2363 (2010).
[15] M. Satoh, T. Matsuno, H. Tomita, H. Miura, T. Nasuno, S. Iga: Nonhydrostatic Icosahedral Atmospheric Model (NICAM) for global cloud resolving simulations, Journal of Computational Physics, the special issue on Predicting Weather, Climate and Extreme events, 227, 3486-3514 (2008), doi:10.1016/j.jcp.2007.02.006.
[16] H. Hasumi: CCSR Ocean Component Model (COCO) Version 4.0, CCSR Report 25 (2006),
available from http://ccsr.aori.u-tokyo.ac.jp/~hasumi/COCO/coco4.pdf (2014/06/26 アクセス)
[17] The Open Group: Systems Management: Data Storage Management (XDSM) API, CAE Specification, Feb 1997.
http://pubs.opengroup.org/onlinepubs/9695979099/toc.pdf
(2014/6/20 アクセス)
[18] Damian Hazen, Jason Hick: GPFS HPSS Integration: Implementation Experiences, LBNL, Sep 2008.
[19]OpenSFS: Lustre® File System Version 2.5 Released, Nov 2013.
http://opensfs.org/press-releases/lustre-file-system-version-2-5-released/
(2014/6/20 アクセス)
[20] 藤本和久，赤池洋俊，三浦健司，村岡裕明：高速ストレージサブシステムの省電力化に向けた新規データ配置制御技術，日本磁気学会 第 177 回研究会，PP.51-58 (2011).
[21] 上村佳史，酒井憲一郎，樋口雄太，小田和友仁，住元真司，宇野俊司，清水正明，石川裕：エクサスケールシステムに向けたファイルシステムデザインの検討，情報処理学会第 142 回 HPC 研究会，No.9 (2013).