サイエンティフィック・システム研究会 2011 年 10 月 19 日
2011 年度 科学技術計算分科会 会合

エクサバイト規模のストレージシステムへ向けて

佐々木 節

高エネルギー加速器研究機構 計算科学センター

[アブストラクト]

高エネルギー加速器研究機構では、KEKB加速器を用い、Belle実験がB中間子のCP保存則の破れを観測することにより、小林・益川両先生のノーベル賞受賞につながる成果を得た。さらに統計精度を上げ、精密測定を行うとともに、素粒子の標準理論を超える新たな物理学へ向け、KEKB加速器をアップグレードしたSuper KEKB加速器の開発が進められている。Super KEKB加速器を用いて行われるBelle II実験においては、2020年には200PB程度のデータを収集すると見込まれている。同程度の統計量のモンテカルロシミュレーションを行うことを考慮すると、エクサバイト規模のストレージステムが必要になる。KEKにおける計算機システムの歴史を紹介するとともに、将来に向けた取り組みについて議論する。

[キーワード]

ストレージ、ハードディスク、テープ、階層化ストレージ管理システム、分散並列ファイルシステム

1. はじめに

茨城県つくば市にある高エネルギー加速器研究機構(KEK)では、長年にわたり、複数の加速器を用いた様々な研究が行われてきた。原子核・素粒子物理学の実験的研究ばかりではなく、物質科学、生命科学など様々な分野の研究が行われている。PF(Photon Factory)や、J-PARC のビームラインは、広く産業界にも利用されており、物質の構造の解明に用いられている。計算科学センターは、データを記録し、解析を行うために必要な計算環境の提供を行っている。

KEK で最大の計算資源需要があるのは、Belle II 実験である。前身となる Belle 実験は、小林・益川理論を裏付ける B 中間子の CP 破れの観測に成功し、両先生のノーベル賞受賞につながった。Belle 実験に用いられた KEKB 加速器を高度化した Super KEKB 加速器の開発が開始されており、現在の 40 倍以上の強度となる予定である。Belle 実験も検出器をアップグレードし、新たなメンバーを諸国から受け入れた Belle II 実験が行われる。観測データのみでも、2020 年には 200PB に達すると考えられており、シミュレーションデータ、2 次データを考慮に入れると、エクサバイト規模のストレージシステムが必要になると考えられる。

一方、2008 年に共用が開始された J-PARC においても、年間 2PB 程度のデータを収集すると見積もられている。J-PARC は、陽子線加速器のコンプレックスであり、原子核・素粒子、物質科学、生命工学など様々な目的に利用されている。Photon Factory の運転も続いており、大量ではないが、データの収集が継続されている。

計算科学センターは、プロジェクト毎に異なる最大時 5 システムを平行して運用を行っていた。それらを統合し、中央計算機とスーパーコンピュータの 2 システムへの統合がまもなく終了する予定である。新中央計算機は、2012 年春に稼働を開始する予定であるが、テープライブラリの容量は16B、ユーザが利用可能なディスク領域は3.5PBとなっている。ここに至るまでの課程を以下に紹介するとともに、エクサバイト規模のストレージシステム開発向けた取り組みについて紹介する。

2. KEK における計算機システムの歴史

KEK のデータ解析用システムは、以前は、最大時で4システムあったものを徐々に統合を進めるとともに、時代に即したアーキテクチャの採用を行ってきた。メインフレームに代わる商用 UNIX による分散環境の構築が 1992 年頃から開始され、2000 年代中頃には、Linux をユーザ環境とするシステムへの置き換えが進み、現在に至っている。

図1にデータ解析システムのユーザが利用可能なディスク領域の量とテープライブラリの容量の伸びを示す。KEKB 加速器、J-PARC の運転開始後、データ需要が急速に伸びたことが見て取れる。

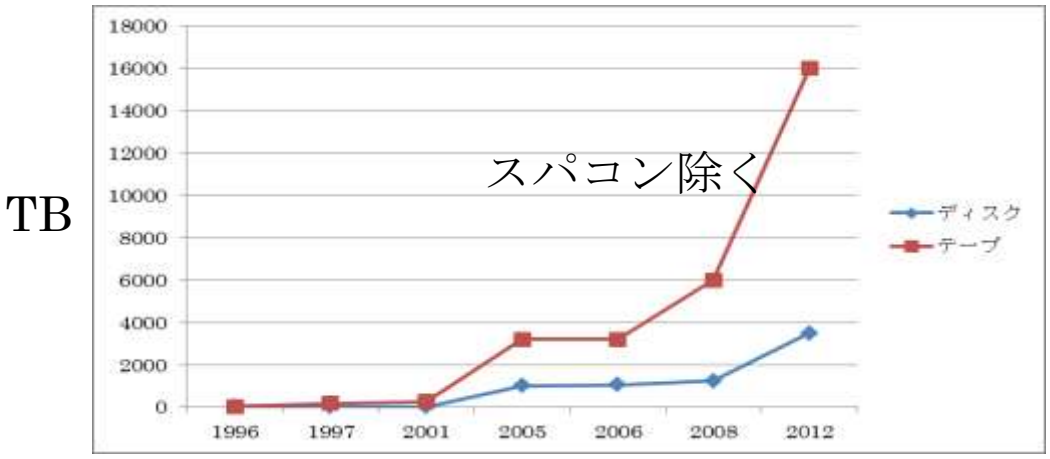

図1 KEK におけるデータ容量の推移

テープ利用の利便性を改善するために、階層化ストレージ管理システムを 1995 年頃から利用している。SONY の Petesite と DoE 研究所共同開発の HPSS が平行して用いられてきたが、新中央計算機システムでは、HPSS に一本化される。ディスクアクセスの高速化と、シームレスにシングルファイルツリーをユーザに提供するために、GPFS が 2008 年から利用されている。

3. エクサバイト規模のストレージに向けて

予定通り加速器の建設と実験が進めば、次回のシステム入れ替え時である 2015 年に 4 年で計算機のレンタルを行うとすると、200～300PB 程度のストレージ容量が必要となる。夏季の加速器シャットダウン中の再解析、また、その次のシステムへのデータ移行を考えると、100GB/sec 程度の総転送速度が要求される。予算の緊縮が続くなか、新たな技術を開発して、困難を克服する必要がある。

中央計算機システム導入のための市場調査で、階層化ストレージ管理システムと分散並列ファイルシステムを連携させたソリューションを国内外に複数確認することができたが、テープを必要とするほど大量のデータを持つ機関の数は RAID の大容量化と共に減っていると考えられており、今後も慎重に動向を見守る必要がある。

国内には、KEK 以外にも、理研、JAXA、国立天文台など大量のデータを抱える機関が多くあり、これらの機関とも情報の交換をし、必要な技術の開発を行っていきたいと考えている。この目的で、Data Intensive Computing 研究会を 2011 年度に私的任意団体として設立した。隔月で研究を開き、お互いの情報交換を行っている。個人的には、DoE 参加の研究所が共同開発を行っている HPSS を参考に、研究機関横断で共通の技術仕様を作成し、大規模ストレージを管理し利用するために必要な技術の開発を協力して行えないかと考えている。

[参考文献]

(1) 高エネルギー加速器研究機構 http://www.kek.jp

(2) Belle2 http://belle2.kek.jp

(3) 「大規模ストレージの今後の課題と展望」サイエンティフィックコンピューティング研究会 2011 年