COLUMN

高エネルギー実験解析システムへのSANの適用

佐々木　節（文部科学省高エネルギー加速器研究機構 計算科学センター）
Takashi.Sasaki@kek.jp

高エネルギー物理学（HEP: High Energy Physics）実験の分野が直面しているデータ処理の現状と課題を紹介し，SAN（Storage Area Network）による解決を検討する．また，広域分散ファイル共有への応用の可能性について述べる．

高エネルギー物理学実験の現状と課題

HEP実験の分野では，巨大な粒子加速器と測定器，大規模な計算機システムが不可欠となっている．現在，HEP実験のための大型加速器が稼働しているのは，CERN（欧州原子核研究機構，スイス），DESY（ドイツ電子シンクロトロン研究所，独），BNL（ブルックヘブン国立研究所，米），FNAL（フェルミ国立研究所，米），SLAC（スタンフォード大学線形加速器センター，米），そして我が国のKEK（文部科学省高エネルギー加速器研究機構）のみである（**図-1**参照）．それぞれの加速器では，多くの場合，複数の実験グループがそれぞれの検出器を用いて研究活動をしている．全世界で約1万人いるこの分野の研究者（大学院生を含む）は，興味のあるテーマが行われている実験グループに所属し研究を行っている．1つの研究グループは，多人数の組織（1,000人規模のものある）で，お互いが遠隔地の大学や研究所にいながら共同研究を遂行している．また，効率的に研究を進めるための道具作り（データ収集や解析に使うソフトウェアの開発）にも熱心である．今やインターネットの代名詞的として使われるまでに発達したWWW（World Wide Web）も，この分野の研究者間での情報交換の手段としてCERNで開発されたことは有名である．

現在のHEPコンピューティングの大きな課題の1つは，データそのものを世界的な規模で分散共有することである．現状では，加速器のある研究所に出向くか，計算機にリモートアクセスしてデータ解析を行うことが多い．今後，各国に地域解析センターを作り，必要なデータを保持して解析を行うモデルが望まれている[1]．これは，加速器を持っている研究所の必要な計算資源が膨大になり，すべてを提供できなくなっていることも理由の1つである．また，研究者を長期にわたり他機関に駐在させることの難しさ，各機関における学生の教育の必要性も理由とされている．さらに，同じ研究グループの中でさえ厳しい競争があるので，独自のアイディアで国際競争に勝つには，手元にデータがあることが重要であると考えられている．

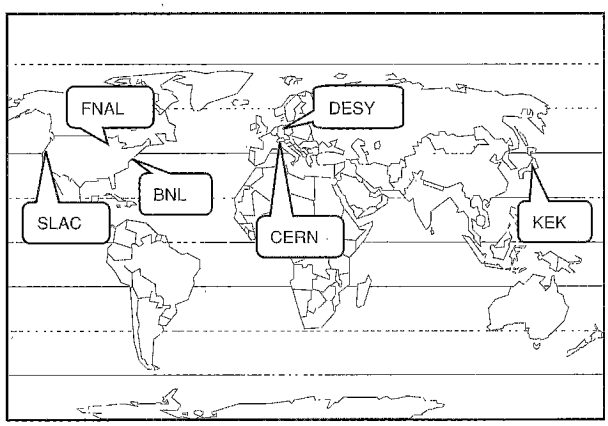


図-1　主要な加速器研究所

地域解析センター構想

各研究機関をより有機的に結び付け，データの共有を図ることで，いわゆるバーチャルグローバルラボラトリーの構築を目指す地域解析センター構想がある．地域解析センターのモデルとしては，**表-1**のような4階層が考えられている．

なお，必要な計算資源の量は，**表-2**のように見積もられている．

現在，KEKで行われているBELLE実験の場合でも年間30TB

Tier 0	加速器を持つ研究所
Tier 1	地域解析センター（各国1または2程度）
Tier 2	大規模大学，研究所
Tier 4	小規模大学など

表-1　地域解析センターの階層

センター	ストレージ	CPU	WAN
Tier 0　CERN	20PB	700K SPECint95	
Tier 1　地域センター	1PB	150K SPECint95	CERNへ1Gbits/sec
Tier 2　大学，研究所等	1TB	20K SPECint95	15 Mbits/sec
Tier 4　その他	未検討	未検討	未検討

表-2　地域解析センターの規模

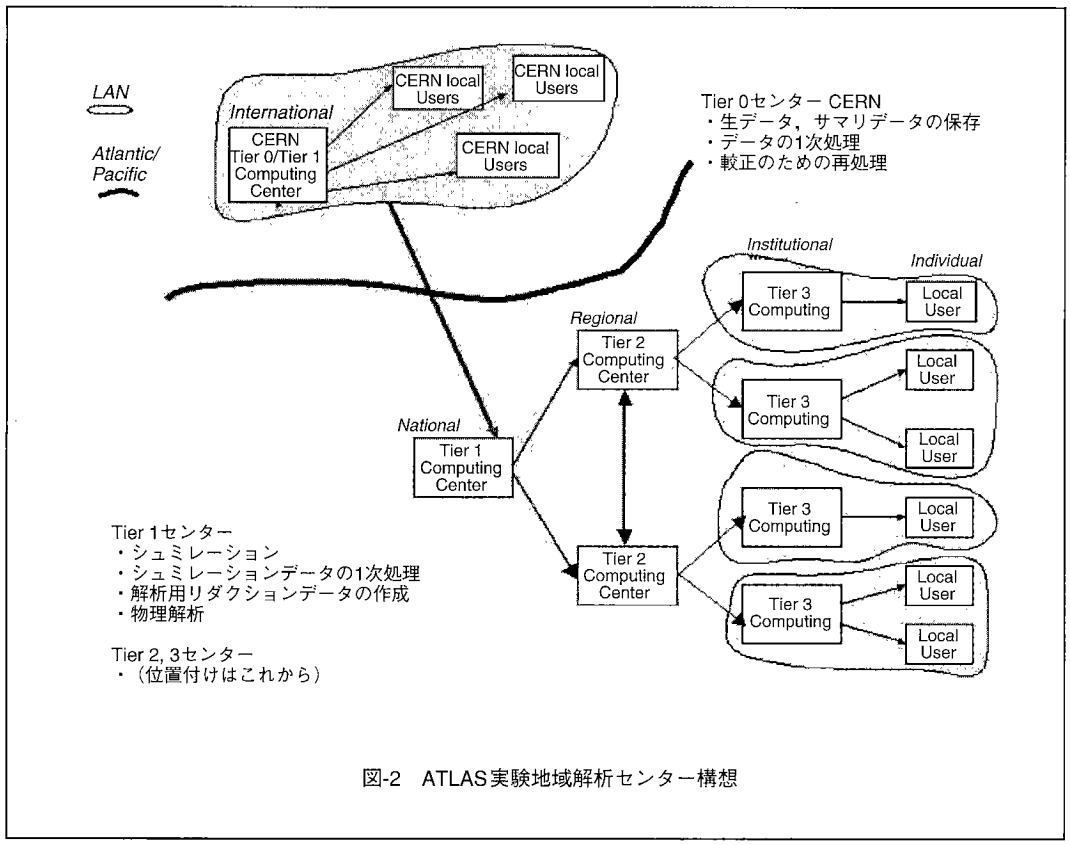


図-2　ATLAS実験地域解析センター構想

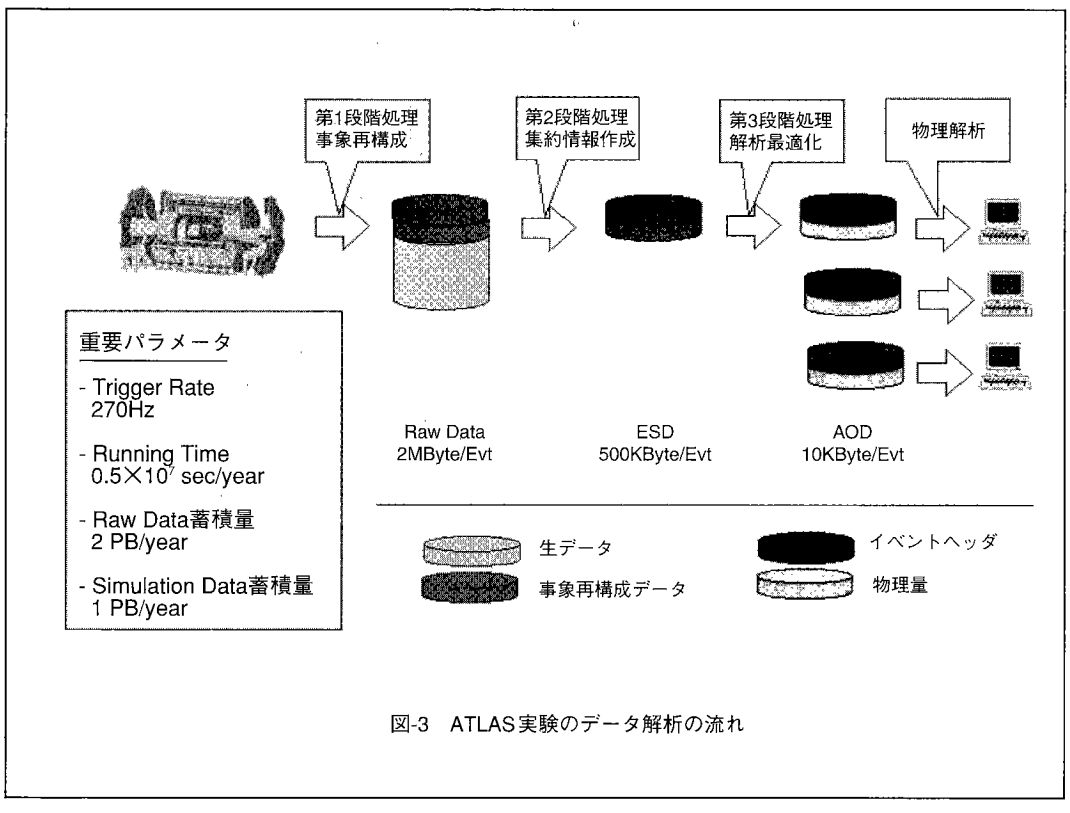
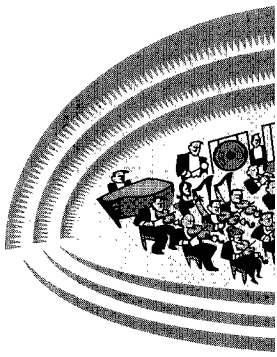


図-3　ATLAS実験のデータ解析の流れ

のデータが収集されている．しかし，2005年にCERNのLHC加速器を用いて開始されるATLAS実験の場合，その量は年間2PBに達する．実験データだけではなくモンテカルロデータも含めると年間3PBとなる．ATLAS実験グループ内で採用されている地域解析センターのモデルを図-2に示す．

ATLAS実験には，日本国内からもKEKをはじめ，多くの研究者が参加しており，国内にもTier 1地域解析センターを1カ所建設することを目標に，技術的な検討を開始し議論を続けている．ATLAS実験でのデータ処理の流れを図-3に示す．

この地域解析センターで用いられるアプリケーションは，いわゆるデータマイニング型である．素粒子同士の反応は，確率事象であり，実際に収集した実験データにはさまざまな種類の事象が含まれている．この中から似



通った性質を持つ事象を選び出し，統計処理を行う．どの種類の素粒子反応によるものかを特定し特徴を調べることで，理論の予言との比較を行う．

実験データの分散共有システムの要件

データベースマネージメント・システム－ODBMS－

HEP実験全体として，オブジェクト指向データベース管理システム（ODBMS）上に実験データを記録し，解析に用いることが検討されている．ATLAS実験ではObjectivity/DBというODBMS製品の採用を予定している．この製品はODMGの定める規格に準拠しており，数百PBまでのスケーラビリティを保証している．ODBMSは商業的に成功しているとはいえず将来性には不安が残るが，機能面で必要な要件を満たす製品がほかに見つかっていない．

Objectivity/DBのアーキテクチャは多くのRDBMSと異なり，ファットクライアント・サーバとなっている．この製品は，OS経由でアクセスできるファイルに記録する．ファイルはローカルディスク上にあっても，分散ファイルシステム上にあってもよい．また，それらの代わりに使用するNFSサーバをスレッド化して高速にした独自の分散I/Oのサーバも提供されている．ロックの完全性を保証するために独自のロックサーバを提供している．ただし，高速化のための特別なファイルシステム等は備えていない．サーバ側の負荷はRDBMSに比べると低く抑えられており，多量のデータに対するスケーラビリティを実現している．

データストレージ・システム－SAN－

KEKでは，Tier 1地域センターに必要とされる大規模なストレージシステムを備えた計算機システムの研究・開発を東京大学素粒子国際センターと共同で行っている．大規模なデータストレージを管理し，高速な分散アクセスを可能にするための手法の1つとしてSANに着目してその適用性の検証を行っている．2005年にはKEKと日本原子力研究所が共同で建設を行うJHFと呼ばれる加速器の稼働も予定されており，ATLAS実験のみならず汎用的に適用できる技術の開発を目指している．なお，ATLASのTier 1地域解析センターの場合，システム全体の総和としては200MB/sec以上の速度でODBMSから必要なオブジェクトを読み出し，40MB/sec程度の速度でオブジェクトを書き込む必要があると見積もられている．この性能を実現するには，複数のI/Oサーバに負荷を分散させる必要がある．1つのI/Oサーバからは，すべてのデータベースファイルにアクセス可能なことが要求される．これを実現する手段として米国DOE傘下の研究所とIBM社の共同開発によるHPSS[2]との連携やSANの応用が考えられる．我々は，すでにHPSSを導入し，SANと両方の適用性の研究を行っている．

製品	排他処理の方法	コメント
A	NFSのサーバとロックの機構を借用	共有ライブラリの入れ替えを行う
B	SCSIのロックコマンドによりデバイスをロック	ロックの粒度が粗い
C	トークン管理方式	排他処理とI/Oは別のネットワークインタフェースを使用する

表-3　SANファイル共有製品

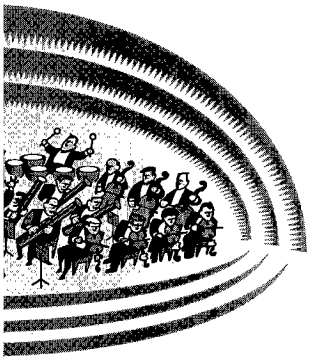
SAN環境下での分散ファイルアクセス

Objectivity/DBをSAN環境下で利用するためには，物理的にディスクが接続されているだけではなく，ファイルシステムとしてすべての計算機からファイル共有が実現されている必要がある．SAN対応と称したファイバーチャネル関連製品は多いが，実際にはミドルウェアのサポートなしには複数の計算機間でファイルの共有を行えない．SANによるファイル共有の利点の1つは，可用性の高いシステムを組みやすいことである．たとえ1つのホストがダウンしても，別のホストからまったくシームレスにすべてのファイルへのアクセスが可能である．したがって，安価な計算機システムを複数導入することでフェイルセーフなシステムの構築が可能となる．

我々は，6台の計算機でファイバーチャネルのディスクを共有するSANのテストシステムを構築した．このシステムを用いODBMSの適用性と性能の評価を行っている[3]．ミドルウェアとしてはUNIXで利用可能な次の3製品の検証を行っている（表-3）．

3つの製品とも単一のアクセスの場合には，スループットとしてほぼディスクの性能を得ることができた．しかし，多重アクセスの場合，多重度を増したときにトータルスループットが急速に下がる製品もあった．どの方式でも排他処理のオーバーヘッドは避けられない．性能の劣化は方式によるものか実装によるものかを判断するのは難しい．また，アクセスパターンにもよるので一概にいうことはできない．1TBのRAIDシステムを導入し，ホスト数をさらに増やし，実用性の検証実験を続ける予定である．他の利用可能な製品のサーベイも続ける．

現在，利用可能な製品はどれもディスクのみをサポートしており，ストレージの階層化はサポートしていない．現時点では1PBに及ぶ容量をディスクだけで構築しようとするとコストがかかりすぎてしまう．容量を低いコストで実現しようとすると階層化ストレージは有効である．HPSS



は，そもそも階層化ストレージシステムとして開発されたのでこの要件は満たしている．しかし，長いレコード長によるシーケンシャルアクセスの性能に最適化されて設計されたシステムなので，データベースが要求するような短いレコード長でのランダムアクセスには必ずしも向いていない．SANによるファイル共有と共存可能な階層化ストレージシステムが実現されるか，ディスクの価格がさらに下がることを期待したい．

ATLAS実験の場合，図-3に示されているESDとシミュレーションの出力およびAODがTier 1地域解析センターの最低限必要とするストレージの容量に相当する．ESDとシミュレーションの処理は，平行して処理を行うことにはなるが，同じデータへの多重アクセスはない．したがって，HPSSのような既存の階層化ストレージシステムとSANによるファイル共有方式の併用の検討も行っている．AODには，多重アクセスが行われるので，SANが有効であると考えている．

広域ネットワーク上でのファイル共有

CERNと日本国内の研究機関との通信を考えるときに，遅延の問題を避けることはできない．日本からCERNまではアメリカ周りでネットワーク接続されており，光ファイバーあるいは電線の長さは最短でも2万キロメートル以上となる．光の速度を超えることはできないので，往復で最小でも120ミリ秒の遅延を生じる．TCP/IPを用いた通信では，この遅延のために効率よく通信することはできない．Windowサイズを大きくするなど，通信のパラメータの工夫である程度は改善も可能だが，実際の運用条件のもとでの根本的解決は難しい．KEKでは，この遅延を人工的に作り出すシミュレータの開発も行い，通信の最適化の研究も行っている[4]．

ATLAS実験の場合，CERNから国内に転送しなければならないデータ量は毎年2PBに達するので，CERNからTier 1センターへの効率的なデータ通信の方法の開発が必要となる．現在，NCITS T11委員会ではST (scheduled transfer) protocolの制定が進められている[5]．もともと，高速なネットワーク上で，転送効率を向上させるための技術である．このプロトコルを利用すると，高遅延のネットワークに対する通信の効率も改善できると期待されている．SCSI on STというプロトコルも含まれており，WANを超えたSAN構築を可能にすると期待される．STプロトコルによりスループットは確保されるであろうが，高遅延の元でのアクセスの排他処理をどうするか，またセキュリティの問題はないかなど，今後解決しければならない課題は多いが，SANのWANへの拡大を期待して注目していきたい．

なお，国内でのTier 1地域センターとTier 2および3地域センターでのファイル共有の枠組みとして2001年度中に東京大学素粒子国際センターとKEKの間で，国立情報学研究所が提供予定のSuper Sinetを利用しテスト環境を構築する予定である．両方の機関にWDM装置が設置され，ギガビットイーサネットによる接続が可能となる．プロトコルは特に制限されないので，STプロトコルを用いWANを介したファイル共有を試みる．

ギガビットイーサネットのスイッチには，ファイバーチャネル・ギガビットイーサネットルータを繋ぎ，ファイバーチャネルのRAIDを直接ネットワークに接続可能にする．計算機はギガビットイーサネットで接続し，そのうえでSTプロトコルを利用する．SCSI over STプロトコルを利用し，OSからはそれらのRAIDがSCSIのRAIDであるかのように認識される．したがって，前に述べたミドルウェアがそのまま利用可能である．3製品のうち，1つに関してはすでに稼働が確認されており，テストを行うための環境は揃っていると考えている．WDMによる国際間のネットワーク接続を実現できる可能性は未知である．今後の技術動向を注意深く見守っていく必要がある．

おわりに

KEKにおけるSANへの取り組みについて紹介した．まだ技術的課題は多く，システムの実用化までには時間を必要とする問題も多い．今後，標準化が進み，ミドルウェアに関してもマルチプラットフォームで相互運用可能な製品が出てくることが期待される．また，ストレージの階層化もサポートされることを望みたい．多くの研究者や技術者の方々に，このような需要もあることを知ってもらい，解決へのヒントをいただけることを期待している．我々が直面している課題は，工学的にもチャレンジングであり，HEP以外の分野への応用が可能な技術であると考えている．

参考文献

1) Models of Networked Analysis at Regional Centres for LHC Experiments (MONARC), CERN/LCB2000-001.
2) http://www.sdsc.edu/hpss/
3) Sasaki, T. et al.: Prototyping an Object Warehouse, CHEP2000, KEK-preprint-99-169.
4) 佐藤博之: 高遅延ネットワークシミュレータ, Linux Conference (2000 Fall).
5) Information Technology — Scheduled Transfer Protocol (ST), NCITS T11.1/Project 1245-D/Rev 3.6.

(平成13年3月5日受付)