スーパー神岡実験解析用電子計算機システム

東京大学宇宙線研究所神岡宇宙素粒子研究施設

■ 執筆者Ｐｒｏｆｉｌｅ ■

小汐　由介

1992 年　京都大学工学部電子工学科卒業
1994 年　東京大学大学院理学系研究科修士過程終了
1997 年　東京大学大学院理学系研究科博士過程、
　　単位取得の上退学
1997 年　東京大学宇宙線研究所・助手採用
1998 年　東京大学・博士号取得(理学)
2007 年　現在 東京大学宇宙線研究所･助教(職名変更)

松崎　義昭

1988 年　弘前大学理学部物理学科卒業
1988 年　富士通株式会社　入社
　　筑波研究学園都市研究機関のコンピュータ
　　システムの構築・サポートに従事
2008 年　現在 科学分野におけるコンピュータシステ
　　ムの企画・ビジネス推進に従事
　　計算科学ソリューション統括部所属

■ 論文要旨 ■

東京大学宇宙線研究所は，神岡宇宙素粒子研究施設を 1996 年に新設し，地下 1,000 m に建設された 5 万トンの超純水を蓄えた水タンク（直径 39.3 m，高さ 41.4 m）とその壁に設置された 11,129 本の光電子増倍管（直径 50 cm）から成るスーパーカミオカンデを利用し，ニュートリノと呼ばれる素粒子の観測を続けている。超新星の爆発など数十年に 1 度 10 数秒ほどしか観測できないケースも確実にとらえるため，観測は 24 時間 365 日続けられ，蓄積されるデータは膨大であり 350 T バイトにも及ぶ。従来のシステムでは，このデータをテープに蓄積し，解析を行っていたため，数年分のデータの読み出しに，半年かかったこともあった。そこで，観測データを大容量ディスク装置に格納し，データアクセスの高速化を行い、データ転送速度 960M バイト/秒のスループット性能を有するスーパー神岡実験解析用電子計算機システムを 2007 年 2 月に導入した。

■ 論文目次 ■

■ 図表一覧 ■

１．はじめに

１．１　神岡宇宙素粒子研究施設の概要

東京大学宇宙線研究所神岡宇宙素粒子研究施設[1]では，ニュートリノの観測，陽子崩壊の探索を通じて，素粒子物理学の研究を行っている。スーパーカミオカンデ[2]（図１）は，ニュートリノの観測装置として，1996 年に建設され，以来 1998 年にはニュートリノ振動という現象を発見し，ニュートリノに質量があることを証明するなど数々の発見を生み出し，ブラックホールや星の誕生の謎の解明に挑戦している。超新星の爆発など数十年に 1 度 10 数秒ほどしか観測できないケースも確実にとらえるため，観測は 24 時間 365 日続けられ，1 日の観測で保存される生データは約 50 G バイトであり，これまでに蓄積されたデータは加工データと合わせて 350 T バイトに及んでいる。この膨大な観測データを大容量ディスク装置に格納し，高速にアクセスして解析するためのスーパー神岡実験解析用電子計算機システム（図２）を 2007 年 2 月に導入した。

１．２　スーパーカミオカンデ

スーパーカミオカンデは，5 万トンの超純水を蓄えた直径 39.3 m，高さ 41.4 m の円柱形水タンクと，その壁に設置された光電子増倍管と呼ばれる 11，129 本の光センサから成り，観測の邪魔になる宇宙線を避けるため，岐阜県・神岡鉱山の地下 1,000 m に設置されている。

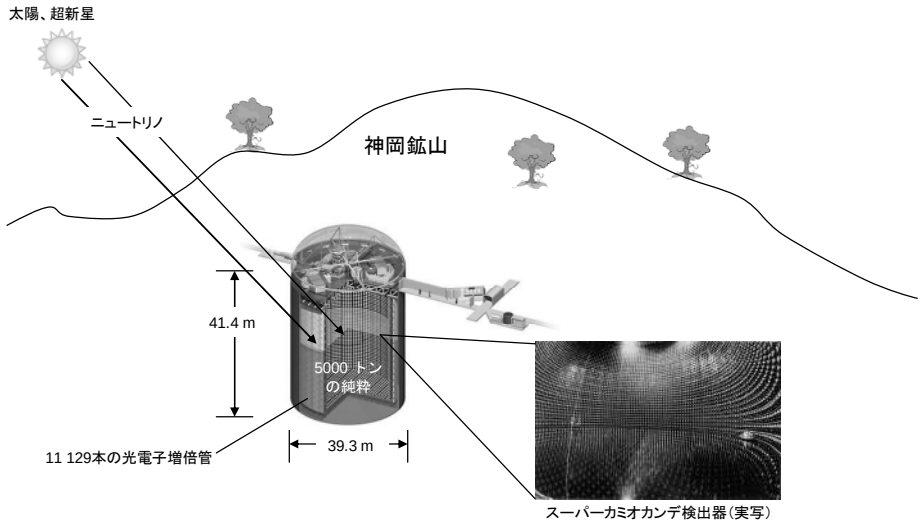


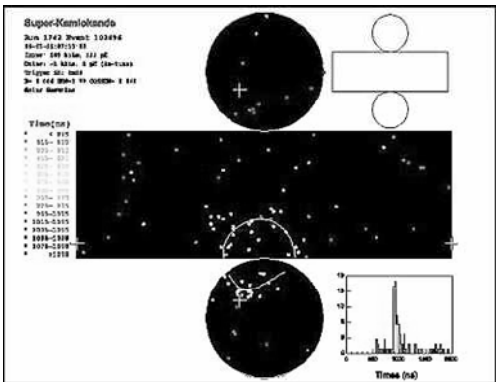
図1　スーパーカミオカンデ

１．３　データ観測と計算機システムへの要求

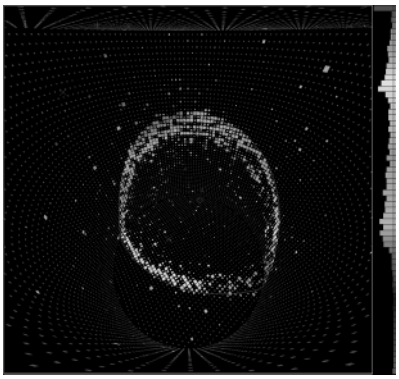
宇宙から飛来するニュートリノには，太陽から来るもの（太陽ニュートリノ（図２）），宇宙線が地球の大気と反応して発生するもの（大気ニュートリノ（図２）），また星の一生の最後に起こす超新星爆発のときに発生するもの（超新星ニュートリノ）などがある。ニュートリノがスーパーカミオカンデに飛び込んでくると，タンク内の水と反応して微弱な青白い光（チェレンコフ光）を発生することがある。この光を光電子増倍管で検出することにより飛び込んできたニュートリノのエネルギー，反応位置，進行方向を計算する。これを事象再構成と呼ぶ。

ニュートリノ観測において特に重要となるのは，バックグラウンド事象の除去である。例えば，太陽から来るニュートリノのバックグラウンドとしては，タンク外から入ってくる環境ガンマ線やタンク内の水中にわずかに残存するラドンなどの放射性物質がある。これらはニュートリノ反応と同様に水中でチェレンコフ光を発生し，非常に紛らわしい事象となる。スーパーカミオカンデでは，観測された粒子の発生位置や進行方向を用いてバックグラウンド事象との区別を行うことができる。観測データのうち，反応位置などから明らかにバックグラウンド事象と判断されたものは，データ取得直後に破棄される。

残ったデータは欧州合同素粒子原子核研究機構（ CERN ： European Organization for NuclearResearch）[3]の世界標準フォーマットである ZEBRA[4]フォーマットへ変換され，実験解析用電子計算機システムへ送られる。これをリフォーマット処理と呼ぶ。1 レコードは約 5 K バイトの長さで，1 日に保存される事象は約 1,100 万事象あるため，保存されるデータ量は 1 日に約 50 G バイトである。スーパー神岡実験解析用電子計算機システムでは，上で残ったすべての事象に対し，まず光電子増倍管の個々の特性に関するパラメタや水質に関するパラメタ（例えば水の透明度など）を用いた補正を行い，さらにリアルタイムに事象再構成を行う。しかし，これらのパラメタは時期的な変動もあり，さらには，事象再構成のアプリケーションも日進月歩で進化しているので，これまで蓄積した生データを基に再解析を行うことが多い。ただ，その再解析を行う生データ量は 110 T バイトもあるため，非常に高速なデータアクセスが要求される。



スーパーカミオカンデでとらえた太陽ニュートリノ事象（展開図）



スーパーカミオカンデでとらえた大気ニュートリノ事象（三次元図）

図2　スーパー神岡実験解析用電子計算機システムによって可視化した事象

１．４　スーパー神岡実験解析用電子計算機システム

スーパー神岡実験解析用電子計算機システムは，スーパーカミオカンデで観測されたデータを収集しフォーマット変換を行うための坑内システムとフォーマット変換されたデータを蓄積し，解析業務を実施するための坑外システム，さらに，日常的に使用される端末やバックアップシステム，監視システム，それらを接続するギガビットイーサネットなどから構成される。

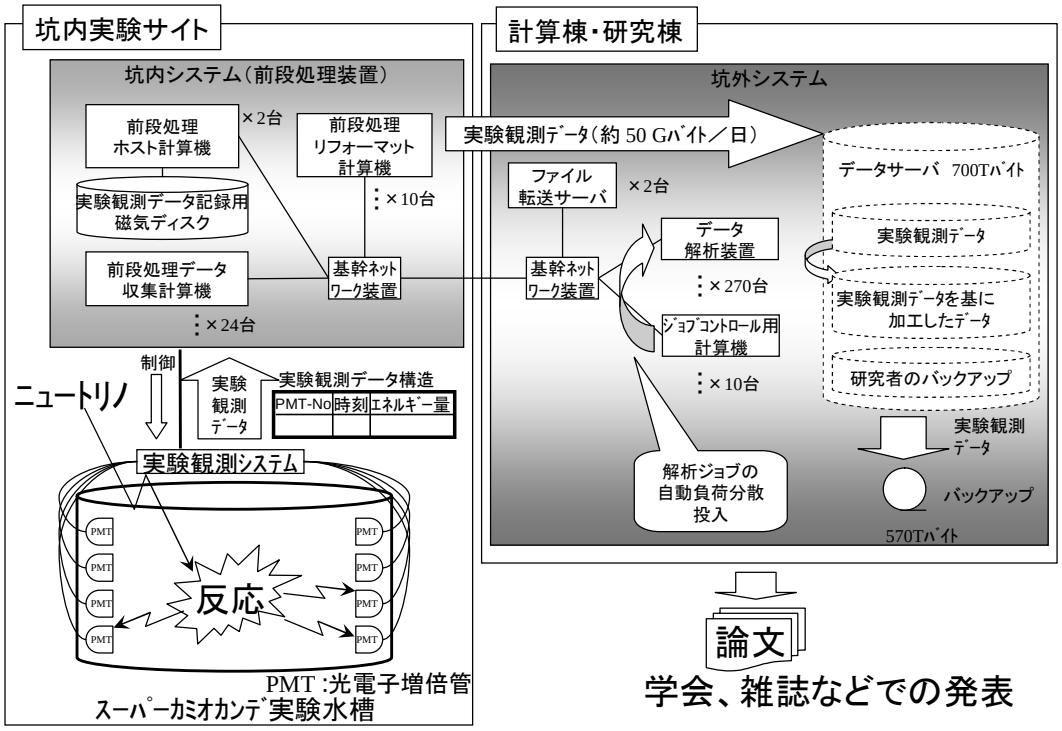


図3　スーパー神岡実験解析用電子計算機システム

２．従来システムの問題点

従来のシステムでは，350 T バイトの膨大なデータをテープに蓄積し，解析を行っていたため，数年分のデータの読み出しに，半年かかったこともあった。そこで，新システム構築において，次節に述べるデータアクセスの高速化を行い、データ転送速度 960M バイト/秒のスループット性能を達成した。

３．データアクセスの高速化とループット性能

３．１　データアクセスの高速化

観測データの蓄積だけであれば 24 時間で約 50 G バイトを書き込めればよい。しかし，蓄積された過去のデータをデータ解析装置で処理する場合，解析処理が滞りなく行えるよう，解析プログラムのデータアクセスに対して，できるだけ高速な入出力性能を提供する必要がある。

これを実現するためにとった対策を，以下に述べる。

３．１．１　ファイルシステム

データ解析装置とファイル転送サーバは，ネットワーク型のファイルシステムを構築することで，どのデータ解析装置からも同じようにファイルを利用可能としている。ネットワーク型のファイルシステムで一般的なものは NFS であるが，経験上 NFS では全データ解析装置から同時に発行された入出力要求を処理することは難しく，かつ，高速性能は望めないと判断し，NFS に代わるファイルシステムとして SRFS を採用した。

また，解析プログラムの入出力モデルジョブを作成し，SRFS を使用して性能測定作業を実施した。この作業の結果，入出力データ長を 8 M バイトと決定し，設計作業では入出力データ長が 8 M バイトの場合に最大性能が発揮できるよう考慮した。

３．１．２　ストレージ

ファイル転送サーバ装置に接続される磁気ディスク装置は，磁気ディスクドライブ 7 個を同時に並列利用するよう構成される。この 7 個の磁気ディスクドライブのセットを物理ボリュームと呼ぶ。この構成をそのまま利用すると，物理ボリュームの最大性能が入出力性能の限界値となる。この限界を超えるために，物理ボリュームを束ねて論理ボリュームを構成し，論理ボリューム利用時に，同時に複数の物理ボリュームが使用される方式を採用した。

論理ボリューム内をいくつの物理ボリュームで構成するか（ストライプ列数），論理ボリュームへの一度の入出力に対して，一つの物理ボリュームへの一度の入出力量をいくつにするか（ストライプ幅）は，論理ボリュームの最大性能を引き出すための重要な設計ポイントである。

具体的には，性能測定の結果から，ストライプ幅は 128 K バイトまたは 256 K バイト，ストライプ列数は 16 または 8 が良い性能を発揮できると判断した。しかし，ストライプ列数を 16 にすると物理ボリューム数が多くなり過ぎ，磁気ディスク装置のハードウェア制限により構成することができないことが判明し，ストライプ列数 8，ストライプ幅 256 K バイトで構成することとした。

３．１．４　ネットワーク

ファイルシステムとして採用した SRFS はネットワークファイルシステムであり，ギガビットイーサネットを媒介して実データとのアクセスを行うが，入出力以外のトラフィックにより入出力性能が低下すること，SRFS 自身が発信するブロードキャストパケットが，ほかの通信を妨害することが予想されたため，入出力専用のネットワークを

構成した。

３．１．５　入出力ライブラリ

入出力長を 8 M バイトに設定した専用の入出力ライブラリを提供することで，利用者が開発した解析プログラムからの高速なデータ入出力を実現した。

３．２　スループット性能

本システムのデータ解析装置では，1 台あたり最大四つの解析プログラムを動作させるため，全データ解析装置から同時に 1,080 個の入出力要求が発行される可能性がある。解析処理を円滑に行うためには，これら多数の入出力要求を滞りなく処理できる高速なスループット性能が要求される。

高スループットを実現するために行った内容と，スループット性能測定結果を以下に述べる。

３．２．１　ネットワークスローダウンの抑止

ネットワークを効率的に利用するには，帯域を使い切るように使用するのが理想的であり，通常ネットワーク帯域を使い切るためには，一つの要求をいくつかに分割して一つのネットワーク上で並行動作させるか，あるいは，複数の要求を一つのネットワーク上で同時実行させる。しかし，本システムのようにデータサーバ側の物理インタフェース数に対して，データ解析装置側の物理インタフェース数が多い場合は，データサーバ側の帯域が足りず，そのままではネットワークスローダウンを招く。このため，1 台のデータサーバの物理インタフェースが受け持つデータ解析装置数を制限することで，ネットワークスローダウンを抑止した。

具体的には，データサーバは 1 台あたり入出力用ネットワークへの物理インタフェースを七つ持っているが，データ解析装置は 270 台のため，その一つの物理インタフェースに対してデータ解析装置を 38 台あるいは 39 台を受け持つよう設定した。この設定によりデータサーバの一つの物理インタフェースの故障で，38 台あるいは 39 台のデータ解析装置が利用不可となるが，スループット性能を重視した設計とした。

３．２．２　通信タイムアウトの抑止

ネットワーク通信においてタイムアウトとリトライ回数は重要な設計ポイントである。タイムアウト値が長すぎると異常の検知と復旧が遅れ，短過ぎるとリトライによる通信が増加し通信性能が低下するからである。

本システムでは異常検知よりスループット性能を重要視し，高負荷時でもタイムアウトせず動作する値を設定したが，この値を計算によって求めることは難しく，最終的には実測によるチューニング作業を実施した。

実測では前章で述べたモデルジョブを 1,080 個同時に実行し，一つのファイルシステムに対して入出力要求を同時に発行させた結果から，タイムアウト値を増減し，これを繰り返し実施した。タイムアウトが発生していると，リトライを行うことからジョブの実行時間にばらつきが生じるが，タイムアウトが発生していない状態では，ジョブの実行時間はほぼ一致すると判断し，最適値を決定した。

３．２．３　スループット性能測定結果

スループット性能測定では，タイムアウト値チューニングに使用したジョブを，270 台のデータ解析装置上で 1,080 ジョブ同時実行し，これらが 3 台のデータサーバに対して入出力を行う際の入出力速度を計測した。この際，最初のジョブの実行開始から 1,080 ジョブが同時に並行動作するまで，ある程度時間がかかることと，ジョブ終了による並列度数の減少を考慮し，一つのジョブを数回連続実行させ，最初と最後の結果を切り捨てることで，多重度が 1,080 に満たない場合の測定値を排除した。

このような条件でデータサーバに配置したデータの Read/Write を行った結果，960 M バイト/ 秒（Read/Write 平均値）のスループット性能を達成した。

４．むすび

本稿では，東京大学宇宙線研究所神岡宇宙素粒子研究施設におけるカミオカンデおよびデータ解析用電子計算機システムを紹介し，さらに，いかにして高速なデータアクセスを実現したかについて背景を交えて説明した。サイトごとのデータ特性やシステム構成は違うので，一概に同じ方策が最適とは言えないが，計算機システム設計の考え方や課題解決のアプローチについて，今後の参考にしていただければ幸甚である。

参考文献

［１］東京大学宇宙線研究所神岡宇宙素粒子研究施設.
http://www-sk.icrr.u-tokyo.ac.jp/index.html

［２］スーパーカミオカンデ.
http://www-sk.icrr.u-tokyo.ac.jp/sk/

［３］CERN.
http://public.web.cern.ch/Public/Welcome.html

［４］CERN：The ZEBRA System.
http://wwwasdoc.web.cern.ch/wwwasdoc/zebra_html3/zebramain.html