

ニュートリノ研究を支えるスーパー神岡 実験解析用電子計算機システム

Computing System for Super-Kamioka Neutrino Detection Experiment

● 岡田英之 ● 大山幸信 ● 塚原知宏

あらまし

東京大学宇宙線研究所様は、スーパーカミオカンデを利用して超新星ニュートリノ、太陽ニュートリノ、大気ニュートリノの観測、および素粒子の大統一理論の検証研究を行っている。スーパーカミオカンデは、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)から成る巨大な観測装置である。スーパーカミオカンデで観測されるデータは1日あたり約45 Tバイトであり、観測データからノイズを除去した約500 Gバイトのデータが保存され、ディスク装置に蓄積されているデータはユーザーデータを含め3 P(ペタ)バイトにも及ぶ。自然現象を対象に観測を行っており再現不可能な事象であるため、24時間365日の確実なデータ取得と長期にわたる観測データの保存が富士通の課題である。富士通ではこの膨大なデータを大容量ディスクに確実に格納、高速にアクセスして解析し、データを安全に長期保存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を2012年2月に導入した。

本稿では、本計算機システムを活用した確実な観測データ取得、高速なデータ処理、長期間のデータ保存のための工夫について紹介する。

Abstract

The Institute for Cosmic Ray Research of the University of Tokyo is observing supernova, solar, and atmospheric neutrinos, while researching the Grand Unified Theory, using the Super-Kamiokande (SK). This is a large-scale observation device, constructed 1,000 m below ground level, with a water tank (41.4 m in height, 39.3 m in diameter) that is capable of storing 50,000 tons of pure water. There are 11,129 photomultiplier tubes (each 50 cm in diameter) mounted on the tank wall. The SK can generate approximately 45 TB of observational data per day, of which approximately 500 GB of data is saved after noise is reduced. The total amount of data stored on the disks is up to 3 PB, including user information. The SK observes natural phenomena, and the challenge for Fujitsu here is to ensure that accurate observational data is captured 24/7, and to store them over a long period of time. In February 2012, Fujitsu installed a computing system in the SK, designed to facilitate the storage of an astronomical volume of data on a large-capacity disk, so that the data can be quickly accessed, and to provide safe, long-term data storage. This paper explains the system's various innovations that realize accurate observational data capturing, high-speed processing, and long-term archiving.

ま え が き

東京大学宇宙線研究所神岡宇宙素粒子研究施設様（以下、東大宇宙線研）⁽¹⁾では、ニュートリノの観測や陽子崩壊の探索を通じて、宇宙や素粒子物理学の研究を行っている。小柴昌俊教授が1987年2月の超新星爆発によるニュートリノ観測により2002年ノーベル物理学賞を受賞されたカミオカンデの後継ニュートリノ観測装置として、スーパーカミオカンデ⁽²⁾は1996年に観測が開始された。1998年には、梶田隆章教授らがニュートリノ振動を確認したことを発表し、2015年度のノーベル物理学賞受賞に結びついており、世界的に注目されている実験の一つである。

富士通は、1993年に本実験の観測データを処理するシステムを導入した。その後、現在に至るまでの間、計算機技術の進歩に合わせてシステムのリプレースを実施し、本実験のデータ解析業務を支えてきた。そして、2012年2月に「スーパー神岡実験解析用電子計算機システム」を導入した。ニュートリノ観測においては、1日に取得されるデータは約45 Tバイトにも及び、このデータからノイズ除去や放射性物質の反応を除いた約500 Gバイトのデータが毎日保存される。これらのデータを保存し、必要なデータをできるだけ速く取り出すために、よりアクセス性能の高い富士通の磁気ディスク装置（FUJITSU Storage ETERNUS DX80 S2）を利用したストレージシステムにリプレースした。また、新たに観測したデータ解析のほか、新しいアプリケーションやパラメーターを変更したアプリケーションにより、過去のデータの再解析を行っている。短期間にこれらの解析を行うために、より高速なデータアクセス性能が必要となる。このため、高速なファイル共有を実現する富士通のスケラブルファイルシステムソフトウェアFEFS（Fujitsu Exabyte File System）を利用し、データアクセスの高速化を図った。

本稿では、スーパー神岡実験解析用電子計算機システムの概要を説明する。

データ観測概要

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本の光センサーから成る。なお、本装置は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。研究の目的は、宇宙から飛来するニュートリノの観測や、陽子崩壊と呼ばれる事象の観測などである。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や、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。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により、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っている。観測されたデータは、欧州合同素粒子原子核研究機構（CERN：the European Organization for Nuclear Research）⁽³⁾が開発しているROOTフォーマット⁽⁴⁾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日に観測される45 Tバイトのデータよりノイズを除去し、チェレンコフ光によるデータを切り出し、バックグラウンド事象を除去したデータ（1日あたり約500 Gバイト）を保存している。スーパー神岡電子計算機システムでは、光電子増倍管の個々の特性に関するパラメーターや水質に関するパラメーター（例えば、水の透明度など）で補正し、リアルタイムに事象再構成を行う。しかし、これらのパラメーターは時期的な変動もあり、更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ため、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。再解析するデータ量は約500 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。

次章からは、坑内実験サイトにある坑内システムと計算機棟・研究棟にある坑外システムの構成を説明し、確実なデータ取得、高速なデータ処理、長

期間のデータ保存を実現した技術について述べる。

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する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、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、それらを接続するネットワークシステムなどから構成される。

ニュートリノの観測は24時間365日行われ、年々データは蓄積されるため、観測データの格納先として膨大な容量が必要となる。スーパー神岡実験解析用電子計算機システムは1993年から稼働しており、現在は5世代目となる。過去のシステムの稼働期間、およびストレージの構成と容量を表-1に示す。

第1～第3世代は、バックアップを含めて磁気テープを採用していた。第4世代以降は、目的に応じて異なるファイルシステムを利用し、高速性を要求

する大容量磁気ディスク領域と、長期保存を目的としたバックアップ領域を分割した。

本実験を支える、第5世代システムは、神岡鉱山の内部（坑内システム）と外部（坑外システム）があり、それぞれ、以下のような構成と特長を有している（図-1）。

● 坑内システム

スーパーカミオカンデでは、宇宙線に関する重

表-1 システムの稼働期間とストレージ

世代	稼働期間 (年)	データ記録 媒体	バックアップ 媒体
第1世代	1993～1997	磁気テープ (CMT) 12.8 Tバイト	
第2世代	1997～2002	磁気テープ (DTF) 200 Tバイト	
第3世代	2002～2007	磁気テープ (LTO1) 440 Tバイト	
第4世代	2007～2012	磁気ディスク 700 Tバイト	磁気テープ (LTO3) 547 Tバイト
第5世代	2012～	磁気ディスク 3 Pバイト	磁気テープ (LTO5) 1.5 Pバイト

CMT : Cartridge Magnetic Tape
DTF : Digital Tape Format
LTO : Linear Tape-Open

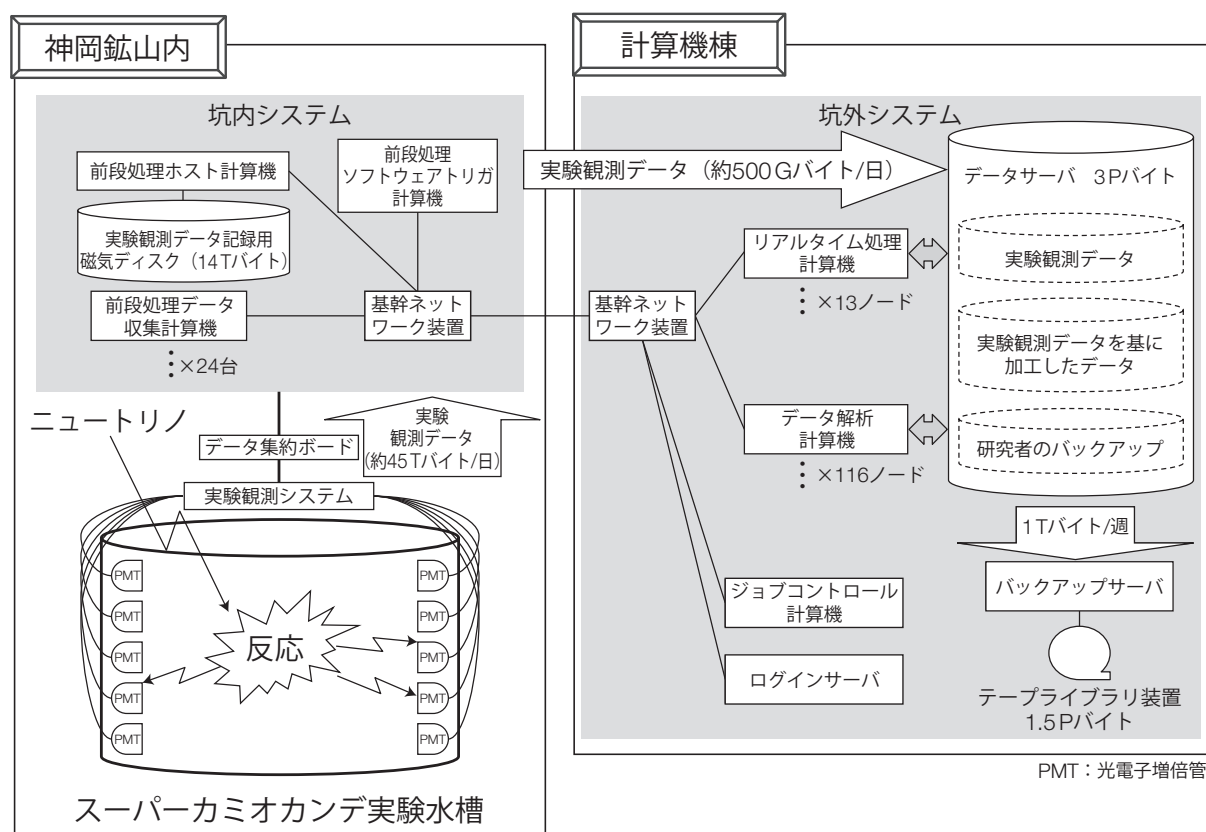


図-1 スーパー神岡電子計算機システム

要なイベントの発生に備え、24時間体制で観測を実施している。そこで使用される坑内システムは、高い稼働性が求められるリアルタイムシステムである。坑内システムは、前段処理データ収集計算機（FUJITSU PRIMERGY RX200 S6：24台）、前段処理ホスト計算機（FUJITSU PRIMERGY RX300 S6：2台、ETERNUS DX80 S2：1台）、前段処理ソフトウェアトリガ計算機（PRIMERGY RX200 S6：10台）、およびそれらを接続するためのオンラインネットワーク（Cisco Systems社製スイッチCatalyst 4506-E：1台）から成る。

実験観測システムと前段処理データ収集計算機は、東大宇宙線研により開発された専用のデータ集約ボード約520枚を用いて接続され、専用アプリケーションでデータ収集を実施している。前段処理データ収集計算機は収集したデータを前段処理ソフトウェアトリガ計算機へ送信し、前段処理ソフトウェアトリガ計算機でデータを時間順に並べ替え、ノイズを除去、事象選別を行う。更に前段処理ホスト計算機へ送信し、実験観測データ記録用磁気ディスクに蓄積する。

● 坑外システム

坑外システムは、前段処理装置から送られてきた観測データの蓄積・解析を行う。本システムは、観測データの解析処理の高速化対応として、最大1,392本（12CPU/ノード×116ノード）のジョブ同時実行が可能となっている。新たに蓄積された観測データの校正や再解析などが常時実行されており、通常500本前後、多いときには実行待ちも含めて3,000本を超えるジョブが投入されている。これら多量のジョブからのデータアクセスを効率良く行うために、ファイルシステムへの高速アクセスが極めて重要である。

（1）坑外システムの構成と特長

坑外システムは、データサーバ（PRIMERGY RX300 S6：8台、ETERNUS DX80 S2：16台）、リアルタイム処理計算機（PRIMERGY BX922 S2：13台）、データ解析計算機（PRIMERGY BX922 S2：116台）、およびこれらを接続するための基幹ネットワーク装置（Cisco Systems社製のCatalyst 4507R+E スイッチ）などから構成される。

データサーバは総容量3 Pバイトのデータ蓄積領域を有し、FEFSでジョブコントロール用計算機お

よびデータ解析計算機に対してファイル共有環境を提供している。リアルタイム処理計算機、データ解析計算機とデータサーバ間はInfiniBand（メラノックス社製のGrid Director 4036）により接続され、ファイルアクセスの高速化を図っている。その結果、多数のジョブからの安定的なファイルアクセスを実現し、快適なデータアクセス環境を提供している。

（2）プログラム開発環境とジョブ制御

利用者によるプログラム開発や解析装置へのジョブ投入は、ログインサーバを利用する。ログインサーバには、複数バージョンのコンパイラが用意されている。また、バッチジョブ運用支援ソフトウェアFUJITSU Software Technical Computing Suite（以下、TCS）環境を用意し、多数のジョブを高速にスケジューリングして実行することを可能としている。これにより、開発～実行～評価など解析業務に必要な全ての作業を一つの端末で行うことができる。

また、本システムのバッチジョブシステムは、リアルタイム系とユーザージョブ系の2系統に分割されている。リアルタイム系は、坑内システムから送られてくるデータのリフォーマット処理、ノイズ除去、事象選別処理を自動的にリアルタイム処理計算機上で実行し、処理結果をデータサーバに保存する。ユーザージョブ系は、データ解析計算機上で実行され、リアルタイム系で処理されたデータを基に、研究者が様々な解析処理を行うために利用されている。このように、バッチジョブシステムの分割には、障害の影響範囲の限定や、系ごとに保守を行うなど、運用へ影響を最小限に抑える工夫がなされている。

TCSでは計算機資源の管理も行っている。ジョブコントロール計算機から投入されたジョブは、TCSによりデータ解析計算機の中から空いているCPUで実行されるため、利用者が空いているリソースを探す必要はない。更に、ジョブで使用していないメモリ領域は、ファイルキャッシュで利用することによりファイルアクセスの高速化が可能である。そのため、できるだけノードを分散し、各ノードのメモリ使用量が少なくなるようにジョブの配置を最適にする設定を行っている。このような工夫により、観測データ解析時間の高速化を図っている。

確実なデータ取得

観測対象が自然現象であるため、構築するシステムに対しては、24時間365日安定して観測データが取得できることが要求されている。

スーパーカミオカンデから、坑外システムのテープライブラリ装置に観測データがアーカイブされるまでの各種サーバやネットワークに関しては、冗長化構成となっており、単体故障にも耐えられるシステム構成となっている。坑外システムのメンテナンスや一時的なネットワーク切断などにより、坑外システムのデータサーバ（最終的な実験観測データの格納先）にデータ転送できない事象が発生しても、観測データの取得は継続される。そこで図-2に示すように、前段処理ホスト計算機にリアルタイムに送られてくる実験観測データを記録するための磁気ディスク装置（14 Tバイト）を備えている。これにより、最長で約30日間、実験観測データを前段処理ホスト計算機で蓄積することが可能となっている。

高速かつ安定したデータ処理

観測データの蓄積だけであれば、1日あたり500 Gバイトを書き込めればよい。しかし、解析プログラムの変更などにより、蓄積された過去のデータをデータ解析計算機で再処理する場合もある。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高速な入出力性能を提供する必要がある。更に、データ解析処理を短時間で行うためには、24時間365日安定してシステム

が稼働することも必要となる。

● ファイルシステム

データ解析計算機とデータサーバ間はInfiniBandで高速に接続され、どのデータ解析計算機からも均一に高速なファイルアクセスを可能としている。データ解析計算機とデータサーバ間は最大で2.5 Gバイト/秒の書き込み性能を有しており、多数のジョブが同時にファイルアクセスを行っても運用に影響しない設計となっている。

ファイルシステムの総容量は3 Pバイトであるが、完全に独立した1.5 Pバイトのファイルシステム2領域の構成となっている。一つは、運用におけるファイルシステムとしてユーザーに公開し利用していただくための領域である。もう一つは、試験環境として運用改善や修正適用の事前検証として利用する領域である。この構成により、安定した運用を実現している。

FEFSを構成するサーバや通信経路は全て冗長化されており、サーバ障害などによるファイルシステム閉塞を防止できる構成となっている。また、FEFSのディレクトリquota設定機能によって、各研究グループに割り当てるディレクトリのファイル使用量の制限を行っている。本機能は、同一ファイルシステム内の特定ディレクトリ配下のファイル使用量に対して制限を行うことができる。各研究グループのファイル使用量に応じて、運用を停止せずに制限値の増減を行うことにより、柔軟な対応が可能となる。

● ストレージ

データサーバの磁気ディスク装置は、磁気ディ

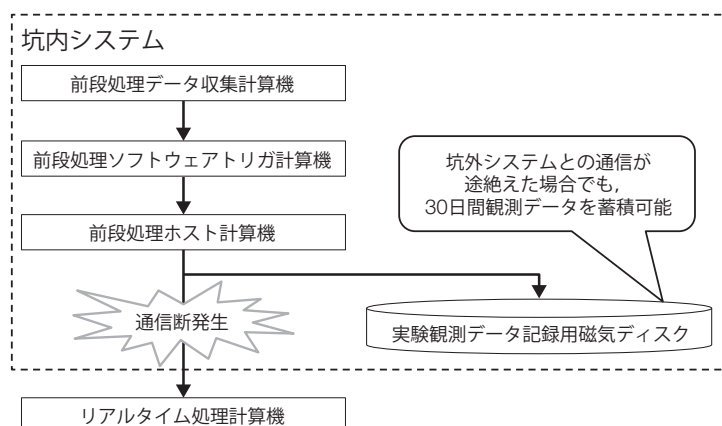


図-2 通信断発生時の観測データ蓄積

スクドライブ14台を同時に並列利用するよう構成される。この14台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。この物理ボリュームをファイルシステムあたり48台使用して負荷を分散することで、大容量で高速なファイルアクセスを実現している。

● バッチジョブ運用支援ソフトウェアTCS

前システムのNQS (Network Queuing System) とはコマンド体系が違っており、利用者が作成した既存スクリプトによる影響を考慮してTCS環境でNQSコマンドが実行できる機能を用意しスムーズな移行を実現した。

TCSでは、各グループ用のジョブキューを用意しているが、実行ジョブの偏りを考慮して、一部のデータ解析計算機は複数のジョブキューに含めている。複数のジョブキューでサーバを共有することにより、特定のジョブキューにジョブがなくなった場合においても、稼働率が低下しないようにしている。

長期間のデータ保存

観測データは取り直すことができないため、確実に保存する必要がある。そのため、坑外システムでは、高速アクセスが可能なデータサーバにファイルを保存すると同時に、自動的にテープライブラリへのバックアップを実行する。非圧縮で1.5 PバイトのLTO5テープが搭載可能なテープライブラリ装置 (ETERNUS LT270: 2台) を有しており観測データのバックアップを取得している。テープドライブの記録速度は、非圧縮で1,120 Mバイト/秒の書き込み性能を有している。

バックアップソフトウェアとしてVeritas Netbackup Standardを採用している。各リアルタイム処理計算機上にクライアントソフトをインストールしており、テープライブラリ装置へデータのバックアップを行っている。また、システム利用者作成のデータを長期保存するためにも、テープライブラリ装置とVeritas Netbackup Standardを利用している。バックアップサーバの特定ディレクトリに置かれたシステム利用者作成のデータを、定期的にテープライブラリ装置にバックアップする運用を行っている。

む す び

本稿では、東大宇宙線研におけるスーパー神岡実験解析用電子計算機システムを紹介した。東大宇宙線研では、今後、更にニュートリノの研究を進められていく。富士通も、より高機能なシステムを提案し続けることにより、本実験に貢献していきたいと考えている。

本稿の執筆に当たり、ご指導、ご協力いただきました東京大学宇宙線研究所神岡素粒子研究施設准教授 早戸 良成様に心より感謝いたします。

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://home.cern/>
- (4) ROOTフォーマット.
<https://root.cern.ch/>

著者紹介



岡田英之 (おかだ ひでゆき)

テクニカルコンピューティング・ソリューション事業本部
計算科学ソリューション統括部
スーパー神岡電子計算機システムをはじめとした、科学分野におけるコンピュータシステムの商談推進、構築、運用支援に従事。



大山幸信 (おおやま ゆきのぶ)

(株) 富士通総研
公共事業部
公共分野のコンサルティングに従事。



塚原知宏 (つかはら とむひろ)

テクニカルコンピューティング・ソリューション事業本部
計算科学ソリューション統括部
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。