特集

スーパー神岡実験解析用電子計算機システム

Super-Kamioka Computer System for Analysis

あらまし

東京大学宇宙線研究所様は，神岡宇宙素粒子研究施設を1996年に新設し，地下1000 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。富士通では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）を2007年2月に導入した。本稿ではスーパー神岡実験解析用電子計算機システム全体の構成とニュートリノのデータをどのように処理しているかを紹介するとともに，いかにしてデータアクセスの高速化とスループット性能の向上を図ったかについて紹介する。

Abstract

The Institute for Cosmic Ray Research (ICRR) of the University of Tokyo newly established the Kamioka Observatory in 1996, and has continued to observe the elementary particles known as neutrinos by using the Super-Kamiokande Neutrino Detection Equipment. This equipment contains a 50 000-ton ultrapure water tank measuring 39.3 meters in diameter, 41.4 meters in height, and located 1000 meters underground. A total of 11 129 photomultiplier tubes (PMTs, 50 cm in diameter) are mounted on the inner wall of the tank. The Kamioka Observatory continues observation 24 hours a day, 365 days a year in order to detect neutrinos observable only for 10 seconds from a supernova explosion which may occur once every dozens of years. The current size of total accumulated data is nearly 350 TB. In February 2007, Fujitsu installed a computer system (known as the “Super-Kamioka Computer System for Analysis”) using mass storage disk drives for saving and rapidly accessing the observed data.

This paper describes the configuration of the Super Kamioka Computer System for Analysis, explains how data is managed and rapidly accessed, and how throughput performance is improved.

万谷　哲
（まんたに　あきら）

計算科学ソリューション統括部 所属
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。

松崎義昭
（まつざき　よしあき）

計算科学ソリューション統括部 所属
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。

山口　靖
（やまぐち　やすし）

計算科学ソリューション統括部 所属
現在，科学分野におけるコンピュータシステムの構築・サポートに従事。

神林康喜
（かんばやし　こうき）

計算科学ソリューション統括部 所属
現在，科学分野におけるコンピュータシステムの構築・サポートに従事。

スーパー神岡実験解析用電子計算機システム

ま え が き

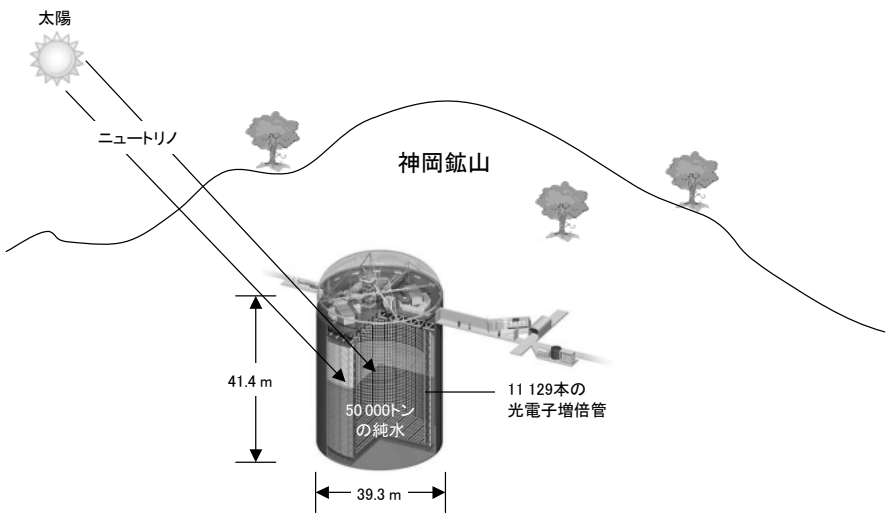
東京大学宇宙線研究所神岡宇宙素粒子研究施設[1]様では，ニュートリノの観測，陽子崩壊の探索を通じて，素粒子物理学の研究を行っている。スーパーカミオカンデ（**図-1**）[2]は，ニュートリノの観測装置として，1996年に建設され，以来1998年にはニュートリノ振動という現象を発見し，ニュートリノに質量があることを証明するなど数々の発見を生み出し，ブラックホールや星の誕生の謎の解明に挑戦している。ニュートリノの観測においては，1日の観測で保存される生データは約50 Gバイトであり，これまでに蓄積されたデータは加工データと合わせて350 Tバイトに及んでいる。これらのデータを保存し，必要なデータをできるだけ速く取り出すために，2007年2月に従来のテープ装置を利用した階層型ストレージ管理システムから，よりアクセス性能の速い富士通の磁気ディスク装置（ETERNUS4000 model500）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の解析のほか，過去のデータを利用して新しいアプリケーションやパラメタを変更した再解析を行っており，短期間にこれを行うために，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Parallelnavi SRFS（Shared Rapid File System，以下SRFS）を利用し，データアクセスの高速化を図った。

本稿では，観測データの概要，解析用システムの説明および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1000 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，陽子崩壊と呼ばれる事象の観測などである。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），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），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，



図版提供：東京大学宇宙線研究所神岡宇宙素粒子研究施設
© Kamioka Observatory, ICRR (Institute for Cosmic Ray Research), The University of Tokyo

図-1　スーパーカミオカンデ
Fig.1-Super-Kamiokande Neutrino Detection Equipment.

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 Nuclear Research）[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1100万事象あるため，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。

実験解析用電子計算機システムの構成を**図-2**に示す。次章からは坑内実験サイトにある坑内システムと計算棟・研究棟にある坑外システムの構成，高速データアクセス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構成

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

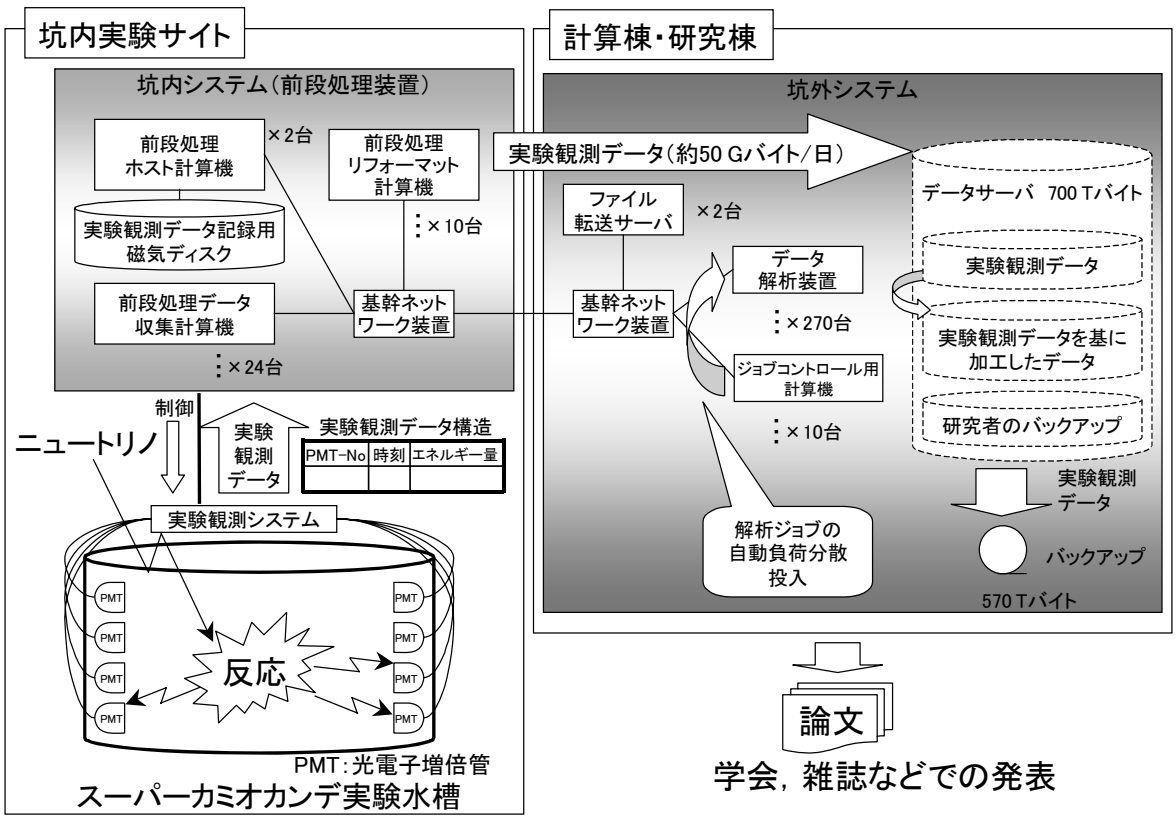


図-2　スーパー神岡実験解析用電子計算機システム
Fig.2-Super-Kamioka Computer System for Analysis.

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。

以下，主な構成システムである，坑内システムと坑外システムについて説明する。

● 坑内システム

スーパーカミオカンデでは，宇宙線に関する重要なイベントが発生した際にそのデータを確実に観測するため，24時間体制で観測を実施している。そこで使用される前段処理装置は，高い稼働性が求められるリアルタイムシステムである。

前段処理装置は，前段処理データ収集計算機（PRIMERGY RX200 S3：24台），前段処理ホスト計算機（PRIMERGY RX300 S3：2台，ETERNUS4000 model100：1台），前段処理リフォーマット計算機（PRIMERGY RX200 S3：10台）の各計算機，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（Catalyst4948：4台，Catalyst2960G：2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究所様により開発された専用のインタフェースで接続され，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ホスト計算機に接続された実験観測データ記録用磁気ディスクに送信し，蓄積する。

前段処理ホスト計算機では，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）との通信が途絶えた場合でもリアルタイムに送られてくる実験観測データが失われるリスクを低減するために，できるだけ大きなファイルシステムを構築する必要がある。ETERNUS4000 model100では，1RAIDグループあたり容量2 Tバイトの制限があるが，Linuxシステムの論理ボリューム管理機能（LVM： Logical Volume Manager）により，複数のRAIDグループを一つのファイルシステムに束ねて，5 Tバイトのファイルシステムを構築した。これにより最長で約100日間，実験観測データを前段処理ホスト計算機側で蓄積することができる。

前段処理データ収集計算機では，前述した事象再構成処理を行い，前段処理リフォーマット計算機によりリフォーマット処理を実行後，必要なデータ（約50 Gバイト/日）をデータサーバに転送する。

● 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・解析を行うシステムである。パラメトリックな解析処理を高速に処理するため，最大1080本（4CPU/ノード×270ノード）のジョブを同時に実行することが可能である。新たに蓄積された観測データのキャリブレーションや，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1080本を超えるジョブが投入されている。これら多量ジョブからのデータアクセスを効率良く行うためには，ディスク装置設計およびファイルシステム設計が極めて重要である。その具体的な注意点などは次章に記述する。

(1) システム構成と主な作用

坑外システムは，データサーバ（PRIMEQUEST 520：3台，ETERNUS4000 model500：6台），ジョブコントロール用計算機（PRIMERGY BX620 S3：10台），ファイル転送サーバ（PRIMERGY BX620 S3：2台），データ解析装置（PRIMERGY BX620 S3：270台），これらを接続するための基幹ネットワーク（Catalyst6509E）などから構成される。各計算機は基幹ネットワークと複数のギガビットイーサネットで接続され，ネットワークアクセスの高速化を図っている。

データサーバは総容量700 Tバイトのデータ蓄積領域を有し，SRFSでジョブコントロール用計算機およびデータ解析装置に対してファイル共有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，ジョブコントロール用計算機を利用する。ジョブコントロール用計算機には，インテルコンパイラやCPUパフォーマンス解析アプリケーションの一つであるVTuneパフォーマンス・アナライザなどの開発環境が用意されている。また，バッチジョブ運用支援ソフトウェアParallelnavi NQS（Network Queuing System，以下NQS）環境を用意し，開発・デバッグしたプログラムを即座に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すべての作業を一つの端末下で実施することができる。

ジョブコントロール計算機から投入されたジョブ

は，NQSによりデータ解析装置の中から空いているCPUで実行される。利用者が空いているリソースを探す必要はない。データ解析装置を構成するブレードサーバ1筐体は10台のブレードで2本の1000BASE-Tインタフェースを共有している。1筐体にジョブが集中してしまうとネットワークのオーバヘッドが大きくなってしまうため，できるだけ筐体を分散してジョブが投入されるようにNQSの環境設定を行っている。

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。

● 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。

● 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。

● 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。

● 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。

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1080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。

● 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。

● 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1080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。

● 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1080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1080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1080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。

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様におけるデータ観測の概要，データ解析用電子計算機システムの紹介をし，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。

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宇宙素粒子研究施設 助教 小汐由介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://public.web.cern.ch/Public/Welcome.html
(4) CERN：The ZEBRA System.
http://wwwasdoc.web.cern.ch/wwwasdoc/zebra_html3/zebramain.html