国立天文台報　第 21 巻，21–31（2021）

大規模観測データ解析システム
システムの概要と構築・設定、性能評価

磯貝瑞希，古澤久徳，山根 悟，田中伸広，巻内慎一郎，小澤武揚，亀谷和久，
大倉悠貴，高田唯史，小杉城治，岡本桜子

（2020 年 6 月 19 日受付；2021 年 3 月 19 日受理）

Large Scale Data Analysis System
System outline, construction/setting, and performance evaluation

Mizuki Isogai, Hisanori Furusawa, Satoru Yamane, Nobuhiro Tanaka,
Sin’itirou Makiuti, Takeaki Ozawa, Kazuhisa Kamegai, Yuki Okura,
Tadafumi Takata, George Kosugi, Sakurako Okamoto

Abstract

Astronomy Data Center constructs a system to analyze a huge data set, such as large-scale survey data obtained by wide-field cameras including Hyper Suprime-Cam at the Subaru Telescope, that require much computing resources for processing, and operates the system in collaboration with the Subaru Telescope since September 2019.

The system consists of a login node, a management node, 35 compute nodes, a storage with 5 PB capacity and high speed I/O, and file servers with redundant configuration. All nodes other than the management node are connected by InfiniBand with bandwidth of 56 Gbps or more. Computing resources of the system are managed by a job scheduler. Interactive use of compute nodes are prohibited.

Although the computing resources at the start of operation were 280 CPU cores and 5 TB memory, the significant addition of compute nodes in spring 2020 expands the total resources to 1,976 cores and 18.5 TB memory. The additional nodes are currently used for the analysis of HSC-SSP that also serves as trial operation. After adjusting of queue settings, the nodes are to be shortly released for HSC open-use observers.

The performance evaluation tests, conducted before the start of operation, reveal that the speed of reading and writing files is 24–26 GB s^{-1}, and that the processing speed of HSC data is four times faster than that on the batch servers of Multi-Wavelength Data Analysis System.

概要

国立天文台天文データセンターでは，同ハワイ観測所すばる望遠鏡の超広視野カメラ Hyper Suprime-Cam（HSC）など，解析処理に多くの計算資源を必要とする大規模観測データ用の解析システムを構築し，2019年9月よりハワイ観測所と共同で運用を開始している．

本システムはログインノードと管理ノード，総数35台の計算ノード，5 PBの容量と高速I/Oを持つストレージ，冗長構成のファイルサーバから構成されており，管理ノードを除く全ノードは帯域幅56 Gbps以上のInfiniBandで接続されている．システムの計算資源はジョブスケジューラによって管理されており，計算ノードの対話的な使用を禁止している．

運用開始時の計算資源は総CPUコア数が280，総メモリ量が5 TBであったが，2020年春に大幅な計算ノードの増設を実施し，それぞれ1,976コア，18.5 TBまで拡充された．増設分は試験運用を兼ねて，HSCによる大規模サーベイプログラムであるHSC-SSPで得られたデータの解析などに使用しているが，キューの調整などを経て間もなくHSC共同利用観測者に開放予定である．

運用開始前に実施した性能評価試験では，ファイル読み書き速度が 24–26 GB s^{-1}であること，またHSCデータの解析速度が多波長データ解析システムのバッチサーバを使用した場合の約4倍であることを確認している．

1　はじめに

大規模観測データ解析システム（以下 大規模解析システム）[1]とは，国立天文台 天文データセンター[2]が構築を行い，ハワイ観測所と共同で運用している共同利用計算機システムである．昨今の望遠鏡ならびに観測装置の大型化によって観測データの容量も肥大化が進んでおり，これら大規模観測装置によって日々生み出される大容量の観測データを高速に解析する環境が天文学コミュニティから強く求められている．本システムはこの目的のために新たに導入された．同じく共同利用計算機システムとして天文データセンターが運用している多波長データ解析システム（以下 多波長解析システム）[3]は，比較的小規模で対話的な処理や論文作成作業に最適化されており，多波長解析システムと大規模解析システムを相補的に使うことで今後の大規模データによる科学活動を促進することができる．運用初期はすばる望遠鏡 Hyper Suprime-Cam（HSC）[4]の共同利用観測者とHSC を用いたハワイ観測所戦略枠観測プログラム（Subaru Strategic Program）であるHSC-SSPへの解析環境提供を目的とし，システムもこの目的に最適化している．この解析環境提供に関してハワイ観測所と共同で運用に当たっている．

本論文では，システムの紹介と天文データセンターが行ったシステム構築・設定および運用開始前に実施した性能評価試験について記述する．2章でHSC観測データの解析とシステムに求められる性能を，3章でシステム構成を，4章でシステム構築・設定を，5章でシステムの利用方法を，6章で性能評価試験を，7章で今後の展望を述べ，最後の8章で本論文の内容をまとめている．

2　HSC 観測データの解析とシステムに求められる性能

2.1　HSC観測の出力データと解析

超広視野主焦点カメラHSCは計116枚のCCDで構成されている．その内訳は，サイエンス用CCDが104枚，フォーカス用CCDが8枚，オートガイダー用CCDが4枚で，観測で生データとして出力されるのはオートガイダー用以外の112枚分である．HSCで取得される生データは，1CCD分で約18 MB，1回の積分（ショットと呼ぶ）で112枚のCCDから出力される総データ量は約2 GBになる．

本システムには，HSCのデータ解析ツールとして，HSC pipeline（hscPipe）[5]がインストールされている．hscPipeは国立天文台，Princeton大学（アメリカ），東京大学カブリ数物連携宇宙研究機構が共同で開発したHSC用の解析ソフトウェアで，Rubin観測所で計画されているLSSTサーベイ用解析pipelineをベースに作成されており，解析で使用するコマンドはpython scriptの様式で記述されている．hscPipeはHSC生データから較正済み画像と天体カタログを作成可能であり，その基本的な手順は表1に示す通り，5つに分かれている[6].

それぞれの処理では，大まかに以下のような作業を行っている．1の1次処理用データの準備では次のCCD解析で必要となる各種データ（Bias, Dark, Flat, Fringe, Sky）を作成する．2のCCD解析では1で作成したデータを使用し，CCD毎に生データの一次処理と天体カタログを作成する．3の位置とフラックススケールの決定では次の天体データの足し合せの準備として，足し合わせるために必要な各CCDの位置関係を表す座標情報と，明るさの関係を表すフラックススケールの情報を導き出す作業を各フィルター毎に行う．4の背景光補正と天体データの足し合せでは3で作成したフラックススケールファイルとCCDの座標ファイルを用いて，天球面座標を平面座標に投影したデータを生成し，観測された全ショットの積分を行う．そして背景光を引いた後に天体を検出しカタログを作成する．5のマルチバンド解析では4の最後で生成された各フィルター毎のカタログを統合し，そのカタログを元に天体測定を行い新たなカタログを生成する．これらの解析処理のうち，3以外はコマンド単体で分散バッチ処理のためのインターフェースを備えており，1つの解析処理をノードを超えた複数のCPUコアを用いて並列処理することが可能である．

2.2　解析環境としてシステムに求められる性能

2のCCD解析までは各CCD毎に処理を行うため，I/O性能がボトルネックにならない限りにおいて，処理時間の短縮には1ショット分のデータを一度に処理することが有効であり，それには112 CPUコア以上が必要となる．4の足し合せと5のマルチバンド解析はCCD毎に処理を行うわけではないが，処理には多くのCPU演算を必要とする．

HSCデータの解析処理，特に2のCCD解析から5のマルチバンド解析までは多くのメモリを必要とする．その理由は解析処理によって異なるが，その中でメモリ消費が最も大きく成り得る要素はマルチバンド解析での天体測定の処理であり，メモリ消費の程度は限界等級の深さや天体の混み合い具合に依存する．HSC-SSPのデータ解析での実績値として，通常で1プロセ

表1：HSCデータ解析処理 基本的な手順.

番号	手順名	解析処理名（6.2節で使用）
1	1次処理用データの準備	Calibデータ作成
2	CCD解析（CCD毎に一次処理と天体カタログを作成）	一次処理
3	位置とフラックススケールの決定	mosaic
4	背景光補正と天体データの足し合せ	Sky補正，stack
5	マルチバンド解析	マルチバンド解析

ス（=1 CPUコア）当たり8 GB程度，少し混み合った領域や限界等級の深い画像の処理では12 GBから20 GB程度，非常に混み合った領域などではそれ以上（極端なもので1プロセスで1 TB超）のメモリを必要とする．よって計算ノードには1 CPUコア当たり8 GB以上，1ノード当たりでは512 GB以上のメモリが求められ，一部のノードではそれ以上のメモリ搭載が望ましい．

以上の解析処理全体に渡って，多くのファイル入出力が実行されるため処理時間の高速化には高いI/O性能を持つストレージが要求される．特に大量の画像データの入出力を伴う1と2，4の解析処理では顕著である．また1つの解析処理を複数のノードで並列に作業できるよう，ストレージは複数のノードから読み書き可能な形で共有できることが求められる．さらにHSCデータの解析では多くのストレージ容量を必要とする．HSCの1ショットで取得される生データは前述の通り約2 GBのデータ量であるが，hscPipeを用いて解析を行った場合，生データ，解析済みデータとカタログ合わせて最終的には1ショット当たり13 GBのデータ量となる（一例として2晩に300ショットの観測を行っていた場合，解析には約4 TBの容量が必要となる）．

上記は1ユーザによる解析で必要な計算資源についてであるが，共同利用の解析環境としては複数ユーザ（10名程度）の同時利用に耐えるだけの計算資源を用意しておくことが求められる．また，本システムはHSC-SSPへの解析環境提供も重要な目的の一つであり，HSC-SSPの解析を行う際に共同利用ユーザの解析処理に影響が出ないよう，HSC-SSPが使用する分（1,000コア以上，約1 PB/年）も合わせたCPUコア数とそれに耐え得るだけのI/O性能と容量を持つストレージおよびファイルシステムが望ましい．

以上より，HSC観測データの解析環境としてのシステムには（1）HSC-SSPの解析処理と複数（10名程度）のユーザによる解析処理を同時に実行することを可能とする多数のCPUコア数（2,000以上），（2）1 CPUコア当たり8 GB以上，1ノード当たり512 GB以上のメモリ，（3）高いI/O性能を持ち，かつ大容量（SSP使用分1 PB/年とユーザ用0.5 PBで最低1.5 PB以上，できれば前年度のSSP解析結果を保管可能な2.5 PB以上）で共有可能なストレージおよびファイルシステム，が求められる．

3　システム構成

本システムはログインノードと管理ノード，計算ノード，大容量かつ高速I/Oを持つストレージと冗長構成のファイルサーバから構成されている（図1）．

計算資源の効率的な利用のため，計算ノードの対話的な使用を禁止し，計算資源をジョブスケジューラによって管理している．ユーザはログインノードにログイン後，ジョブ投入することで計算ノードを利用可能である．各ノードのスペックと役割は表2にまとめてある．また，システムの構成を図2に示している．

図1：システム全景.

3.1　計算ノード

2020年6月現在，計算ノードは35台あるが，そのスペックは表2の通り4種類に分類され，納入時期も異なる．導入初年度（2018年度）に導入された5台はIntel製CPUを1台当たり4つ搭載しており，1ノード当たりのCPUコア数は56，メモリ量が1 TBである．翌2019年度にはAMD製CPUを搭載するノード26台を導入した．うち2台は第1世代のEPYC CPUを1台当たり2つ，残りの24台は第2世代のEPYC CPUを1つ搭載し，どちらも1ノード当たりのCPUコア数は64，メモリ量は512 GBである．最後に納入されたのは2 TBのNVMe-SSDを搭載する4台で，Intel製CPUを1つ搭載し，1ノードあたりのCPUコア数は 8，メモリ量は128 GBであるが，SSDをswap領域に設定することでCPUコア数よりも高クロック数を求める解析処理や，1プロセスで1 TB超のメモリを必要とするような他の計算ノードでは実行できない解析処理に特化している．

3.2　管理ノードとストレージ

管理ノードは，ユーザ認証，時刻，メール中継，ログ集約，システム監視などの役割を担っている．ストレージは2台のストレージコントローラと9台のJBOD† で構成されており，ファイルシステムはIBM Spectrum Scale（旧称GPFS: General Parallel File System，以下SS）[7]，総容量は5 PBである．なお，ファイルシステムと総容量はこちらで指定した仕様に基づき実施された競争入札により，総容量を含めた性能と導入費用との総合評価で最も高い評価を得た提案が採用されている．

3.3　計算資源とCfCAクラスター

運用開始時の計算資源は総CPUコア数が280，総メ

† Just a Bunch Of Disksの略，ストレージ拡張ユニット．

表2：システム情報.

ノード名	台数	OS	CPU	メモリ	備考
ログインノード	1	RHEL	Intel Xeon Silver 4114 2.2 GHz 10 core × 2	128 GB	PBSサーバ
管理ノード	1	CentOS	Intel Xeon E5620 2.4 GHz 4 core × 1	24 GB	LDAPミラー他
計算ノード	5	RHEL	Intel Xeon Gold 6132 2.6 GHz 14 core × 4	1 TB	
	2	CentOS	AMD EPYC 7601 2.2 GHz 32 core × 2	512 GB	
	24	CentOS	AMD EPYC 7742 2.25 GHz 64 core × 1	512 GB	
	4	CentOS	Intel Xeon W-2145 3.7 GHz 8 core × 1	128 GB	2 TBのNVMe-SSD搭載
ファイルサーバ	2	RHEL	Intel Xeon Gold 5122 3.6 GHz 4 core × 1	64 GB	SSサーバ

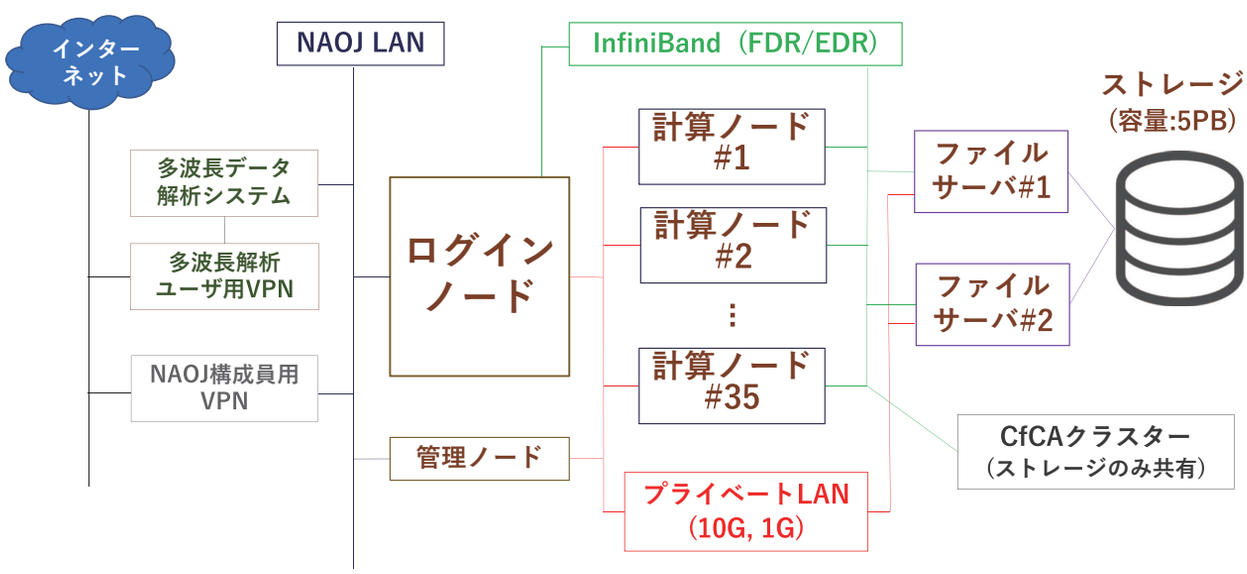


図2：システム構成図.

モリ量が5 TBであったが，2020年春に大幅な計算ノードの増設を実施し，現状の 総CPUコア数 1,976，総メモリ量 18.5 TBまで拡充されている．本システムに備わる計算ノードに加えて，本台天文シミュレーションプロジェクト（CfCA）[8]との協力体制のもと，CfCA管理のPCクラスター（総CPUコア数1,056）をSSクライアントとして追加することで，HSC-SSPデータ解析専用の拡張計算資源として利用している．

3.4　ネットワーク

本システムを構成するノード・サーバで国立天文台ネットワーク（図2の NAOJ LAN）に直接接続されているのは，ログインノードと管理ノードである．それ以外のノード・サーバはシステム内のプライベートネットワークにのみ接続されている．プライベートネットワークは一般系と管理系の2系統あり，一般系は10 Gbps，管理系は1 Gbpsの帯域幅である．一般系はファイルシステム（SS）以外のノード間高速通信を担い，home領域の共有やジョブスケジューラなどが利用している．管理系はユーザ認証，メール，IPMI，システム監視（snmp）などが利用している．

3.5　InfiniBand

管理ノード以外の全ノードは帯域幅56 Gbps以上のInfiniBand（FDR/EDR）で接続されており，ノード間の超高速データ通信が可能である．InfiniBandを使用するのはSSのみとしている．

3.6　多波長解析システムとの共有

本システムは多波長解析システムとアカウント情報およびVPNを共有している．このため，本システムの利用者は多波長解析システムのアカウントが必須となる．また，国立天文台外のネットワークからの本システムへのアクセスにはVPNを使用する必要があるが，ユーザが国立天文台構成員ではない場合，多波長解析システムのVPNサービスを利用可能である．

3.7　OS

本システムのOSは Red Hat Enterprise Linux（以下 RHEL）7 [9] およびその非商用版クローンであるCentOS 7 [10]を採用している．運用開始前に導入されたノード・サーバ（ログインノードと計算ノード，ファイルサーバ）がRHELで，管理ノードと運用開始後に増設した計算ノードがCentOSである．運用開始前に導入されたノード・サーバのOSでRHELが採用された主な理由はファイルシステム（SS）のサポート対象OSだからであり，運用開始後に増設した計算ノードでCentOSを採用した主な理由はhscPipeのサポートおよび開発実証環境のOSだからであるが，計算ノード増設前にCentOSで試験システムを構築し，ファイルシステムその他の動作や運用に影響がないこと，また運用システムで計算ノードのOSを混在させても同じく影響がないことを確認している．

3.8　ジョブスケジューラ

ジョブスケジューラは多波長解析システムで運用経験のあるPBS Professional（以下PBS）のオープンソース版を採用している．ログインノードがPBSサーバ兼ジョブ投入ノード，計算ノードがPBS実行ノードである．

4　システム構築・設定

本システムを構成する機器は複数回の調達によって納品されており，その形態・設定の程度も様々であるが，大きく分けて，部材納品，OSインストール済み納品，SSインストール・設定済み納品の3種類にまとめることができる．それぞれの納品の内訳は，部材納品はAMD EPYC CPU搭載の計算ノード全26台，OSインストール済み納品はログインノードと計算ノード7台（Intel Xeon Gold 3台とXeon W-2145 4台），SSインストール・設定済み納品は計算ノード2台とファイルサーバである．

部材納品の機器については組立から（図3），OSインストール済みの機器やSSインストール・設定済みの機器についてはその後の設定から我々が担当した．また管理ノードについては既存サーバをHDD交換の上でOSインストールから実施した．

OSインストール後に行う主な設定としては，全ノード共通ではセキュリティ強化，時刻同期，メール配送設定などがあり，加えてログインノードと計算ノード共通ではSSのインストールおよびクライアント設定，ユーザ認証設定やhome領域のNFSオートマウント設定，さらにログインノードでPBSサーバ設定，計算ノードでPBS実行ノードの設定がある．

管理ノードはメール中継，時刻，ログ集約，システム監視，LDAPミラー，ウイルス対策ソフトのミラーサーバなど各種サービスを提供するための設定を行っている．ユーザ認証は多波長解析システムのLDAP認証を共有し，追加のグループ設定を行うことで大規模

図3：AMD EPYC ノード組立後の様子.

解析システムへのアクセス許可を与えている．管理ノードが多波長解析システムのLDAPサーバのミラーとなり，ログインノードおよびプライベートネットワーク下の計算ノードの参照先としている．

4.1　SSの導入と設定変更

ログインノードと計算ノードはSSのインストール後，SSクラスターへの追加やSSクライアントとしての登録を行い，ノードによって異なるInfiniBandデバイス名とポート（RDMA‡で使用する）や，同じくノードによって異なるメモリ量に合わせたページプール設定など，SSに関する設定変更を実施した．更にSS領域の使用状況をユーザ毎に把握するため，SSのクォータ機能を有効にする設定変更も行っている．

4.2　home領域共有のための設定変更

home領域はストレージコントローラの機能によって，ストレージ内のSSとは別の領域がNFSプロトコルで提供されている．この領域をログインノードと計算ノードで共有するよう設定している．ストレージコントローラは設置時から極力設定を変更しないようにしているが，クォータ機能の有効化とユーザの初回ログイン時にホームディレクトリが自動生成されるよう変更を行っている．

‡ Remote Direct Memory Accessの略．OSを介さずにあるノードから別のノードのメモリへ直接アクセスする技術.

4.3　PBSの導入と設定

ログインノードと計算ノードにPBSを導入し，それぞれサーバ用と実行ノード用の設定を行った後，ジョブの実行結果の標準エラー出力をユーザ領域にファイル出力できるようにするため，PBSサーバと実行ノード間でSSHホストベース認証の設定を行っている．PBSに関する設定のうち，ノードの登録や設定，キューの作成や設定などはPBSサーバ上でqmgrコマンドを実行することで行っている．PBSは実行中のジョブが使用している計算資源を常に監視しており，あらかじめ指定した割当量を超える計算資源を使用したジョブの扱いは各実行ノード上の設定ファイルで定義可能である．本システムではCPUコア数とメモリ量について指定した割当量を超過したジョブを強制終了するよう設定している．

5　システムの利用方法

前述の通り本システムは計算資源の効率的な利用のため，計算ノードの対話利用を禁止しジョブスケジューラによって計算資源を管理している．ユーザが対話的に利用可能なノードはログインノードのみであり，ユーザはこのノードにログインし，ジョブ投入することで計算ノードを利用可能である．ログインノードは，国立天文台内のネットワークからであれば直接SSHでログイン可能である．台外ネットワークからの場合，国立天文台構成員であれば情報セキュリティ室が提供しているVPNサービスを，それ以外のユーザは多波長解析システムが提供しているVPNサービスを利用することで，外部からのSSHによるログインが可能である．

ジョブは，ユーザが用意したジョブ投入用のスクリプト（以下 ジョブスクリプト）をqsubコマンドで実行することで投入可能である．ジョブスクリプトの記述は，本システムが採用したジョブスケジューラの商用版を採用している多波長解析システムと基本的には同じである．HSC観測データの解析処理であれば，hscPipeが出力するジョブスクリプトを元にPBS指示文を必要に応じて追加して完成させる．ジョブ投入に使用するキューはユーザが利用可能な計算資源によって分かれており，ハワイ観測所との協議によって決定されるHSC共同利用観測者であればqmキューを使用する．このキューは2020年6月現在の設定では 最大で

112 CPUコアと 約2 TBのメモリを1ジョブ当たり15日間使用可能である．なお，キュー構成と利用可能な計算資源については今後の計算ノード増設分の開放に伴い変更の可能性がある．

6　性能評価試験

システム全体の性能を評価するため，ファイルシステムに関する2種類の性能評価（ファイルの読み書き速度とメタデータ操作速度，システムにおけるファイルシステムのトータルスループットの測定）と実際のHSC観測データの処理時間を測定する試験を実施し，多波長解析システム開発系での結果と比較した．本試験はユーザを排除して行う必要があるため，運用開始前に計算ノードが2台構成および5台構成で実施した結果を掲載している．既存HSC観測者がデータ解析を行ってきた環境と比較することが望ましいが，従来ハワイ観測所が提供してきた小型PC（32 CPUコア）はユーザの活動があり，またユーザが他に利用可能な多波長解析システムも運用系については同様の状況のため，条件を満たす多波長解析システムの開発系を，必要に応じて運用系を模した構成にして使用した．

多波長解析システム運用系は，対話型解析サーバ32台，バッチ型解析サーバ2台と総容量1.6 PBの共有ストレージ他で構成されている．対話型解析サーバはkaim系とkaih系の2種類あり，両者は搭載メモリ量（kaim系は192 GB，kaih系は256 GB）とローカルストレージの容量（kaim系は12 TB，kaih系は51 TB）および仕様が違うものの，搭載CPUは同じ（Intel Xeon E5-2667v4, 3.2 GHz 8コアを2つで1台あたり16コア）である．バッチ型解析サーバはローカルストレージがない点以外，kaim系サーバと同じ仕様である．

多波長解析システム開発系は運用系のkaih系対話型解析サーバと同じCPUとメモリ量（256 GB）を搭載したサーバ3台から構成されている．図4に示す通り，試験に用いたストレージは1台のサーバと16 Gbpsのファイバーチャネル2本で接続されており，領域の容量は12 TB，ファイルシステムはXFSである．この領域は10 GbpsのLANを通してNFSバージョン3プロトコルで他の2台のサーバと共有されている．

ファイルシステムの性能評価試験では大規模解析シ

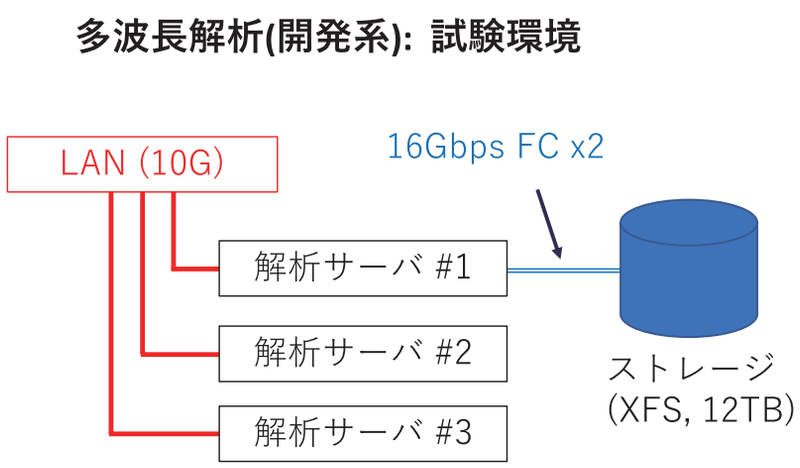


図4：多波長解析システム（開発系）試験環境.

表3：性能評価試験 試験環境比較一覧.

試験環境名	大規模解析（5ノード）	大規模解析（2ノード）	多波長解析（開発系）
CPU	Intel Xeon Gold 6132 2.6 GHz 14core × 4		Intel Xeon E5-2667v4 3.2 GHz 8 core × 2
メモリ	1 TB		256 GB
台数	5	2	3 (2*)
総コア数	280	112	48 (32*)
ファイルシステム	SS		XFS + NFSv3

*「HSCデータ処理速度」試験での使用台数と総コア数.

ステムと並列数を近づけるためにサーバ3台全てを使用し，HSCデータ処理速度の試験では，多波長解析システム運用系のバッチサーバと同じ台数（2台，図4の解析サーバ#1と#2）を使用した．

以上3種類（大規模解析システム 5ノードと2ノード，多波長解析システム開発系）の試験環境を比較した一覧を表3にまとめている．

6.1　ファイルシステムの性能評価

（1）ファイル読み書き速度

分散ファイルシステムの性能評価に適したベンチマークソフト IOR [11]のバージョン2.10.3を使用して，ファイルシステムのWriteとRead速度を測定した．図5がその結果で，単位は$\mathrm{MB\,s^{-1}}$，縦軸はログスケールである．I/Oの並列数は総CPUコア数と同数とし，大規模解析システムでは計算ノード5ノードで280，2ノードで112，多波長解析（開発系）では48である．I/Oの1プロセスで1ファイルを読み書きすることとし，1ファイルのサイズは解析中のHSCデータの典型的なファイルサイズである 100 MiB（1 MiB = 2^{20}バイト）に，また 1 I/O Call当たりのデータ転送サイズは1 MiB とした．試験開始前に，試験に使用する全計算ノードのキャッシュをクリアし，測定はWrite，Readを1セットとしたものを10セット連続で実施している．本試験により，大規模解析5ノード（SS）のWriteおよびRead速度は24–26 $\mathrm{GB\,s^{-1}}$と並列数の増加（112から280への）による性能劣化が見られず，ファイルシステム納品時と同等の性能を保持し，かつ多波長解析（開発系，NFS，並列数48）とReadで10倍以上，Writeで約80倍以上であることを確認した．

（2）ファイルメタデータ操作速度

分散ファイルシステムの性能評価に適したベンチマークソフト mdtest [11]のバージョン1.9.3を使用してファイルシステムのファイルメタデータ操作（作成，状態確認，読み込み，削除）速度を測定した．図6がその結果で，単位は操作数$\mathrm{s^{-1}}$，縦軸はログスケールである．操作の並列プロセス数はファイル読み書き速度と同じで，1プロセス当たりのファイル作成数は1,000 としている．ファイル読み書き速度と同様，試験開始前に試験に使用する全計算ノードのキャッシュをクリアし，測定は4つの操作を1セットとしたものを3セット連続で実施している．ファイルメタデータの操作速度は操作内容によって大きく異なるが，全ての操作で並列数の増加（112から280への）による性能劣化が見られず，ファイルシステム納品時と同等の性能を保持し，かつ多波長解析（開発系，NFS，並列数48）の20倍を超えていることを確認した．

6.2　HSCデータ処理速度

HSCの観測生データからマルチバンド解析による検出天体カタログ作成までの一連の解析処理を行い，その処理時間を測定した．図7がその結果で，単位は秒である．解析はHSCデータの解析処理ソフトhscPipeのバージョン6.7を使用し，生データはHSC解析講習会のウェブページ[12]で公開されているもの（112 CCD版，3バンド）を用いた．BiasとDarkは5ショットを合成して作成した（全バンド共通で1 CCD当たり5枚）．また，Flat とサイエンス画像は各バンド毎に5ショットを合成して作成した（1バンド1 CCD当たり5枚）．図7の各解析処理と2.1節の基本的な手順との対応関係は表1に示す通りである．これらの処理のうち，Calibデータ作成内のFlat作成とFringe作成，Sky作成，mosaic，stackが1バンドにつき1ジョブ（3バンドで3ジョブ，Fringe作成のみ3バンドのうちの1バンドが対象のため1ジョブ）の実行に対し，Calibデータ作成内のBias作成，Dark作成と一次処理，Sky補正，マルチバンド解析が3バンドで1ジョブの実行である．各解析処理での投入ジョブ数と使用したキューの一覧は表4にまとめている．ここでキュー名は「q」+最大使用可能コア数を意味しており，q112 であれば最大で112 CPUコアを使用可能で，試験ではジョブ投入時に用いたキューの最大使用可能コア数まで使用している．多波長解析（開発系）での試験ではq16キューを使用している．これは運用系でユーザが利用可能な複数のキューの中で，最も多くのCPUコア（16コア）を使用可能である．大規模解析（5ノード，総コア数280）で1バンドにつき1ジョブ投入する解析処理では，q112キューを用いたジョブ2つとq56キューを用いたジョブ1つを同時に投入している．これは，3つのジョブ全てでq112キューを用いると計算資源不足で1つのジョブが待ち状態になるのを回避するためである．

図7より，大規模解析システム5ノードでの解析は運用系を模した構成である多波長解析（開発系）のバッチサーバ（2台，32 CPUコア）を使用した場合の約4倍の速さで完了する事を確認できる．一方で大規模解析システム同士での比較では，総積算処理時間ではCPUコア数の差（5ノード:280に対して 2ノード:112）ほどの速度差が見られない．これは，Calibデータ作成内のFlat作成やSky作成，一次処理，mosaic，stackのように用いた総CPUコア数に応じて処理時間が短くなる解析処理がある一方で，Calibデータ作成内のBias作成，Dark作成，Fringe作成やSky補正のように5ノードと2ノードで同じ条件（キューと投入ジョブ数）で実施

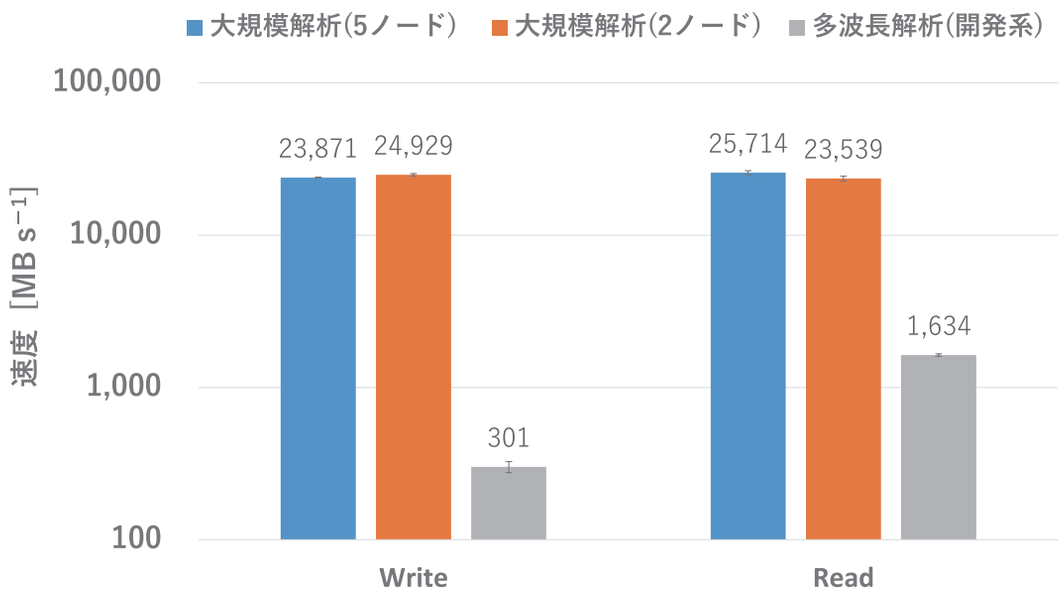


図5：ファイル読み書き速度の結果.

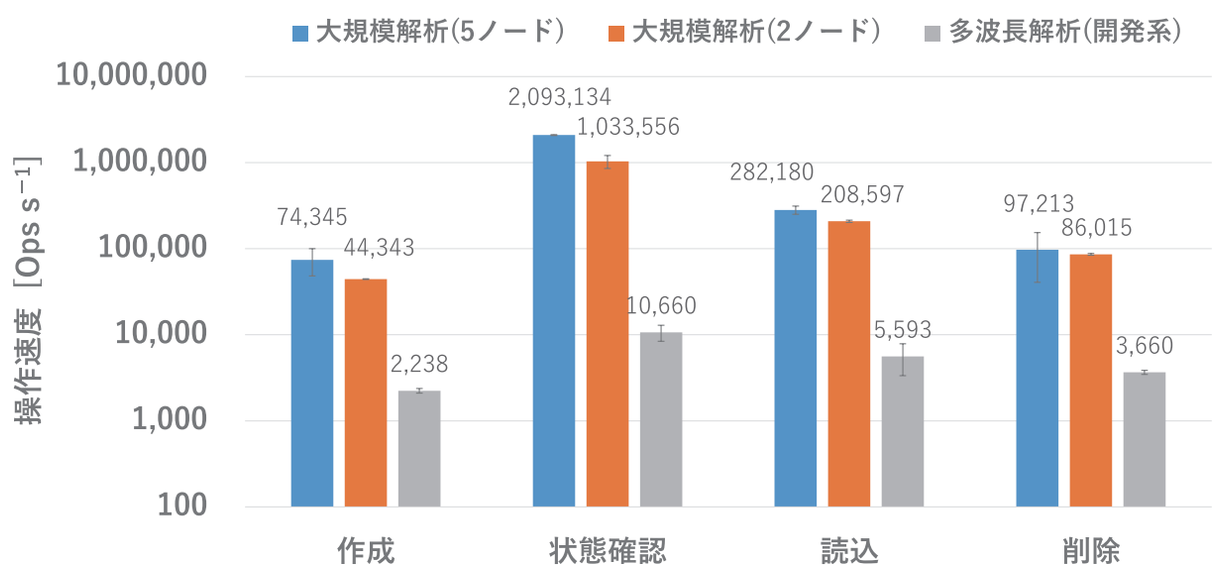
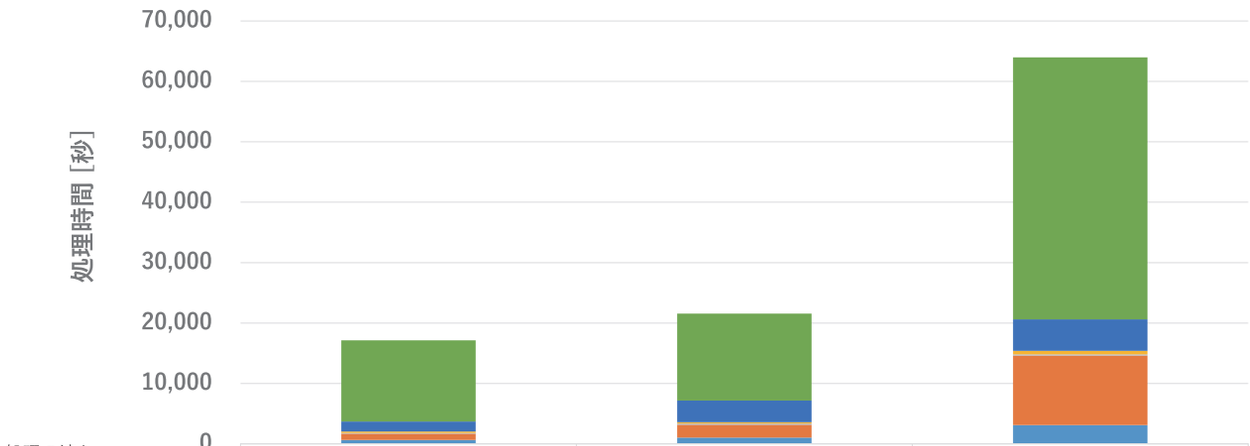


図6：ファイルメタデータ操作速度の結果.

している解析処理や，マルチバンド解析のようにより多くのCPUコア数（5ノードのq280に対して2ノードのq112）を用いているにも関わらず，処理時間に使用コア数の差程違いが出ない処理が含まれているためである．特に総処理時間で多くを占める解析処理はマルチバンド解析であるが，大規模解析システムの2ノードから5ノードで解析処理に用いるCPUコア数を2.5倍にしても，処理時間はわずか7％しか減少していない．その理由の一つとして考えられるのは，マルチバンド解析内では4つの解析サブ処理（カタログ統合，天体再測定，新カタログ構築，最終カタログ作成）が行われており，それぞれの解析サブ処理は並列処理が可能であるが，その並列数には限りがあること（フィルター数×分割天域数で，本試験では233．なお天域の分割はフィルター毎に行われるため，本試験のように分割天域数がフィルターによって異なる場合がある），またある解析サブ処理内の全ての並列処理が完了してから次の解析サブ処理に進むため，並列処理のどれか一つで処理時間が長い場合，並列数を増加させても処理時間の減少にそれほど寄与しなくなることが挙げられる．その程度は解析処理を行う画像データ次第であるが，CPUコア数を多く使用しても処理時間にさほど寄与しない場合があるというのは本試験で明らかとなった重要な結果の一つである．



	大規模解析（5ノード）	大規模解析（2ノード）	多波長解析（開発系）
マルチバンド解析	13,424	14,404	43,389
stack	1,675	3,572	5,197
Sky補正	238	235	582
mosaic	108	186	202
一次処理	1,040	2,143	11,507
Calibデータ作成	642	1,028	3,109

図7：HSC データ処理速度の結果.

表4：HSC データ処理速度測定 ジョブ投入数と使用したキューの一覧.

処理名	投入ジョブ数	ジョブ投入に使用したキュー		
		大規模解析（5ノード）	大規模解析（2ノード）	多波長解析（開発系）
Calibデータ作成	9	q112 × 7, q56 × 2	q112	q16
一次処理	1	q280	q112	q16
Sky補正	1	q112	q112	q16
mosaic	3	q56	q56	q16
stack	3	q112 × 2, q56 × 1	q112	q16
マルチバンド解析	1	q280	q112	q16

7　今後の展望

7.1　増設計算ノードのHSC共同利用観測者への開放

2020年6月現在，増設計算ノードの利用はHSC-SSPデータの解析処理などに制限しているが，間もなくHSC共同利用観測者へ開放する予定である．現状ではそれぞれで専用のPBSキューを設定しているが，既存計算ノードと増設ノードではコア当たりのメモリ量が異なるため，開放にあたり検討の上キューの調整を行う可能性がある．

7.2　HSC共同利用観測者以外の受け入れ

計算ノードの増設により本システムの計算資源はCPUコア数で7倍，総メモリ量で3倍超と大幅に拡充され，HSC-SSPデータなどに対する大規模な解析処理が実施されない期間では，システム負荷が下がると期待されるので，計算資源を有効に活用することが重要である．そこで，利用状況を見ながらではあるものの，受け入れ対象を当該セメスターのHSC共同利用観測者だけでなく，過去のHSC観測者やHSCアーカイブデータ利用者，さらにはそれ以外の大規模な計算処理用途の希望者などへ広げることを検討中である．ただ

し後者については，必要なソフトウェアの導入やジョブスケジューリングの調整など，2020年6月現在HSCの解析処理に最適化されているシステムの設定変更が必要とされるため，受け入れまでにある程度の検討と準備期間が必要と思われる．

8　まとめ

国立天文台天文データセンターでは，同ハワイ観測所すばる望遠鏡のHSCなど，解析処理に多くの計算資源を必要とする大規模観測データ用の解析システムを構築し，2019年9月よりハワイ観測所と共同で運用を開始している．本システムはログインノードと管理ノード，総数35台の計算ノード，容量5 PBのストレージと冗長構成のファイルサーバから構成されており，システムの計算資源は2020年6月現在，総CPUコア数が1,976，総メモリ量が18.5 TBである．本システムは同じく共同利用計算機システムとして天文データセンターが運用する多波長解析システムと相補的な関係にあり，両システムでアカウント情報を共有している．このため，本システム利用者は多波長解析システムのアカウントが必須となる．運用開始前の計算ノード2台および5台構成時に実施した性能評価試験で，ファイル読み書き速度が 24–26 GB s^{-1}であること，またHSCデータの解析速度が多波長解析システムのバッチサーバを使用した場合の約4倍であることを確認した．2020年6月現在，本システム利用者をHSC共同利用観測者とHSC-SSPデータの解析処理に制限しているが，今後受け入れ対象を広げることを検討中である．

謝辞

HSCデータ解析環境やジョブスケジューラを含む本システムの構築にあたり，ハワイ観測所や天文シミュレーションプロジェクト，天文データセンターの皆様から助力と助言を頂いたことに感謝する．特に天文シミュレーションプロジェクトの伊藤氏には，本システムの導入前の設計段階から構築後の運用やPCクラスターの連携に至るまで多大なる助力と助言を頂いており，深く感謝の意を表する．また本システムの構築には多くのオープンソースソフトウェアを利用している．これらの開発者の皆様に感謝する．

参考文献

[1] 大規模観測データ解析システム, https://www.adc.nao.ac.jp/LSC/
[2] 国立天文台 天文データセンター, https://www.adc.nao.ac.jp/
[3] 多波長データ解析システム, https://www.adc.nao.ac.jp/MDAS/
[4] Hyper Suprime-Cam (HSC), https://www.naoj.org/Observing/Instruments/HSC/index.html
[5] Bosch, J., et al: The Hyper Suprime-Cam software pipeline, *PASJ*, **70**, Issue SP1, id.S5 (2018)
[6] HSC pipeline manual, https://hsc.mtk.nao.ac.jp/pipedoc/
[7] IBM Spectrum Scale, https://www.ibm.com/products/spectrum-scale
[8] 天文シミュレーションプロジェクト, https://www.cfca.nao.ac.jp/
[9] Red Hat Enterprise Linux, https://www.redhat.com/ja/technologies/linux-platforms/enterprise-linux
[10] CentOS, https://www.centos.org/
[11] IOR, https://ior.readthedocs.io/
[12] HSC解析講習会用チュートリアル, https://hsc.mtk.nao.ac.jp/HSC_Training_tutorial/