国立天文台・天文データセンター　大規模観測データ解析システム III

○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(ADC)、
大倉悠貴、岡本桜子(ハワイ観測所)、髙田唯史、小杉城治(ADC)

概要(Abstract)

国立天文台天文データセンターでは、ハワイ観測所すばる望遠鏡の超広視野カメラ 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、 運用を開始している。本システムは大容量かつ高速 I/O を持つストレージと総コア数 1,976 の計算ノード他から構成されており、演算性能は今年度実施の増設で大幅に増強された。本講演ではシステムの概要、計算ノードの増設、増設後に実施した性能評価試験、現在の運用状況と今後について報告する。

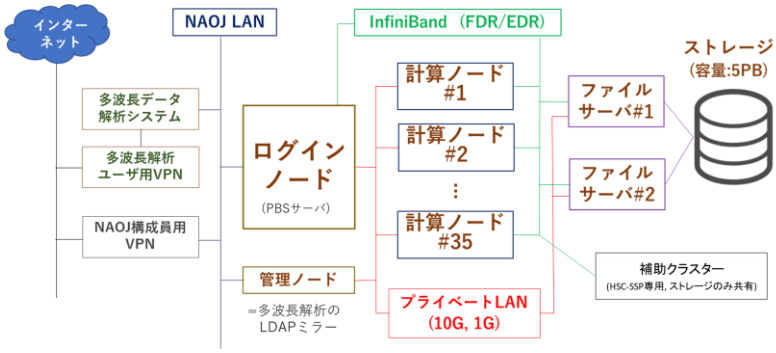
1. システムの概要

大規模観測データ解析システム(以下本システム、図 1)とは、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で、HSC を用いたハワイ観測所戦略枠観測プログラム(HSC-SSP)を含む HSC 共同利用観測者への解析環境提供が初期の主な目的である。このため、初期運用中はシステムを HSC 観測データの解析処理に最適化し、ユーザは HSC 共同利用観測者(PI/CoI)と HSC 観測データ(過去の観測やアーカイブを含む)解析者(以下一般ユーザ)に限定している。本システムの構築は天文データセンター(ADC)、運用は ADC とハワイ観測所(HSC 共同利用+SSP 分)が担当している。

図 1 大規模解析システムラック列

本システムは、ログインノード 1 台、計算ノード 35 台、ファイルサーバ 2 台、ストレージ、管理ノード 1 台で構成される。OS は Red Hat Enterprise Linux 7 またはその非商用版クローンの Cent OS 7 である。計算ノードは仕様の異なる 4 タイプの計算機で構成されており、全 35 台の総コア数は 1,976、総メモリ量は 18.5TB である。ストレージは容量 5PB でファイルシステムは IBM 社の Spectrum Scale である。図 2 にシステム構成図を、表 1 に計算ノードの仕様一覧を示す。

図 2 システム構成図



本システムは、ADC が運用し国内外の研究者へ共同利用サービスとして解析環境を提供している「多波長データ解析システム(以下多波長解析)」とユーザ情報を共有しており、システムの利用には多波長解析のアカウントを必要とする。また、計算資源の効率的な利用のため、本システムでは計算ノードの対話的使用を禁止し、計算資源はジョブスケジューラで管理している。ユーザはログインノードからジョブを投入することで計算ノードを使用する。ユーザが利用可能な計算資源とその割り当ての優先度・期間は表 2 に示す通りユーザタイプによって異なり、HSC 共同利用観測者は優先度が中（または高）で、優先利用期間は利用宣言開始から 1 年間(インテンシブプログラムの場合はプログラム最終セメスター終了から 1 年間)、一般ユーザは優先度が低、利用期間は最大 1 年間(ただし更新可)である。ジョブ投入の際に使用するキューもユーザタイプに応じて用意しており、現在のキュー構成は表 3 に示す通りで、HSC 共同利用観測者は 1 ジョブ当たり最大で 112 コア、1,800GB のメモリを 15 日間利用可能な qm キューを、一般ユーザは 1 ジョブ当たり最大で 32 コア、450GB のメモリを 15 日間利用可能な ql キューを使用可能である。上記キュー以外にも、1 プロセスで 1TB 超のメモリを必要とする解析用の qhm キューやテスト用の qt キューを用意しており、これらのキューは全ユーザが使用可能である。

表 1　計算ノード仕様一覧

ノード名	台数	OS	CPU	メモリ	総コア数	総メモリ量
an[01-05]	5	RHEL7	Intel Xeon Gold 6132 2.6GHz 14core x4	1TB	280	5TB
an[06-07]	2	CentOS7	AMD EPYC 7601 2.2GHz 32core x2	512GB	128	1TB
an[08-31]	24	CentOS7	AMD EPYC 7742 2.25GHz 64core	512GB	1,536	12TB
an[91-94]	4	CentOS7	Intel Xeon W-2145 3.7GHz 8core	128GB +2TB swap	32	512GB +8TB swap
総数	35				1,976	18.5TB +8TB swap

表 2 利用可能な計算資源・利用期間

ユーザタイプ	アカウント取得方法	利用可能な計算資源	資源割当の優先順位	利用可能なキュー	利用可能期間
HSC-SSP	---	~ 2,000コア	高	qssp, qt, qhm	~2か月x2/年
HSC共同利用観測者（インテンシブ含む）	ハワイ観測所経由	112コア	中 - 高	qm/qh, qt, qhm	利用宣言後1年間（プログラム終了+1年）
一般（過去のHSC観測者、アーカイブ利用者など）	ADCへ利用申請（随時）	32コア	低	ql, qt, qhm	最大1年間（更新可）

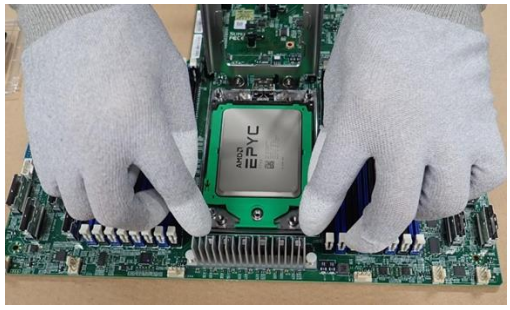
表 3 システムのキュー構成（現状）

キュー名	優先順位	CPUコア数		メモリ量 [GiB]		同時実行可能ジョブ数		実行可能ノード
		最大	デフォルト	最大	デフォルト	ハード	ソフト	
qssp	最高	1,944	56	16,200	450	---	---	01-31
qh	高	280	56	4,500	450	---	---	
qm	中	112	56	1,800	450	---	1	
ql	低	32	28	450	225	---	1	
qt（テストキュー）	最高	4	4	64	64	1	1	
qhm（要1TB超メモリの解析専用）	中	32	8	7,960	1,990	---	1	91-94

2. 計算ノードの増設

本システムは 2019 年 10 月に計算ノード 5 台構成で運用を開始し、その後 2020 年 4 月に計算ノード 30 台を増設、約 3 か月の HSC-SSP 専有利用またはユーザを限定した試験運用の後、同 7 月に HSC 共同利用観測者に増設分を開放している。この増設分 30 台のうち CPU に AMD EPYC を採用した 26 台はパーツを調達・自作(図 3)し、筐体なしで運用することで購入費用を圧縮している。この 26 台分の導入にあたり、2019 年度前半に試験用として 2 台分のパーツを調達し、組立・構築・動作試験で動作確認と経験を積んだ後、2019 年度後半に 24 台分の調達・組立・構築・動作試験を実施している。更に 2019 年度末には運用中および調達済みの計算ノードでは実行できない、1 プロセスで 1TB 超のメモリを必要とする解析

図 3 CPU 取付作業の様子

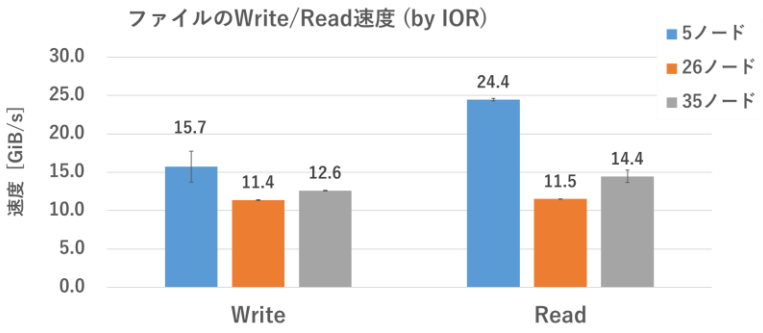


用に、2TB の NVMe SSD を搭載した計算機 4 台を調達・構築し、SSD を swap 領域とした上で動作試験を実施、ユーザを限定した試験運用の後に全ユーザへ開放している。

3. 性能評価試験

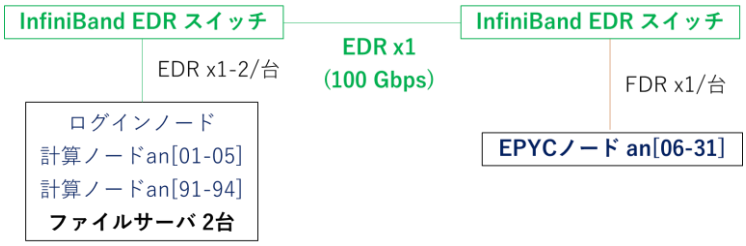
計算ノード増設後に、並列処理数増加による I/O 性能への影響確認を目的とした性能評価試験を実施した。試験は分散ファイルシステムのベンチマークソフト IOR を使用したファイル読書速度測定で、総コア数と同数のファイル(容量:100MiB)のシーケンシャル書込/読込を並列実行し、この実行を 10 回繰り返したものを 1 セットとし、それを 2 セット試行している。試験を実施するノードは、増設前から運用していた 5 台、増設した 30 台のうち AMD EPYC CCD を搭載した 26 台、全 35 台の 3 種類に分けている。図 4 がその結果である。

図 4 性能評価試験結果



左側が Write の結果、右側が Read の結果で、縦軸は GiB/s 単位の速度である。この図より、並列数を増加させると Write/Read ともに速度が低下するが、最も速度が低いのは最も並列数が多い全 35 台ではなく、全 EPYC ノード 26 台であることがわかる。この速度低下の原因は、システム内で使用している 2 台のインフィニバンドスイッチ(図 5 に示す通り、一方に EPYC ノード 26 台が、もう一方にファイルサーバを含み管理ノードを除く残りの全ノードが接続されている)間の接続が EDR 規格のケーブル 1 本で、その帯域幅(100Gbps、実効転送レートでは 11.3GiB/s)で制限を受けているためとほぼ特定できている。この改善は今後の課題であるが、実際の運用においては試験のように多数の I/O 処理のタイミングが一致し、実効転送レートの上限で制限を受ける状況がまだ起きていないことをシステム監視等で確認している。

図 5 インフィニバンド接続図



4. これまでの運用状況と今後

2020 年 10 月より HSC データ解析に限定した一般ユーザの受け入れを開始しており、現在のアカウント保持者:は 14 名である。2019 年 10 月の運用開始から 2020 年 12 月末までの 15 か月間で、17,459 ジョブが実行され、ジョブ実行の積算 CPU 時間は 18,749 日、12 月末時点でのストレージ使用量は総容量 5PB に対し 2.9PB の利用状況であった。

今後については今年度中に計算ノードを 5 台増設予定であり、また HSC に限定しないデータ解析希望者の受け入れを検討中である。