

大規模観測データ解析システム(LSC)

現状と検討課題

磯貝瑞希 (国立天文台天文データセンター)

LSC運用チーム

内容

☆ 大規模観測データ解析システムの概要と現状

(システムの特徴、構成、優先権とリソース、キュー構成...)

☆ 検討課題

大規模観測データ解析システム(LSC)とは

- ☆ 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
- ☆ 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。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観測+HSC-SSP)が担当
- ☆ 試験運用はHSC観測データの解析処理に最適化し、計算ノード5台で開始。計算資源の問題でユーザは当該セメスターのHSC共同利用観測者に限定。その後計算資源の拡充とともに受け入れ対象を「アーカイブ含むHSC観測データ解析者」、「HSCに限定しないデータ解析者」に拡大
- ☆ アカウントはMDASと共有。効率的な運用のため、計算資源はジョブ管理ソフト(OpenPBS)で管理 → **ユーザによる計算ノードの対話的利用を禁止**

大規模解析システムラック列



LSC 拡充の推移

- 2019年 9月 計算ノード 5 台(総CPUコア数:280)構成で試験運用開始
(S19Bよりプロポーザル採択者(内の利用希望者)の受入開始)
→ 計算資源の問題で、当該セメスターのPI・CoIのみ利用申請可
- 2020年 4月 計算ノード30台(SSD搭載4台を含む)を増設
- 2020年 7月 増設分(30台)をHSC観測ユーザに開放 → 計算ノード35台体制
- 2020年10月 本運用開始。HSCデータ(アーカイブ含む)解析希望者の受け入れ開始
- 2021年 4月 計算ノード5台を増設。現在の40台体制
- 2022年 1月 HSCに限定しないデータ解析希望者の受け入れ開始

システム情報

種類	台数	OS	CPU	メモリ[GB]	総コア数
ログインノード	1	RHEL7	Intel Xeon Silver 4114 2.2GHz 20core (10core x2)	256	20
計算ノード	40	CentOS7	Intel Xeon/ AMD EPYC	1024/512/12 8	2,296
ファイルサーバ	2	RHEL7	Intel Xeon Gold 5122 3.6GHz 4core	64	8

+管理ノード

ストレージ:

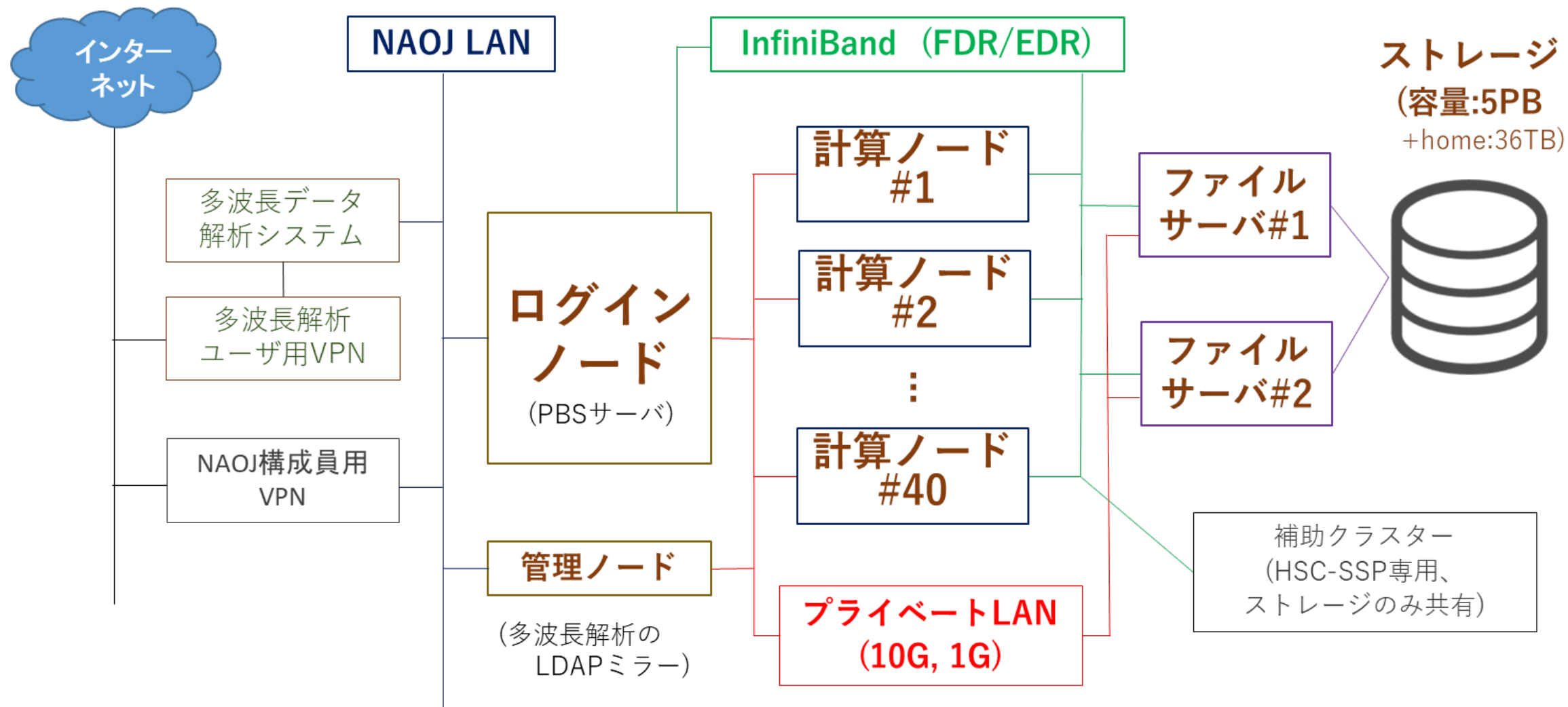
- home領域: 36TB (quota: 200GB)

ファイルシステム: XFS+NFSv3

- 作業領域: **5PB** (quota:現状未設定)

ファイルシステム: IBM Storage Scale (旧GPFS)

システム構成図



計算ノードの内訳

ノード名	台数	OS	CPU	メモリ	総コア数	総メモリ量
an[01-05]	5	CentOS7 (7.6/7.7)	Intel Xeon Gold 6132 2.6GHz 56core (14Cx4)	1TB	280	5TB
an[06-07]	2		AMD EPYC 7601 2.2GHz 64core (32Cx2)	512GB	128	1TB
an[08-36]	29		AMD EPYC 7742 2.25GHz 64core	512GB (22) 1TB (7)	1,856	19TB
an[91-94]	4		Intel Xeon W-2145 3.7GHz 8core	128GB +2TB swap (SSD)	32	512GB +8TB swap
総数	40				2,296	25.5TB +8TB swap

an[91-94]: 他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用

利用可能な計算資源・利用期間

利用可能な計算資源とその割当の優先度はユーザタイプによって異なる

ユーザタイプ	利用申請先	利用可能な 計算資源	優先順位	キュー	計算ノード	(優先)利用可能期間
HSC-SSP	---	～ 1,500コア	高	qssp	an[13-36] (an12は試験機化)	～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	112コア	中	qm	an[01-11]	利用宣言後1年間 (プログラム終了+1年)
一般	ADC (随時)	32コア	低	ql	an[01-11]	最大1年間 (更新可)

キュー構成

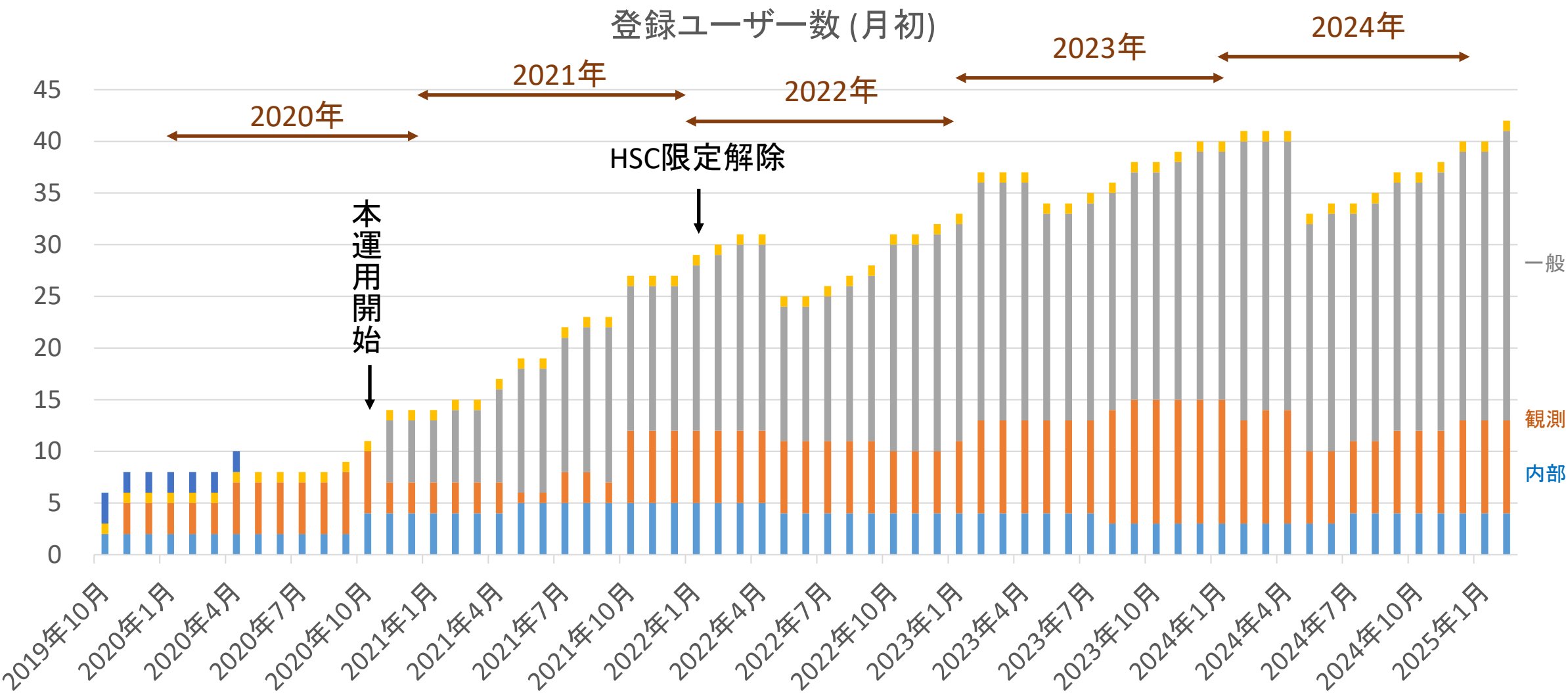
黒太字: ユーザが利用可能なキュー
試験用キューは割愛

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	1,600	64	11,250	11,250	---	---	13-36
qh	高	280	56	4,500	450	---	---	01-11
qm (観測)	中	112	56	1,800	450	---	1	
ql (一般)	低	32	28	450	225	---	1	
qt (テストキュー)	最高	4	4	64	64	1	1	01-36
qhm (要1TB超メモリの解析用)	中	32	8	7,910	1,940	---	1	91-94

同時実行可能ジョブ数(ハード)は、qt以外**現状**無制限

これまでの利用状況

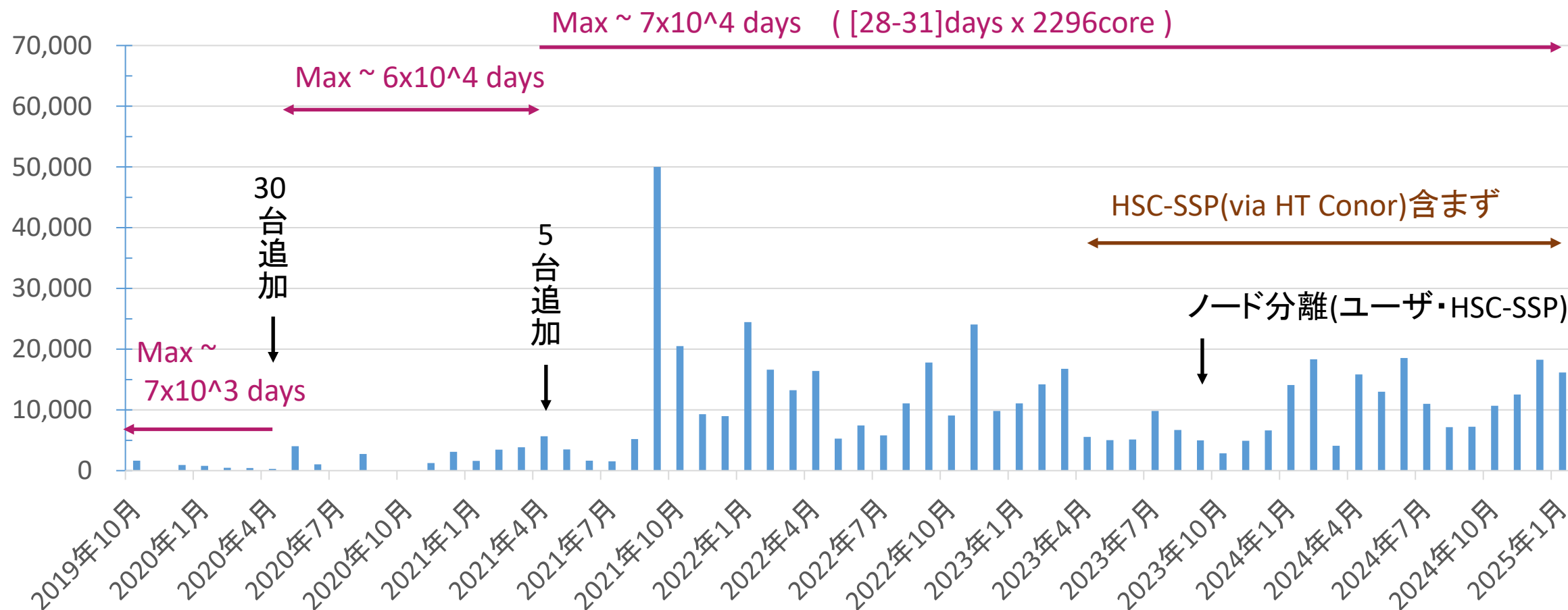
1. 登録ユーザー数推移



これまでの利用状況

2. ジョブCPU時間 [単位:日]

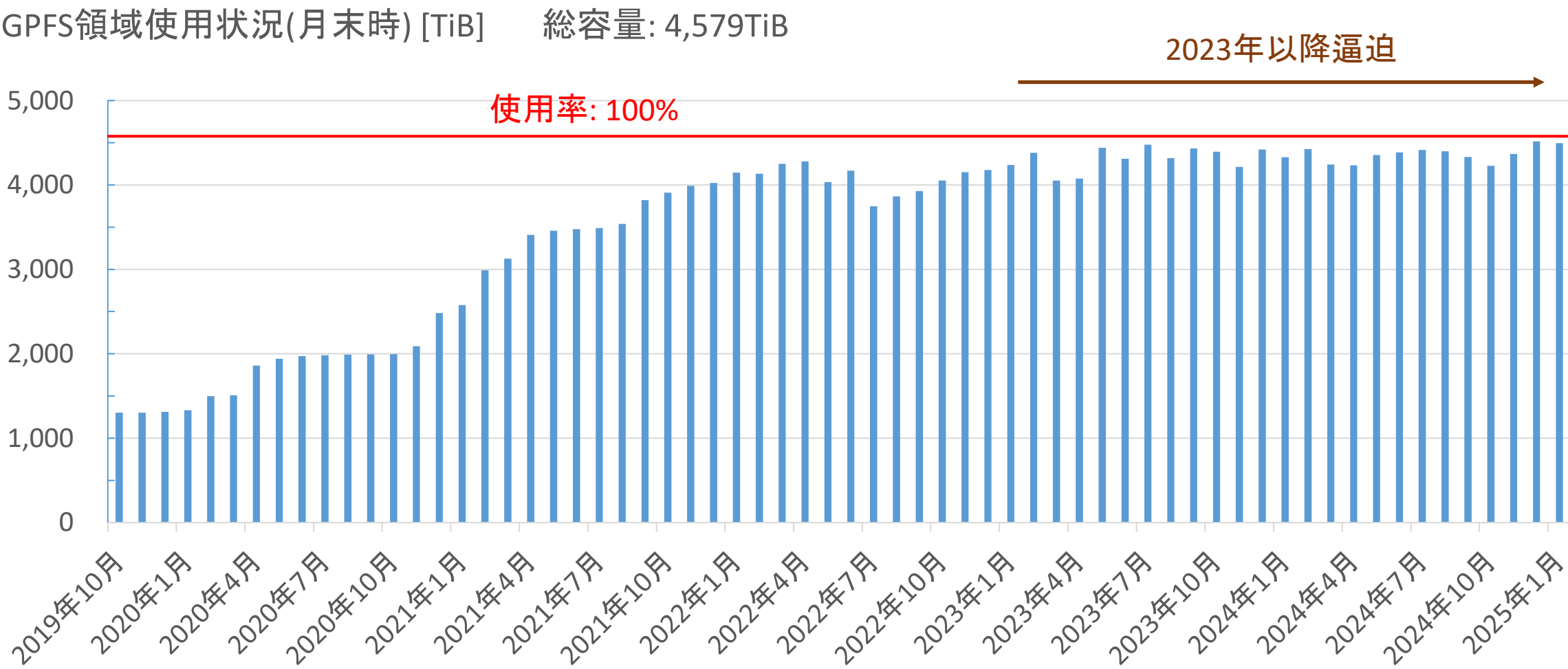
ジョブCPU時間 (all) [days]



補足: CPUに空きがあってもメモリに空きがなければジョブは実行されない。

これまでの利用状況

3. GPFS領域使用量



検討課題

ご意見をお寄せください

- ☆ quotaの導入
- ☆ PFSデータ解析(ハワイ観測所業務)でのLSC利用
- ☆ 実行ノードの分離
- ☆ 同時実行可能ジョブ数の制限(ハードリミット)
- ☆ OS更新
- ☆ 定期メンテの導入

検討課題

ご意見をお寄せください

☆ quotaの導入

GPFS領域はこの2年ほど容量逼迫が続いている。システム・HSC-SSPでこれ以上空きを作るのは難しい。lustreの新領域(2PB、うちユーザ:1PB?)を準備中だが、多くのユーザが利用できるようGPFS領域と新領域にquotaの導入を検討中

- ・ 上限: 30TB/領域
- ・ 導入時期:
 - lustre新領域: 利用開始
 - GPFS領域: 一定の猶予期間(3か月程度を想定)を設けた上で開始
- ・ データ削除: 領域の使用率が閾値(例えば90%)を超えたとき

インテンシブ観測や利用目的上多くの容量が必要な場合は別途相談

補足: GPFS領域は老朽化が進んでいるので、いつまで提供可能かは未知数

検討課題

ご意見をお寄せください

☆ PFSデータ解析（ハワイ観測所業務）でのLSC利用

ハワイ観測所の業務としてのPFSデータ解析でLSCを利用予定

- データ解析専用ユーザ(pfs)・専用キュー(qpfs)の導入
- PFSデータ解析**専用**ストレージの導入
- PFSが利用するCPU資源: ~ 1000 core

詳細は岡本さんの講演で.

検討課題

ご意見をお寄せください

☆ 実行ノードの分離

HSC-SSP用ジョブによって(観測者・一般)ユーザジョブの長期サスペンドが発生するようになったため、HSC業務とユーザで実行ノードを分離。

また、新領域準備などの目的で計算ノード1台を試験機化。

(HSC業務: 1600 → 1536 core, ユーザ: 664 core, 試験機: 64 core)

今後も業務とユーザで実行ノードを分割することを検討中

- ・ HSC+PFS ~ 1,500 core
- ・ ユーザ ~ 700 core
- ・ 試験用 ~ 64core (計算ノード1台分)

検討課題

ご意見をお寄せください

☆ 同時実行可能ジョブ数の制限(ハードリミット)

一般ユーザを中心にユーザが増えてきている。多くのユーザがストレスなく利用できるよう、ソフトリミットに加えてハードリミットの導入を検討中

ハードリミット設定案: ql: 2, qm: 2

(参考: 1ジョブ当たりの最大使用可能CPUコア数: ql:32core, qm:112core)

検討課題

ご意見をお寄せください

☆ OS更新

RHEL/CentOS 7は2024年 6 月末日でEOLとなった。今後は脆弱性が発見されても修正パッケージが提供されないため、OSの更新を計画中

時期: HSC-SSP解析終了後

ダウンタイム: 約 2 週間(システムメンテ)

新OS: Rocky Linux 8

hscPipe: 最新版(8.5.3)のみ導入

OS更新後、MDASソフト導入を検討中。

検討課題

☆ 定期メンテの導入

定期メンテの導入を検討中

頻度: 3か月ごと

開始時期: 未定(OS更新後、MDASと1か月程度ずらす)

ご意見をお寄せください