大規模観測データ解析システム(LSC)

Large-scale Data analysis system (LSC)

現状と検討課題

Current Status and Issues

磯貝瑞希 (国立天文台天文データセンター)

LSC運用チーム

2025年11月18日　　国立天文台天文データセンター　共同利用ユーザーズミーティング

内容 / Contents

☆ **大規模観測データ解析システムの概要と現状/**
Overview and Current Status of the LSC
（システムの特徴、構成、優先権とリソース、キュー構成...）
/ (system feature, configuration, priorities and resources, queue configuration ...)

☆ **前回UM以降の変更点 / Changes since the last UM**

☆ **検討課題 / Issues for Consideration**

大規模観測データ解析システム(LSC)とは / What is LSC?

☆ 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

☆ 計算ノード5台で試験運用を開始。計算資源の制限でユーザを当該セメスターのHSC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拡大 (2022年以降: 「HSCに限定しない」データ解析希望者)

☆ 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

システム情報 / System info.

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	**39**	**CentOS7**	**Intel Xeon/ AMD EPYC**	**1024/512/12 8**	**2,240**
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)

- 作業領域: 5PB (quota: 30TB)

- 新作業領域: 3.5PB

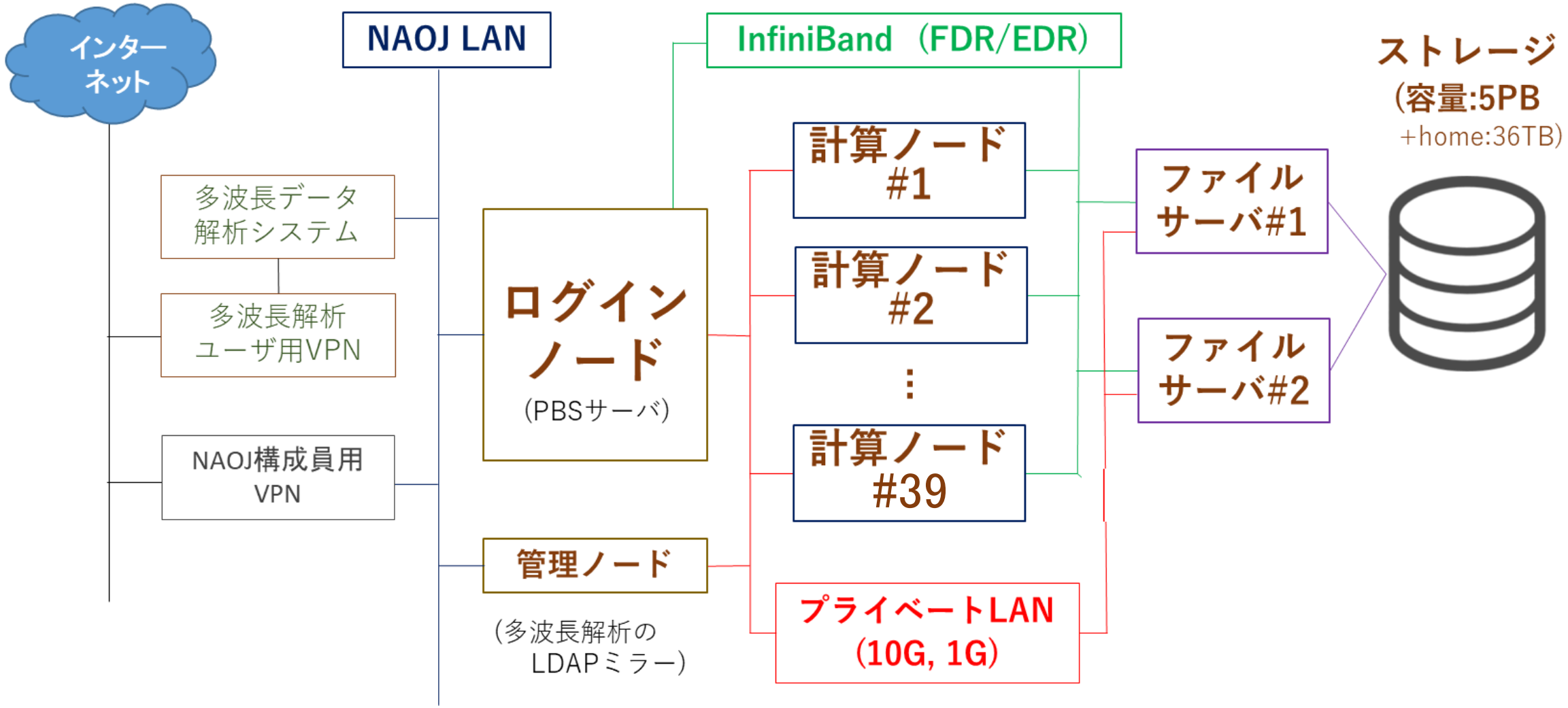
2025年6月に導入

ファイルシステム: XFS+NFSv3

ファイルシステム: IBM Storage Scale (旧GPFS)

ファイルシステム: Lustre 準備中

システム構成図 / System configuration diagram



2021年9月　MDAS共有領域のマウント開始@ログインノード

計算ノードの内訳 / Compute Node Details

an02 2025年4月に故障

node name	number of nodes	OS	CPU per node	memory per node	total number of CPU cores	total memory
an[01,03-05]	4	CentOS7 (7.9)	Intel Xeon Gold 6132 2.6GHz 56core (14Cx4)	1TB	224	4TB
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	**39**				**2,240**	**24.5TB +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

計算ノードの分離 / Separation of compute nodes

現在、HSC業務、PFS業務でもLSCを使用しているが、ユーザへ長期サスペンドなどの影響が出ないようそれぞれの用途でジョブ実行ノードを分離している。

用途 Usage	ノード名 node name	ノード数 Number of nodes	CPUコア数 Number of CPU cores	メモリ量 memory amount
一般・観測ユーザ	**an[01,03-11]**	**10**	**608**	**7TB**
試験	an12	1	64	0.5TB
PFS業務	an[13-20]	8	512	4TB
HSC業務	an[21-36] (SSP: an[13-36])	16 (24)	1024 (1536)	13.5TB (17.5TB)
全員	**an[91-94]**	**4**	**32**	**0.5TB + 8TB(swap)**

利用可能な計算資源・利用期間 /
Available Computing Resources and Duration of Use

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 user type	application contact	available compute resource	priority	queue	compute nodes	(優先)利用可能期間 Priority Usage Period
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[01,03-11]	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[01,03-11]	最大1年間 (更新可)

キュー構成 / queue configuration

黒太字: ユーザが利用可能なキュー

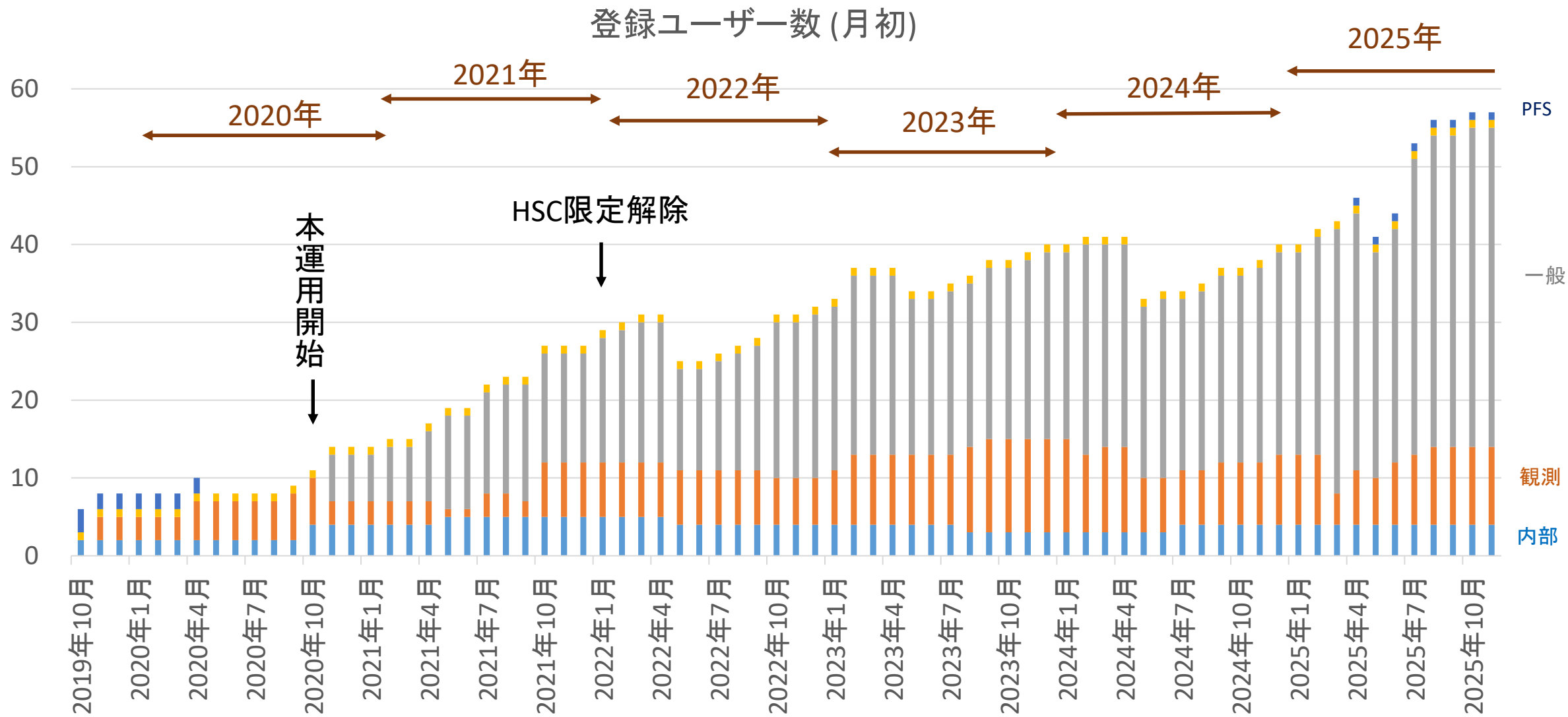
業務用・試験用キューは割愛

キュー名 queue name	優先順位 priority	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 exec. nodes
		最大 max	デフォルト default	最大 max	デフォルト default	ハード hard	ソフト soft	
qm（観測）	**中**	**112**	**56**	**1,800**	**450**	**---**	**1**	**01, 03--11**
ql（一般）	**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

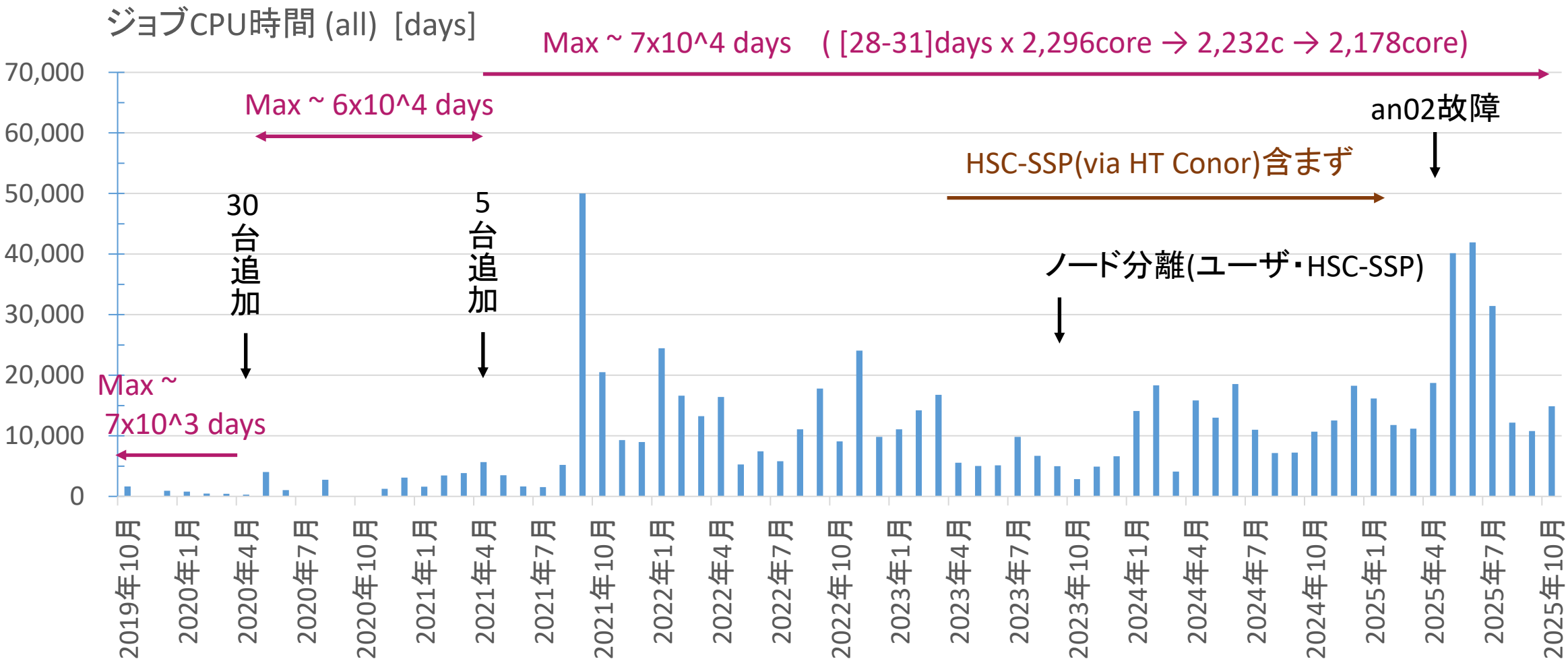
これまでの利用状況 / Past Usage Status

1. 登録ユーザ数推移 / Transition in the Number of Registered Users



これまでの利用状況 / Past Usage Status

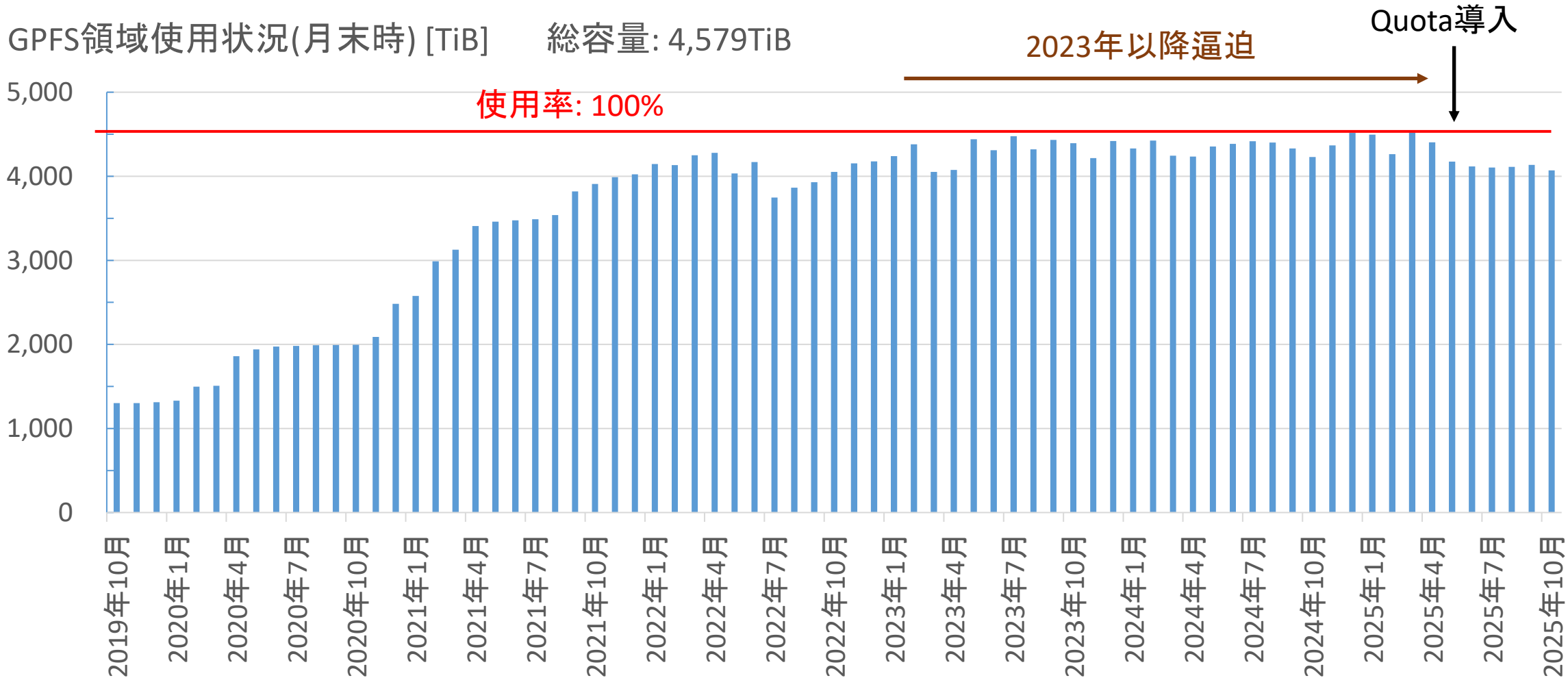
2. ジョブCPU時間　[単位:日] / Job cpu time [unit:days]



補足:「実際に使用された」CPU時間。ジョブは要求CPUコア数とメモリの両方に対して計算資源に空きがなければ実行されない。つまり、CPUに余裕があってもメモリに空きがなければ実行されない。

これまでの利用状況 / Past Usage Status

3. GPFS領域使用量 / GPFS volume usage



前回UM以降の変更点 / Changes since the last UM

☆ **GPFS領域のquota / Implementing quota on the GPFS file system**

→ 2025年6月に導入。基本30TBだが、足りないユーザは応相談。

☆ **定期メンテの再開 / Resumption of regular maintenance**

→ 約3か月に1回のペースで実施中。今後も継続予定だが、実施日は変更予定。

☆ **PFSデータ解析(ハワイ観測所業務)としてのLSC利用開始 /**

Start of LSC utilization for PFS data analysis (Hawaii Observatory work)

→ 計算ノードan[13-20]を使用中。一般・観測ユーザと実行ノードを分離。

PFSのLSC利用は今後も継続予定。

検討課題 / Issues for Consideration

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) /**

Limitation of the number of simultaneously executable jobs

☆ **OS更新 / OS update**

☆ **実行ノード分離の解除 / Release of execution node separation**

検討課題 / Issues for Consideration

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) /**

Limitation of the number of simultaneously executable jobs

一般ユーザを中心にユーザが増えてきている。また一部で大量のジョブを投入する
ユーザが見られ、ジョブスケジューラに負荷がかかっている。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を導入予定。

時期: できるだけ早期に (本UM後 ～ 次回12月メンテ)

ハードリミット設定値: 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)

キュー構成 / queue configuration (ハードリミット導入後)

黒太字: ユーザが利用可能なキュー

業務用・試験用キューは割愛

キュー名 queue name	優先順位 priority	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 exec. nodes
		最大 max	デフォルト default	最大 max	デフォルト default	ハード hard	ソフト soft	
qm（観測）	**中**	**112**	**56**	**1,800**	**450**	**2**	**1**	**01, 03--11**
ql（一般）	**低**	**32**	**28**	**450**	**225**	**2**	**1**	
qt（テストキュー）	**最高**	**4**	**4**	**64**	**64**	**1**	**1**	
qhm（要1TB超メモリの解析用）	**中**	**32**	**8**	**7,910**	**1,940**	**---**	**1**	**91-94**

検討課題 / Issues for Consideration

ご意見をお寄せください

☆ **OS更新 / OS update**

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、早急な更新が必要である。
更新に伴うシステム運用停止期間短縮のため、ログインノードの代替機を注文。
以下は現時点での想定:

時期: 2026年2〜3月（ログインノード代替機の納品・構築完了後）
運用停止期間: 約1週間 (システムメンテ)
新OS: Rocky Linux 8 が第一候補 (MDASとの親和性 + OpenPBSの構築)
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 / Issues for Consideration

ご意見をお寄せください

☆ **実行ノードの分離解除 / Release of execution node separation**

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していたが
SSP解析業務がひと段落ついたため、実行ノードの分離の解除を検討中。