

国立天文台天文データセンター 新多波長解析システム の構築と性能評価

磯貝 瑞希、
共同利用運用G(ADC)

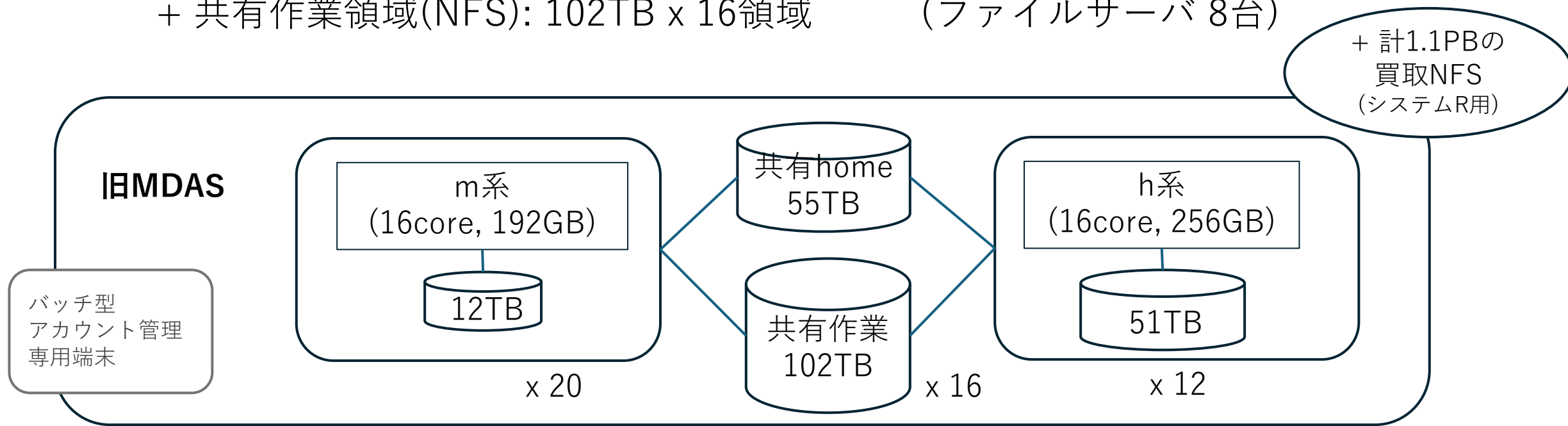


多波長解析システムとは

- 天文データセンターが運用する共同利用計算機システム
- 国内外の天文および関連分野の研究者へデータ解析環境を提供
- ユーザ ~ 340名
- 導入ソフトウェア: CASA, IRAF/PyRAFなどの解析ソフトウェア、言語(IDL, Python...)、コンパイラ(intel oneAPI, gcc...)...など多数
- 2024年7月にシステムリブレース。レンタルからリース、システム構築および運用の計算機作業も天文台に。

旧システム

- システム構成: (対話型:32台, バッチ型:2台)
 - 解析サーバ(16core, m系:192GB/h系:256GB) 計32+2台 (m:20+2台、h:12台)
 - ストレージ:
 - + ローカル作業領域: 12TB/対話型m, 51TB/h
 - + 共有home領域: 55TB x 2領域 (ファイルサーバ専用機)
 - + 共有作業領域(NFS): 102TB x 16領域 (ファイルサーバ 8台)



旧システムの課題

- 共有作業領域が細切れで使いにくい(~100TB x 16領域) & 頻繁に逼迫
- 共有作業領域のファイル読み書き(IO)性能が高くない(NFS、ローカルより低)
- 計算資源が細切れ(台数は多いが、1台の計算資源:小)
- ローカル作業領域は共有作業領域よりもIO性能が高いがサーバ固有で使いにくい

新システムの方針:

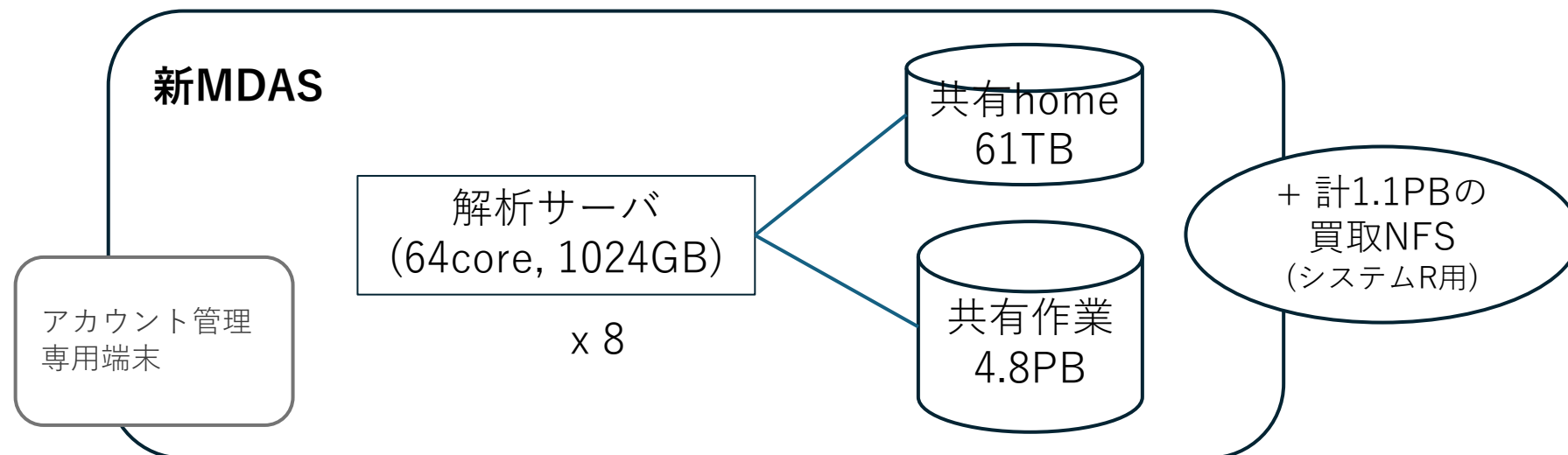
- システムの計算資源はほぼ同等。作業領域の容量は増加
- 解析サーバの台数を削減し、1台あたりの計算資源を増やす
- ローカル作業領域の廃止、バッチ型解析サーバの廃止(→ LSC)
- 大容量かつ高速な共有作業領域の採用 → Lustre FSの採用
- 高価な専用機ではなく、汎用機+ソフトウェアによる高可用性の実現

新システム

- システム構成:
 - 解析サーバ(64core, 1024GB) 計8台
 - ストレージ(ファイルサーバ6台):
 - + 共有home領域(NFS): 61TB
 - + 共有作業領域(lustre): 4.8PB

資源の新旧比較:

	旧	→	新
計算資源:			
CPU:	512+32core	→	512core
Memory:	~7TB	→	~8TB
home+			
作業領域:	2.6PB	→	4.9PB



Lustreファイルシステム(FS)とは

複数のサーバで一つのファイルシステムを構成する「クラスターFS」の一つ。
OSSでパフォーマンスに優れ、HPC分野で多くの採用実績がある。

Lustre FSの構成:

- MGS マネージメントサーバ Lustre FSの管理サーバ。システムに一つ。
MDSと兼用可
- MDS メタデータサーバ メタデータ格納領域(MDT)を管理するサーバ
1領域=1MDT。冗長構成可
- OSS オブジェクトストレージサーバ オブジェクト格納領域(OST)を管理するサーバ。
1サーバで複数のOSTを管理可。冗長構成可

Lustre FS@新MDAS: 1 x MGS兼MDS(1xMDT) + 4 x OSS(6xOST)

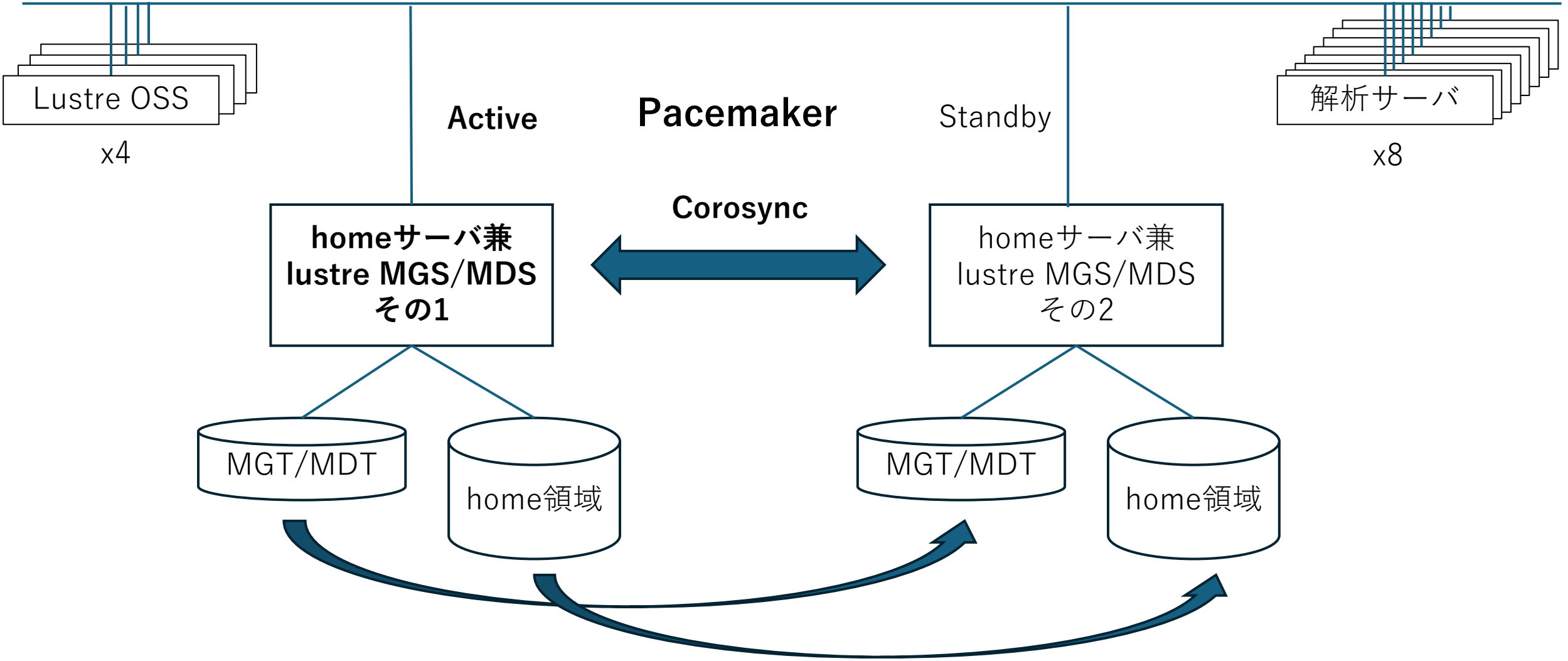
高可用性の実現

- ・ 共有home領域: 障害が起きると全解析サーバが使用不可
- ・ Lustre MGS兼MDS/MDT: 障害が起きるとLustre領域が使用不可

耐障害性を高めるため、DRBD+Pacemakerで冗長構成化:

- DRBD(=Distributed Replicated Block Device) ストレージレプリケーションソフトウェア。サーバ間でブロックデバイスの内容をミラー
- Pacemaker クラスターリソースマネージャ。 クラスタ制御ソフトウェア
Corosyncと組み合わせて、Active-Standby構成のクラスタを構築可

SAN(PNW)



DRBDによるレプリケーション

性能評価試験

旧システムと新システムで2種類の試験を実施。

- **画像処理:** 天文データ解析ソフトウェアPyRAF/IRAFを使用
1chip 2048 x 4096 pixの一次処理(flat:16枚、object:32枚) x 64chip分
→ 主にファイル読み書き性能を評価
- **天体検出:** 天体検出ソフトウェアSExtractorを使用
33416 x 30720 pix (1pix=4byte) のFITS画像データからの天体検出 x 64データ
→ 主に演算性能を評価

試験実施前に、ファイルサーバおよび実行サーバでキャッシュクリアを実施

性能評価試験結果 (処理時間[s])

画像処理 (PyRAF)

旧システム 16core x 4hosts		新システム core x host(s)			
lfs07	lfs14	64 x 1	32 x 2	16 x 4	8 x 8
998	1006	89	54	36	28

旧→新で速度は
10倍超に向上

天体検出 (SExtractor)

旧システム 16core x 4hosts		新システム core x host(s)			
lfs02	lfs09	64 x 1	32 x 2	16 x 4	8 x 8
322	340	185	117	95	93

旧→新で速度は
1.7倍以上に向上

旧→新 ファイル読み書き性能の向上が顕著

まとめ

- 多波長解析システムはデータ解析環境を提供する共同利用計算機システム
- 2024年7月のシステムリプレイスで新システムに移行。天文台が構築
- 計算資源は旧システムとほぼ同等。台数の削減(1/4)で計算資源の細切れを解消
- 共有作業領域はLustreの採用により旧システムの課題(低IO性能、領域の細切れ)を解消
- DRBD+Pacemaker/Corosyncでhome領域とlustre MGS/MDSの高可用性を実現
- 性能評価試験では旧システムの10倍超(画像処理)または1.7倍以上(天体検出)と特にファイルの読み書きで大幅な性能向上を実現