ALMAの大規模データ処理

国立天文台ALMA推進室
小杉　城治

ALMAの大規模データ処理

(Atacama Large Millimeter/Submillimeter Array)

- ALMAの概要と現状
- データアーカイブ
- データ処理
- データ解析の高速化

ALMAとは
(Atacama Large Millimeter/Submillimeter Array)

- 日・米・欧共同でチリ・アタカマ砂漠に建設中の巨大電波望遠鏡(ミリ波・サブミリ波干渉計)
- 今年3月に最初のCall for Proposal
  - 初期科学運用観測開始
  - 初期運用時の望遠鏡16台でも、感度において、サブミリ波で世界最高性能
- 2012年度に本格運用開始予定
  - 望遠鏡66台
  - 運用期間は30年程度を予定

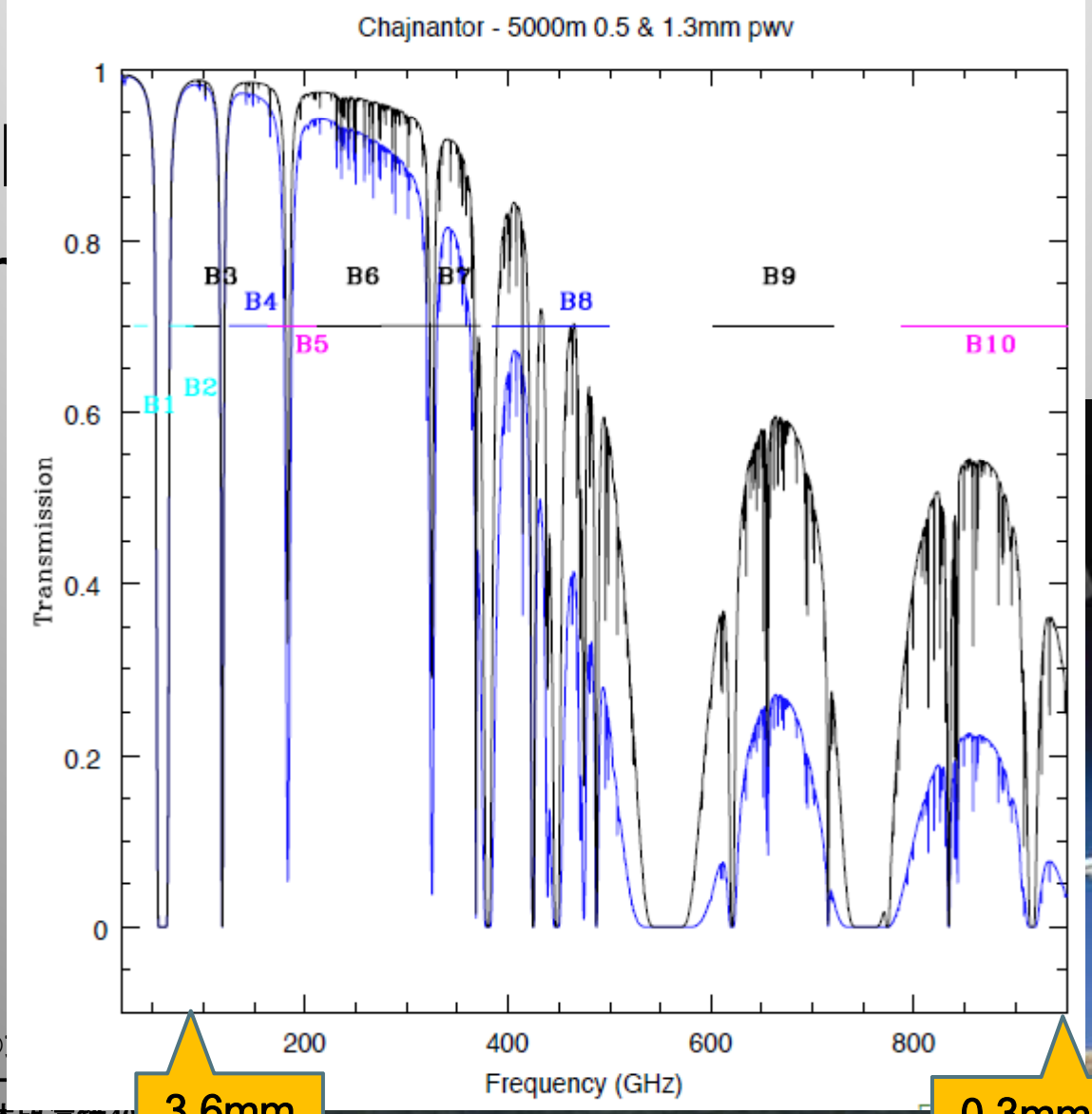
ALMAの主な仕様

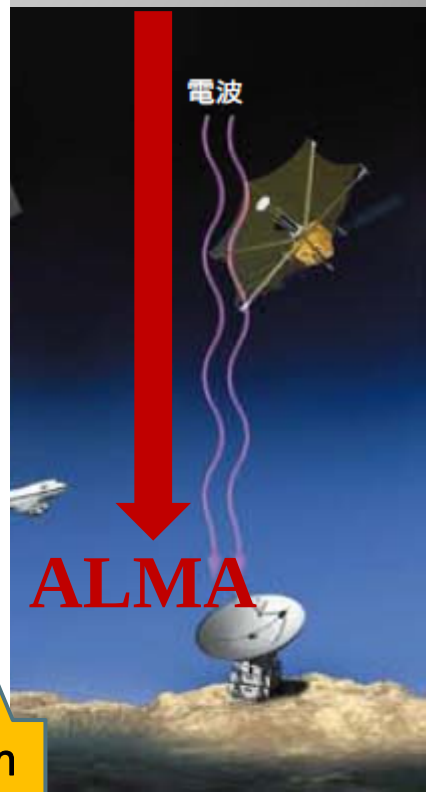
- 望遠鏡（計66台）
  - メイン干渉計用 12m x 50台
  - コンパクト干渉計用 7m x 12台 (ACA)
  - 単一鏡観測用 12m x 4台 (ACA)
- 干渉計最大基線長　0.15–16 km
- 角度分解能　0”.5–0”.005 (@900GHz)
- 視野　7” (@900GHz)
- 観測波長域　0.3–3.6mm (80–950GHz)

観測波長域

◆80G

◆0.3m





的情報

F:山麓施設(標高 2,900 m)

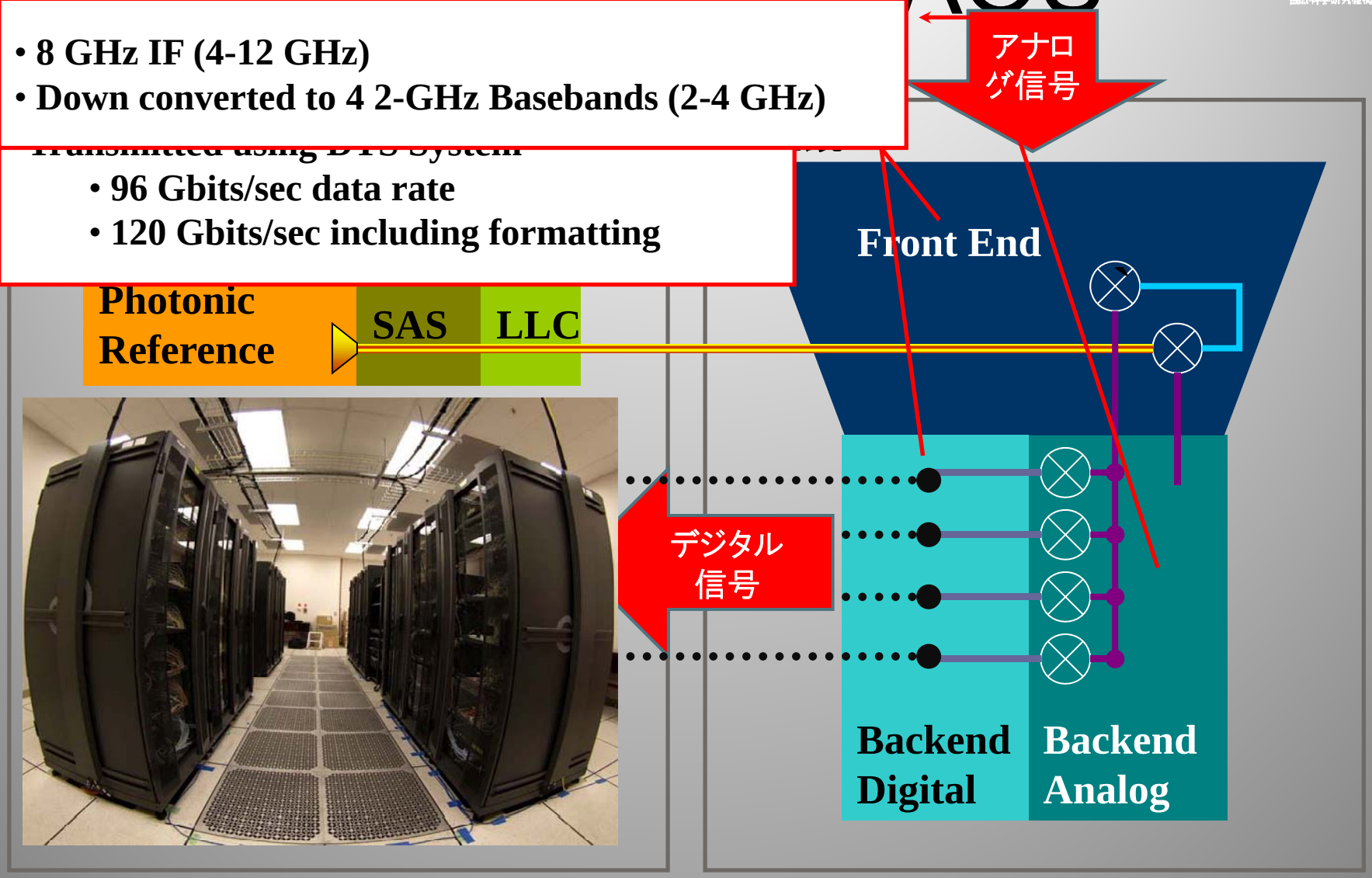
S:山頂施設(標高 5,000 m)

Signal Path at AOS

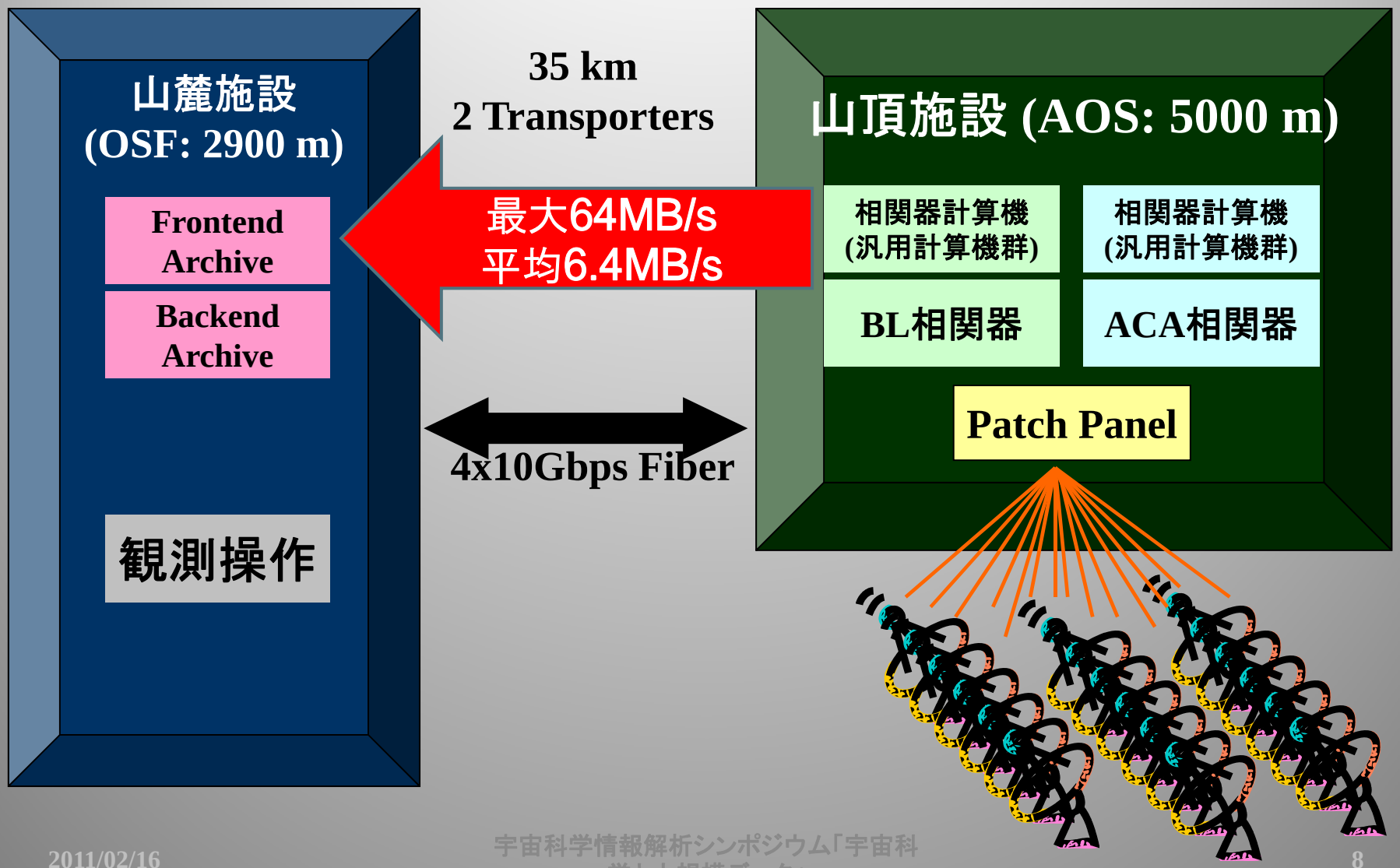
- 8 GHz IF (4-12 GHz)
- Down converted to 4 2-GHz Basebands (2-4 GHz)

Transmitted using DTS System

- 96 Gbits/sec data rate
- 120 Gbits/sec including formatting



Signal Path: AOS - OSF



Data Flow: ~SCO

(Santiago Central Office)

山頂施設
(AOS: 5000m)

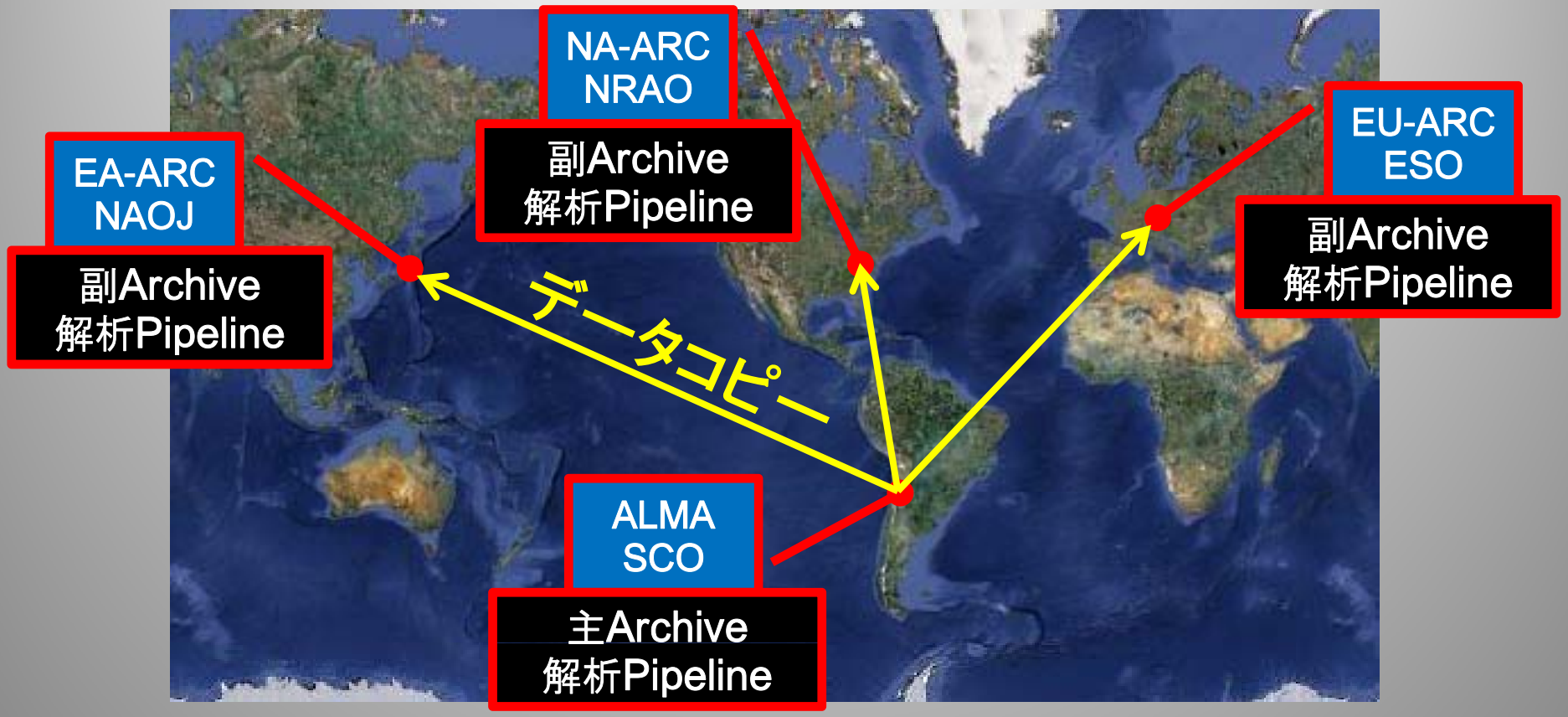
最大64 MB/s

山麓施設
(OSF: 2900m)

平均6.4 MB/s

サンチャゴ施設
(SCO)

チリALMA観測所とARC
(ALMA Regional Center)



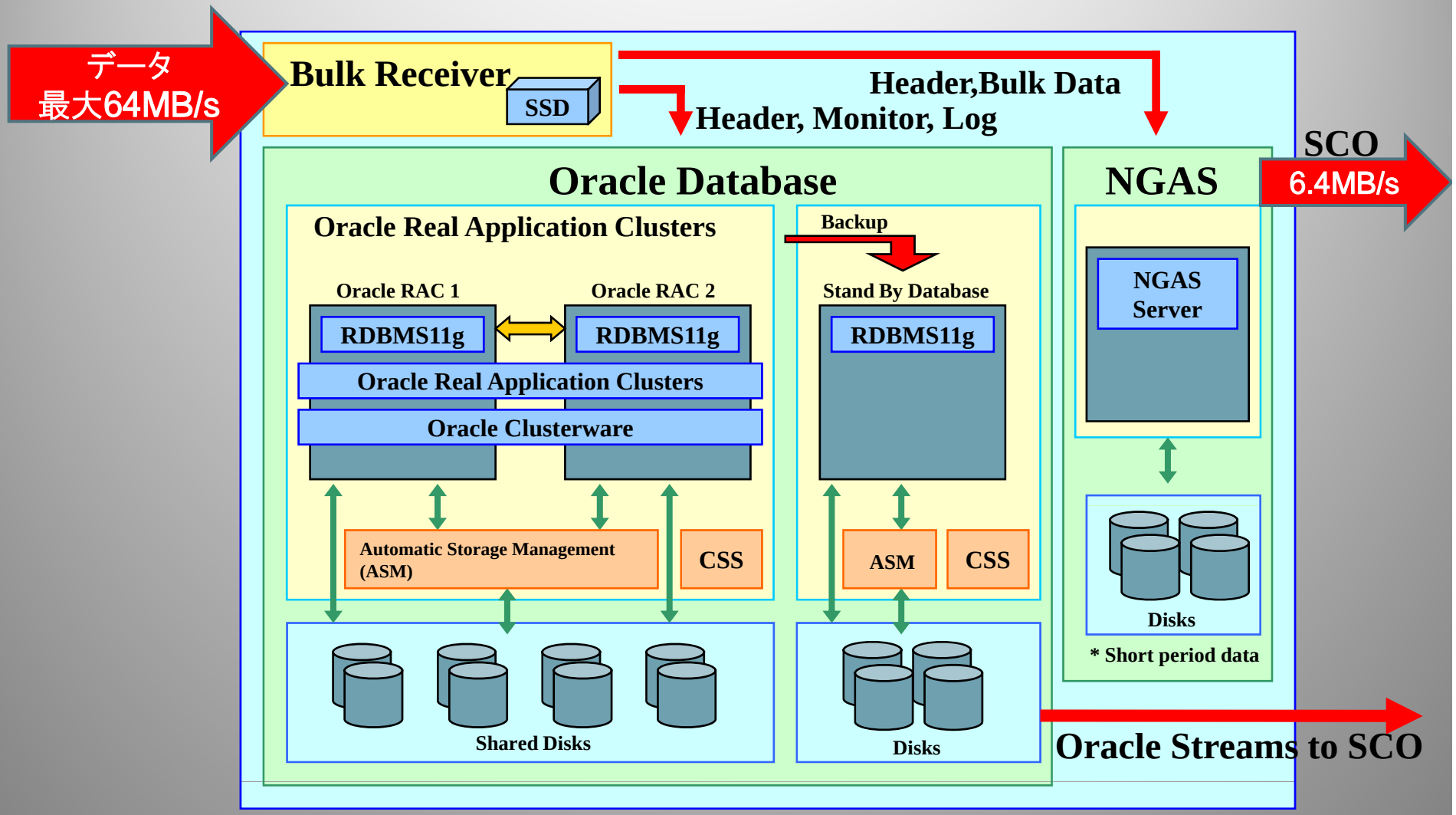
観測者は基本的に解析パイプライン済みデータを
各ARCから受け取る

ALMAのアーカイブシステム

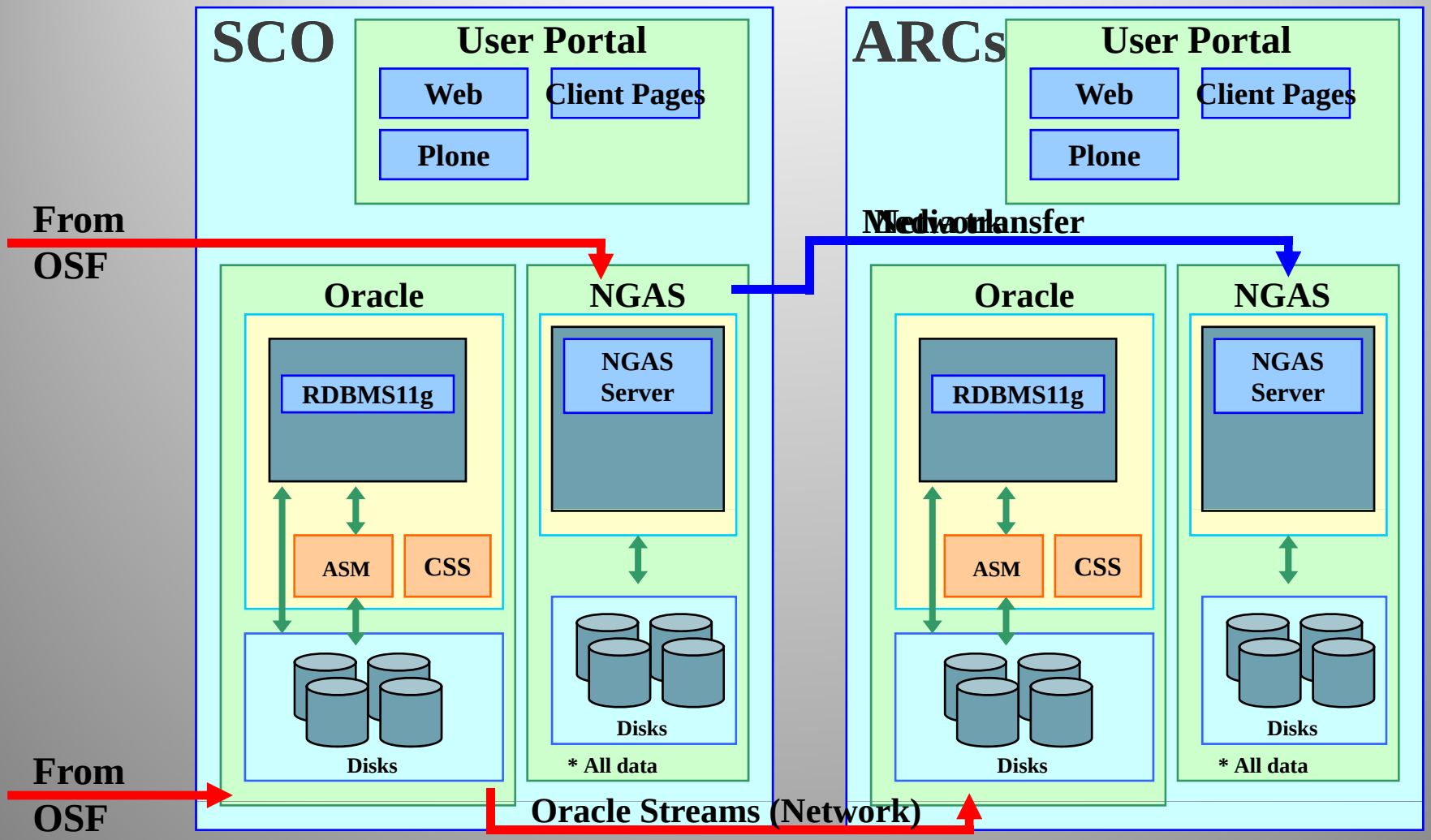
- Frontend Archive in OSF
  - 最大64MB/sの観測データを取りこぼしなく
  - 観測ログや機器のステータス情報もDB化
- Backend Archive in OSF
  - 平均6.4MB/sでSCO Science Archiveへ転送
- Primary Science Archive in SCO
  - データ量:200TB/year @本格運用時
  - 各ARCのScience ArchiveへReplication
- Secondary Science Archives in ARCs
  - データ量: 200TB/year

注:Science ArchiveとOSF Archiveの違いは、一般ユーザーがアクセスできるかどうか。

Frontend/Backend Archive



SCO/ARC Science Archive



ALMAのデータ処理

- ALMAのデータ
  - 年間200TB
  - 典型的な1つのデータセットサイズ 25 – 250 GB

  (データセットが解析の処理単位)
- ALMAの主要なデータ処理
  - 相関処理(相関器、相関器計算機)
  - パイプライン処理
  - インタラクティブ処理(CASA)

相関処理

- 相関器（リアルタイム専用計算機）

- 相関器計算機（準リアルタイム汎用計算機）

ACA相関器のデータ処理（1/2）

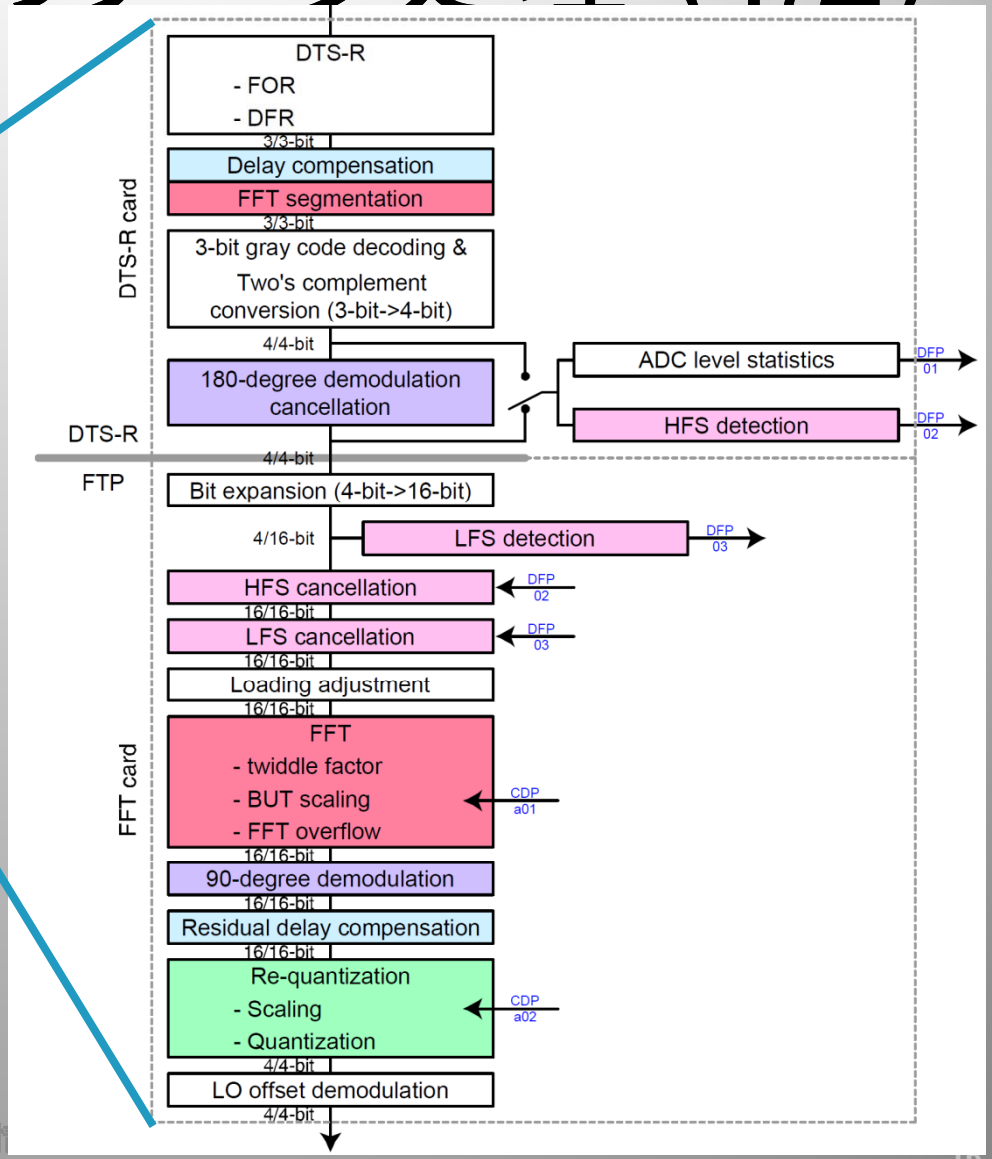
デジタル
信号
12Gbits/s/望遠鏡 × 12

ACA相関器

相関データ
2.4GB/s

相関器計算機
64MB/s/計算機

補正済み
データ
3.6MB/s



ACA相関器のデータ処理（2/2）

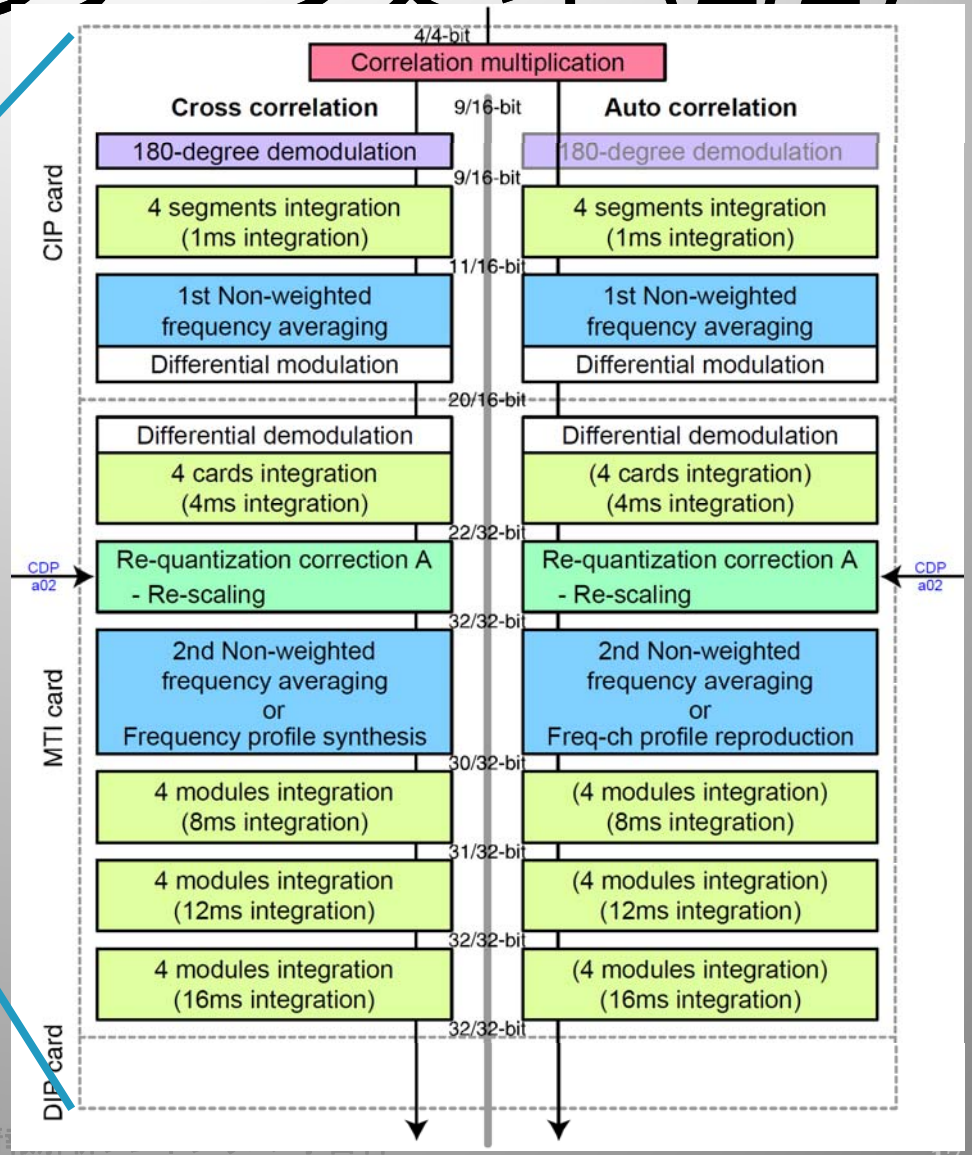
デジタル
信号
12Gbits/s/望遠鏡 × 12

ACA相関器

相関データ
2.4GB/s

相関器計算機
64MB/s/計算機

補正済み
データ
3.6MB/s



相関器計算機のデータ処理

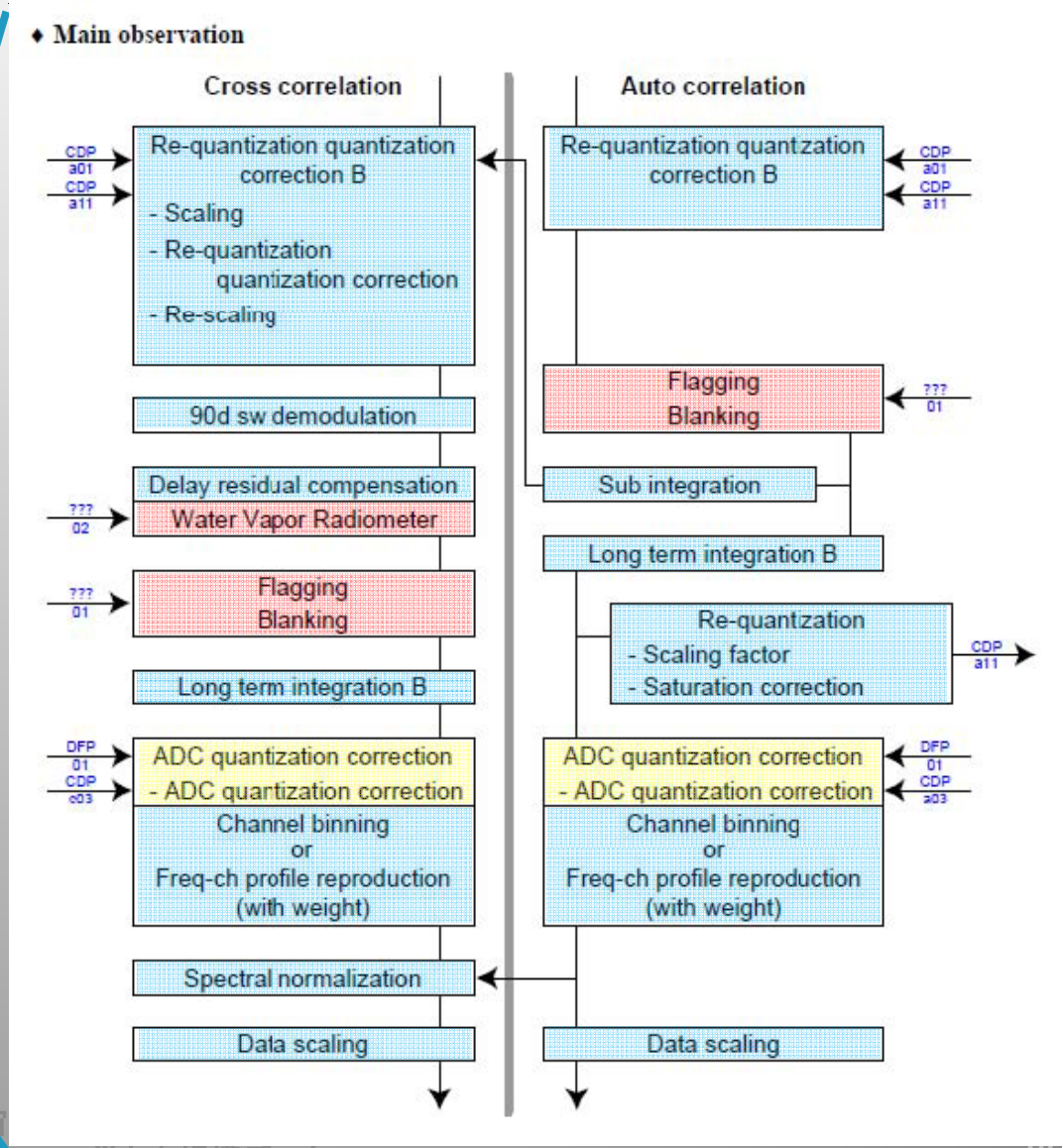
デジタル
信号
12Gbits/s/望遠鏡 × 12

ACA相関器

相関データ
2.4GB/s

相関器計算機
64MB/s/計算機

補正済み
データ
3.6MB/s



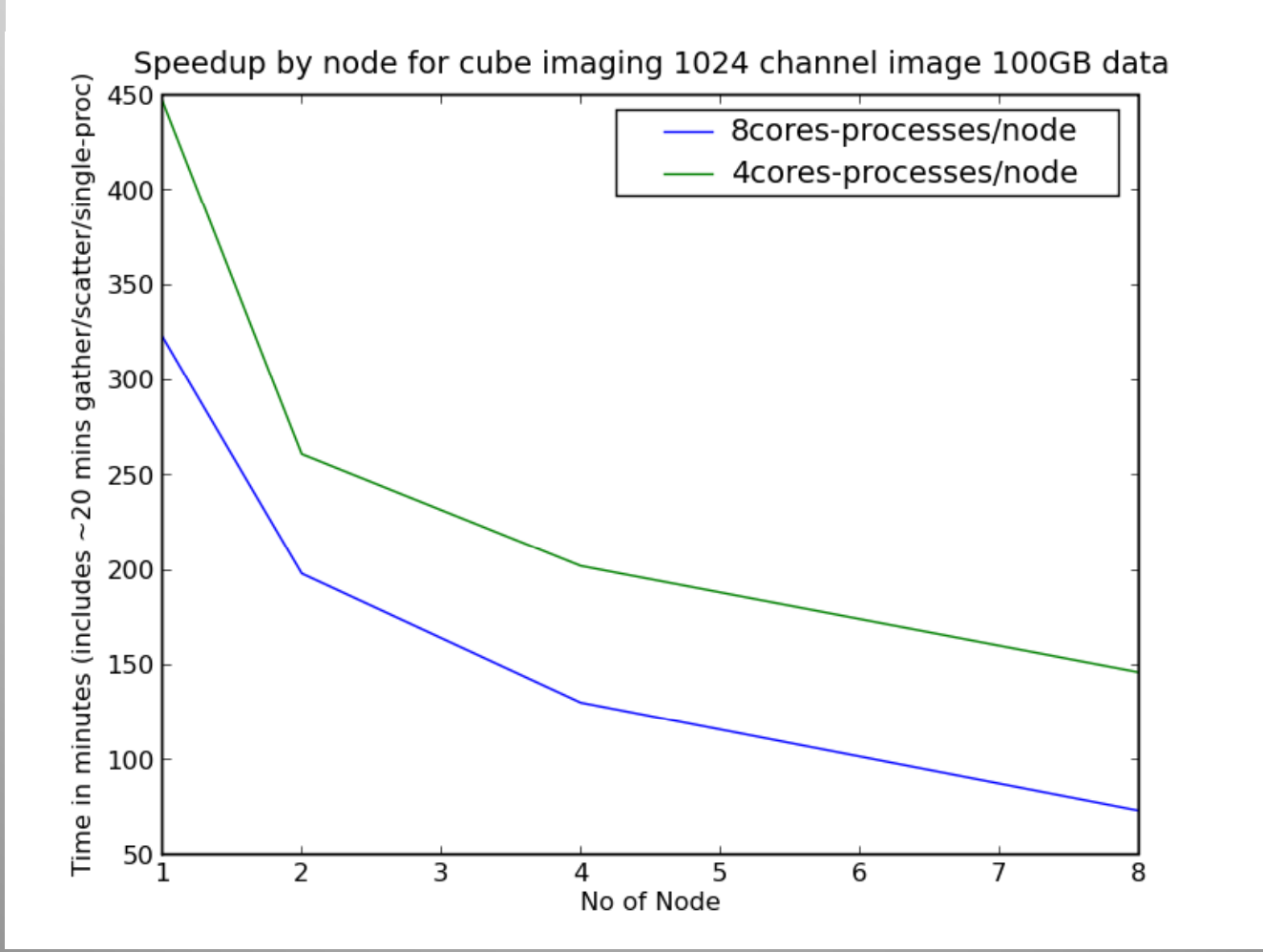
相関処理

- 相関器(リアルタイム専用計算機)
  - 高度な高速化が要求される部分
  - 処理の詳細までわかっている部分
- 相関器計算機(準リアルタイム汎用計算機)
  - データフォーマット変更など、手間がかかるがさほど速度を要求されない部分
  - 処理詳細が定まっていない(運用を通じて詳細化されるキャリブレーションなど)部分
  - 望遠鏡などの制御ソフトと依存関係の深い部分(運用で実装が変わりうる)

データ解析ソフトウェア：CASA
(Common Astronomy Software Applications)

- 電波観測データ用解析ソフトウェア
- AIPS++から進化
- C++ Core LibraryをPythonでラップ
- 日米欧、15人体制で開発中
- ALMA PipelineはCASAの機能を解析エンジンとして利用
- プロセスの並列化(Python Level)の実装中
- データへの同時（並列）アクセス(C++ Level, 並列アクセス可能なデータ形式)の実装中
  - 天文観測データは概してI/Oネック

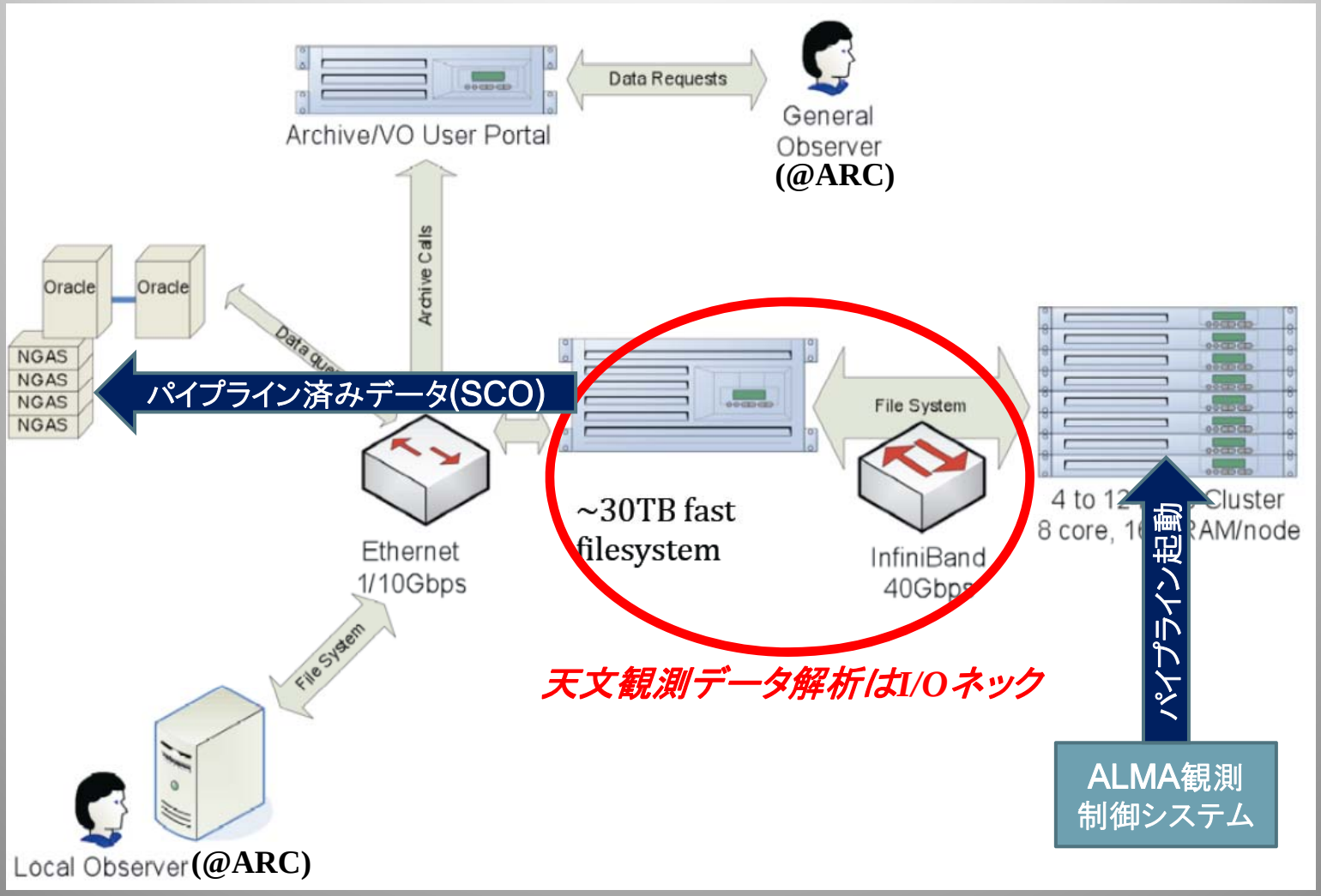
CASAプロセス並列化
データアクセス並列化（実測例）



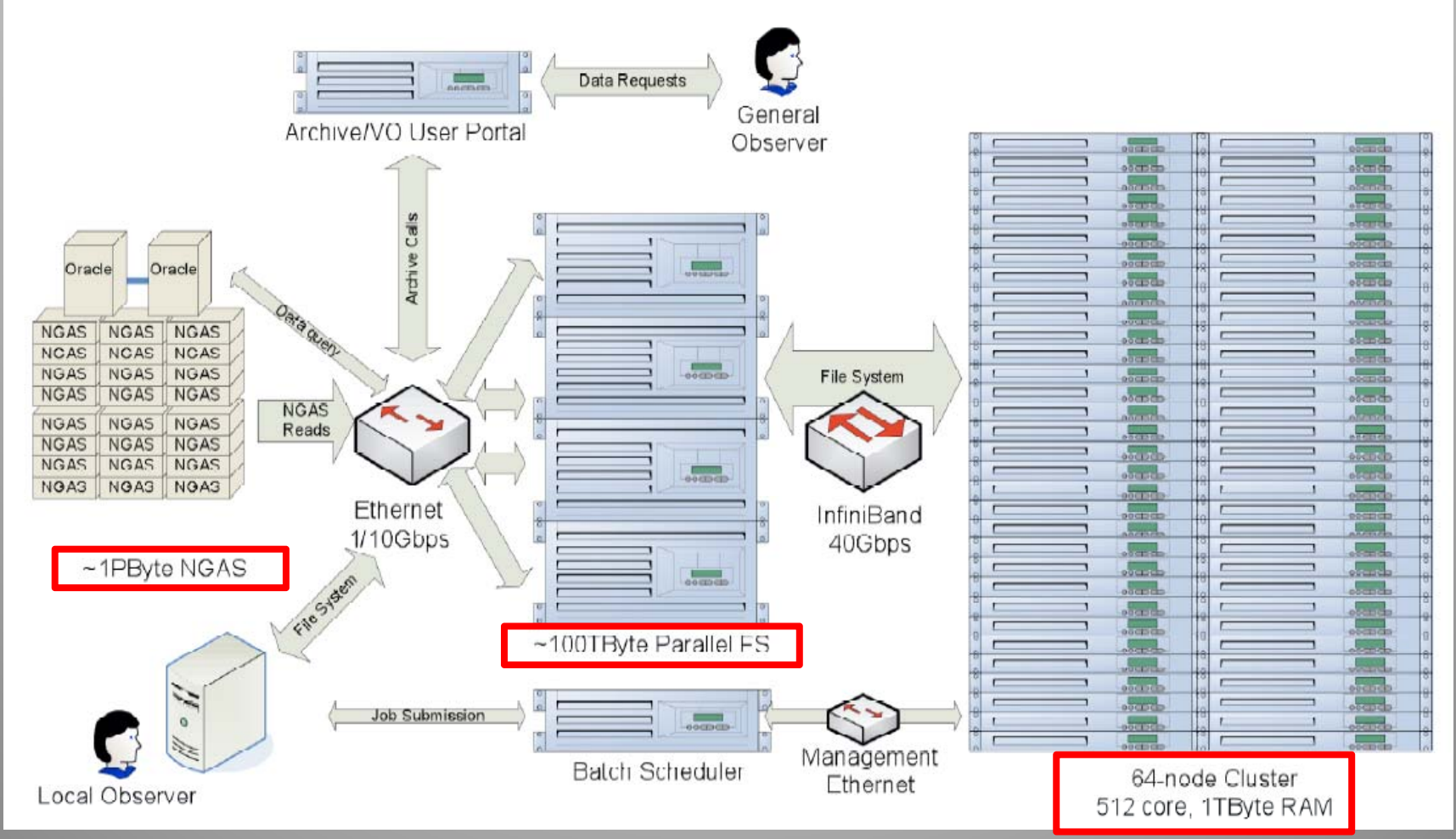
パイプライン処理

- SCO Data-Driven Pipeline
  - データセット(ターゲットデータと関連するキャリブレーション)が揃った時点で動作開始
  - 平均データレート(6.4MB/s)以上の高速処理
- ARC User-Driven Pipeline
  - ユーザーが再解析(例:解析パラメータを変更)のためにインタラクティブに起動
  - バーストパフォーマンスが必要(ユーザーは性急)

初期パイプラインクラスター（計画）



本格運用時のパイプラインクラスター（計画）

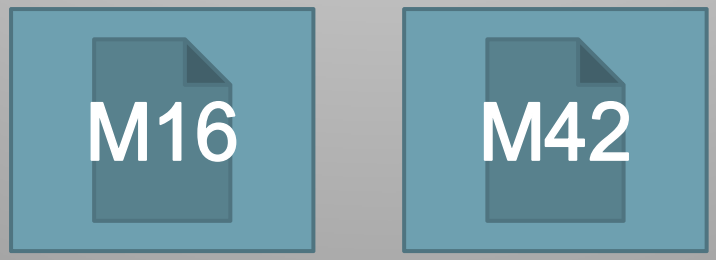


データ解析高速化のためにALMAは

- 解析プロセスの並列化
- データへの並列アクセスが可能なハードディスク上のデータフォーマット
- 高速ネットワーク、高速ファイルシステムを使用

大規模データ処理の可視化（≠データの可視化）

- ユーザーはパイプラインの処理内容に問題がなかったどうか判断したい

そこで、

- ALMAパイプラインは解析ログを解析済みデータとともにアーカイブし、ユーザーに提供する