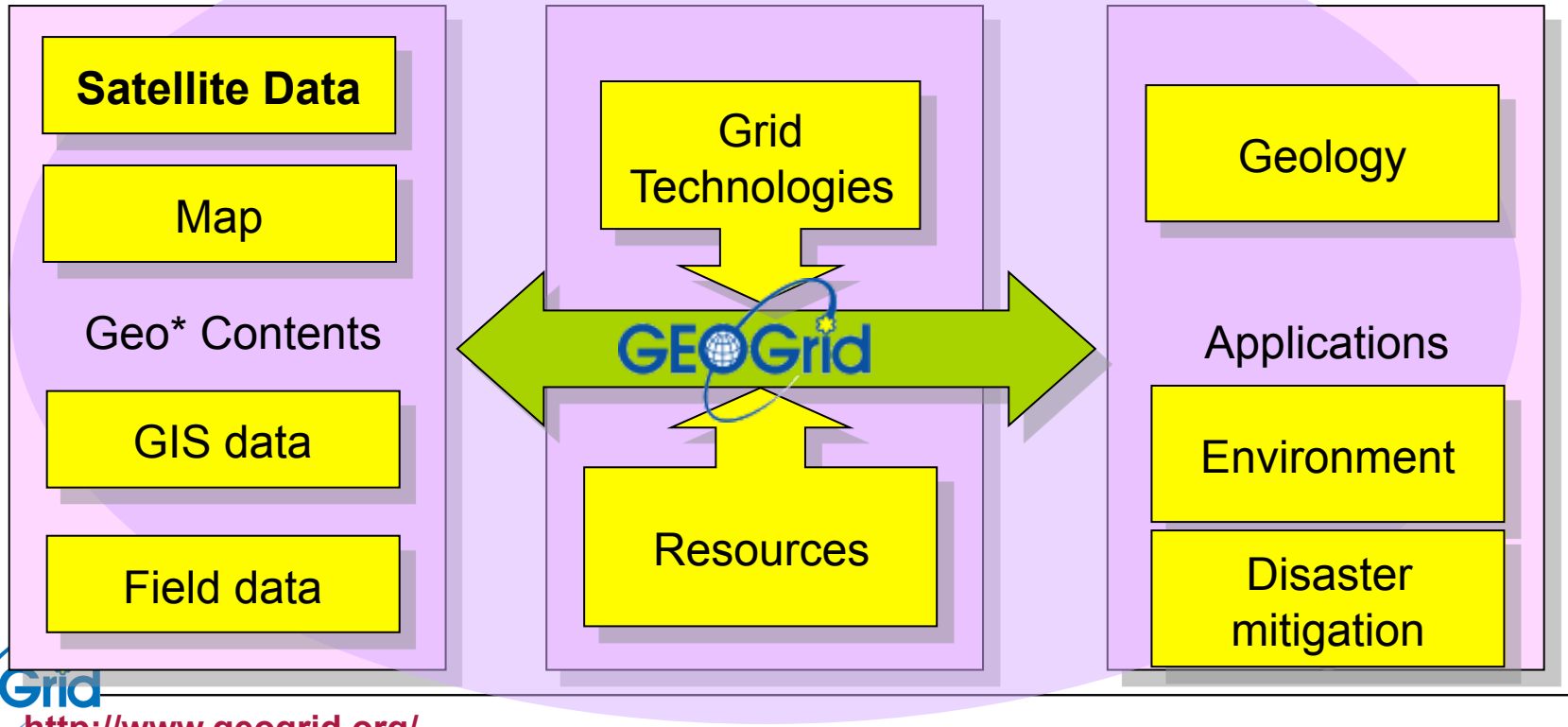
地球観測グリッド(GEO Grid)について

山本 直孝
産業技術総合研究所
情報技術研究部門
サービスウェア研究グループ

H22年度宇宙科学情報解析シンポジウム「宇宙科学と大規模データ」
H23年2月16日 @ISAS

GEO (Global Earth Observation) Grid

- 2005年にスタートした産総研のプロジェクト
- グリッド技術を用いて様々な地球観測データを統合し、環境、防災など様々な応用分野に適用しようという、デザインかつコンセプト
- グリッドのキラーアプリとして、また、産総研の分野融合課題として、積極的に推進してきた。

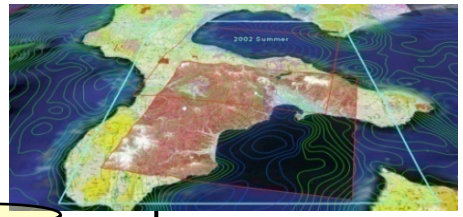


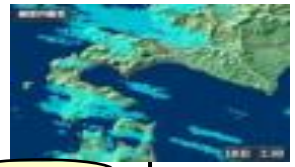
シナリオ例： ハザードマップ作成

Visualize Hi-Resolution Hazard Map

GIS Data

Rain Sensors

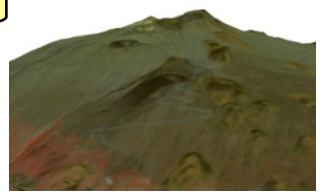


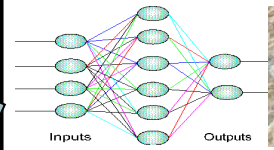


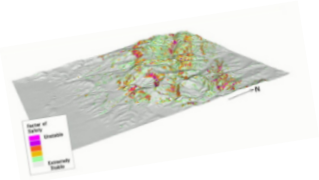
GEO

MET

ASTER







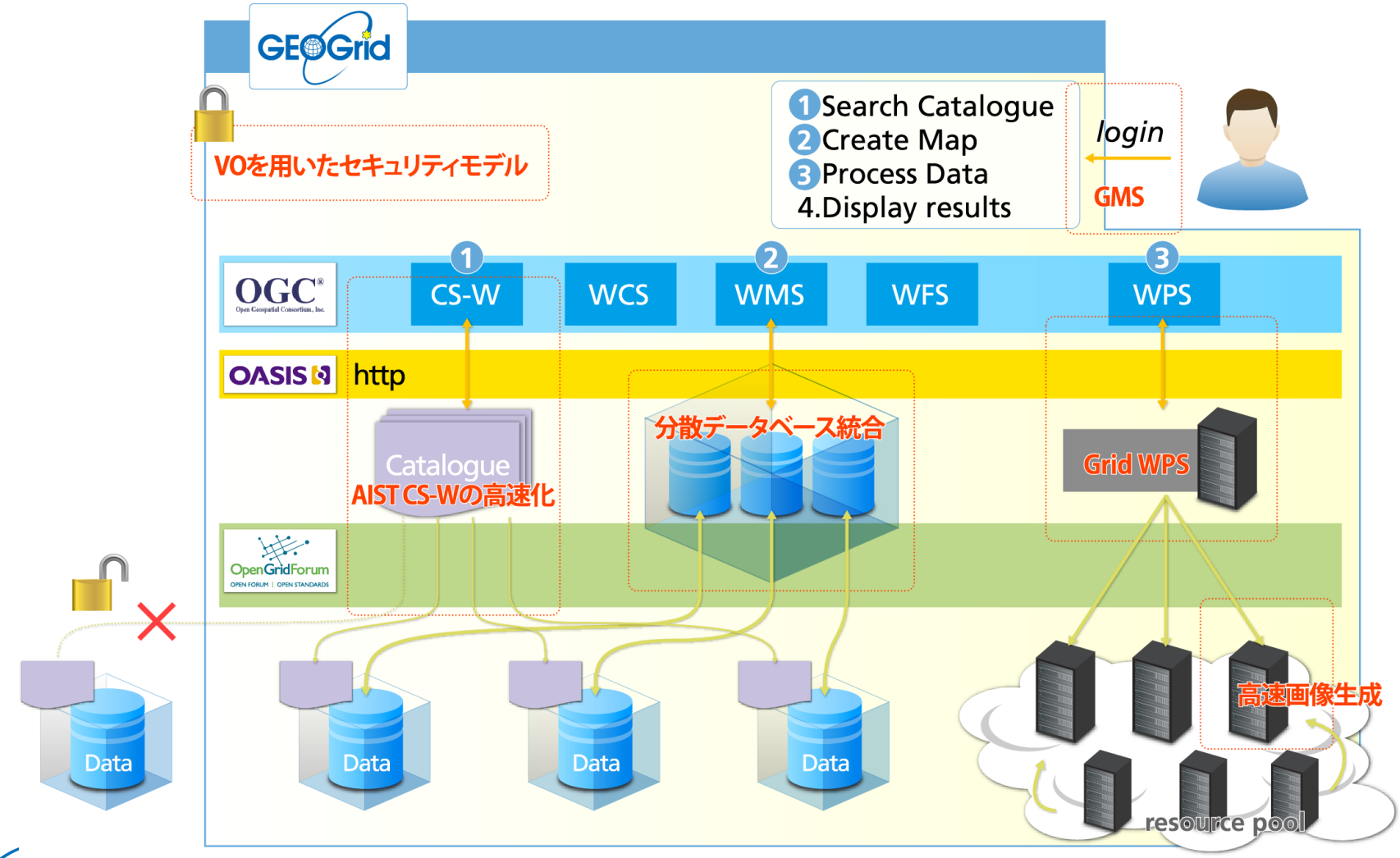
Digital Elevation Map
Created by
ASTER Satellite Images

Neural-Map based Data
Mining / Landslide Simulation
on Large Cluster Computers

要求機能

- 検索
  - 複数データベースへの共通化されたアクセス
- データ転送
  - ユーザがデータをダウンロードする必要の無い第3者転送
- アプリケーション実行
  - 様々なプラットフォームでのアプリケーション実行
- セキュリティ
  - 単一認証（シングルサインオン）
  - アクセスコントロール
  - 権限委譲 (delegation)

GEO GridのGridたる所以



DBからみた特徴

- ファイル　on Gfarm 約800TB、約180万シーンのASTERが中心
  - DB＝PostGIS&独自のカタログ実装(メタデータおよび構造化データの格納)
  - 全量の同時処理・解析＝あることはあるが、あまりない
    - 固定したワークフローが多く、あまり改造・変更がない
    - パラメータなどの変更による再処理。
      - 再処理化のパラメータ等のメタデータの生成・管理・次の検索に対する利用が重要
  - 検索した結果の解析、興味のある対象を選択しての処理＝とても多い
    - データを如何に簡単に探せるかが鍵

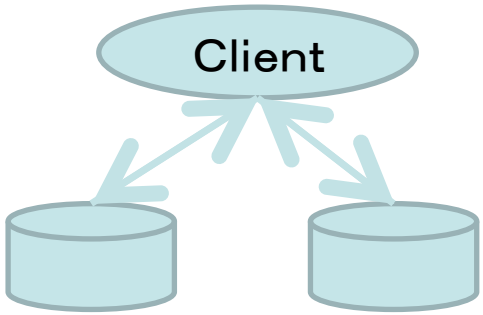
目次

- (異種)分散データベース統合
- メタデータ統合
- データ統合
- データインテンシブなのか？
- クラウドへ

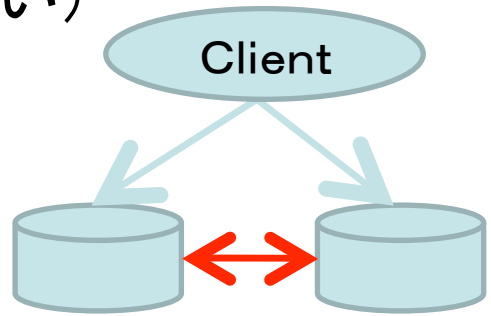
1）（異種）分散データベース統合

DB統合/連携の要件

- 異種のデータが同じフレームワークで扱えること
  - XML,RDB,RDF,web etc
- サイト「間」処理によるデータ統合
  - 同じ問合せを全サイトへ（簡単）



  - 分散結合（最適化が入るのでややこしい）

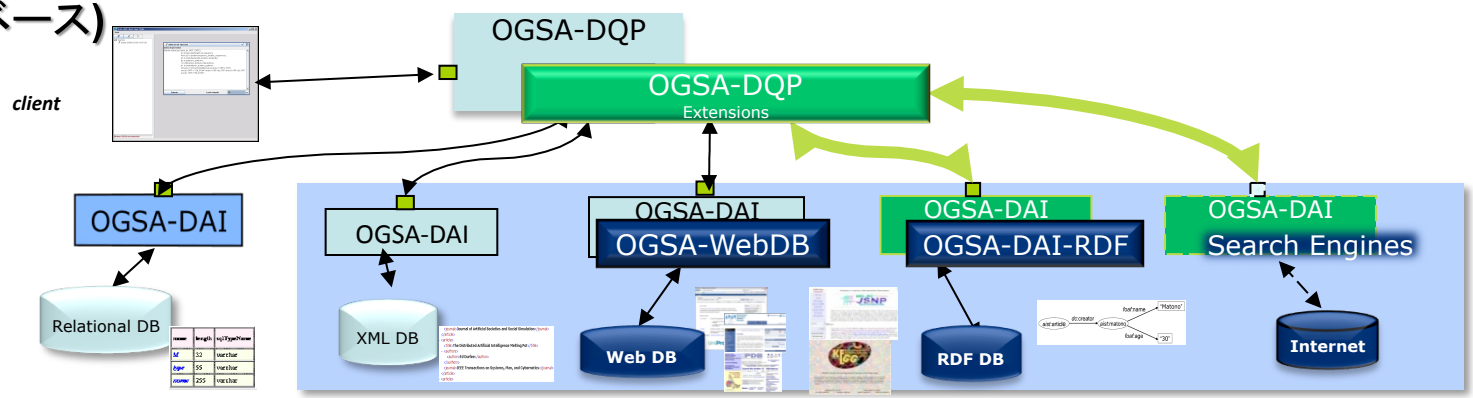


データベース連携・統合技術に関する研究

異種の分散データベースを連携・統合するミドルウェアの研究開発

- 宣言的な言語(SQLやSPARQL)で複数のDBを連携
- DBの異種性(全文検索やRDF)を吸収する機能
- 分散処理の実行順序の(動的な)最適化機能

プロトタイプ
(OGSA-DAI
ベース)



2）メタデータ統合

メタデータにおける事情

- 標準が多過ぎ
  - 地球観測だけでも、、
    - Dublin Core
    - ISO, GML, JMP
    - EO (Earth Observation) profile
- それでもさらにフォーマットが必要、、
  - 衛星やセンサ固有の項目を入れてほしい、、
- その割に項目が埋まってないぞ、、

全文検索エンジンに基づく実装

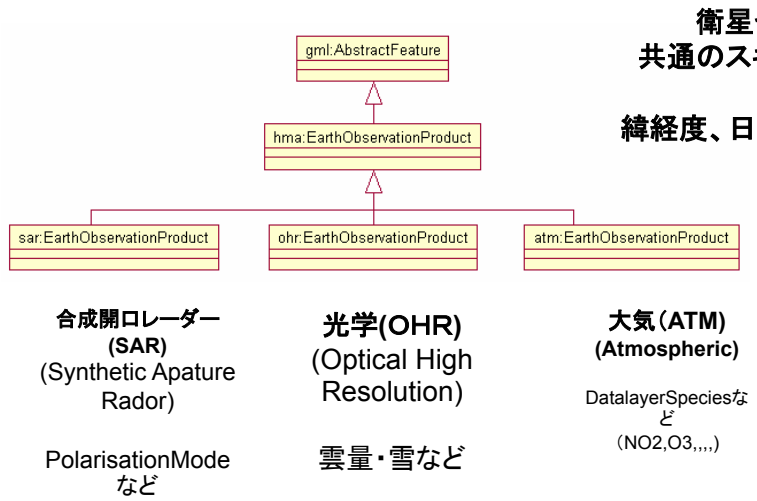
DBMS（**PostGIS**とか）を使わない

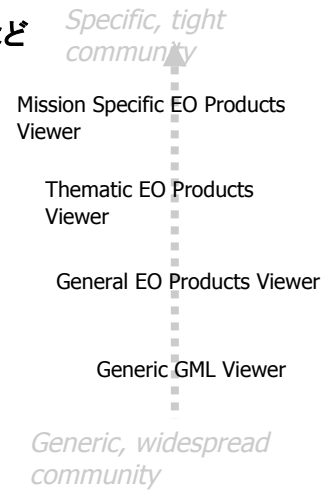
遠目には、、ジオメトリ検索（overlap、cover等）ができる全文検索

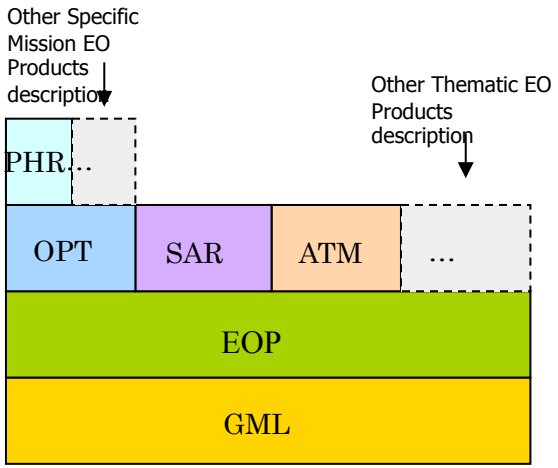
- 様々なXML形式の格納：
  - スキーマレス
- 全文検索：検索結果の数に対する応答性能が高い。
  - ページングなど。
- 地球観測メタデータ
  - 更新がほとんどなく追加のみ。複雑な索引構造でもOK.
  - 構造（時刻や場所）に基づいた検索も可能 (WiSE, Lucene)
- 全文検索
  - やっぱり便利

GEO Gridにおけるメタデータ設計の考え方

- ebRim(e-bussinessの情報モデル)に基づく
  - Webサービスの(OASIS)標準
  - オブジェクト間の関連で記述(拡張性が高い)
  - OGC標準
- モデルの拡張性(スキーマ間の階層関係)

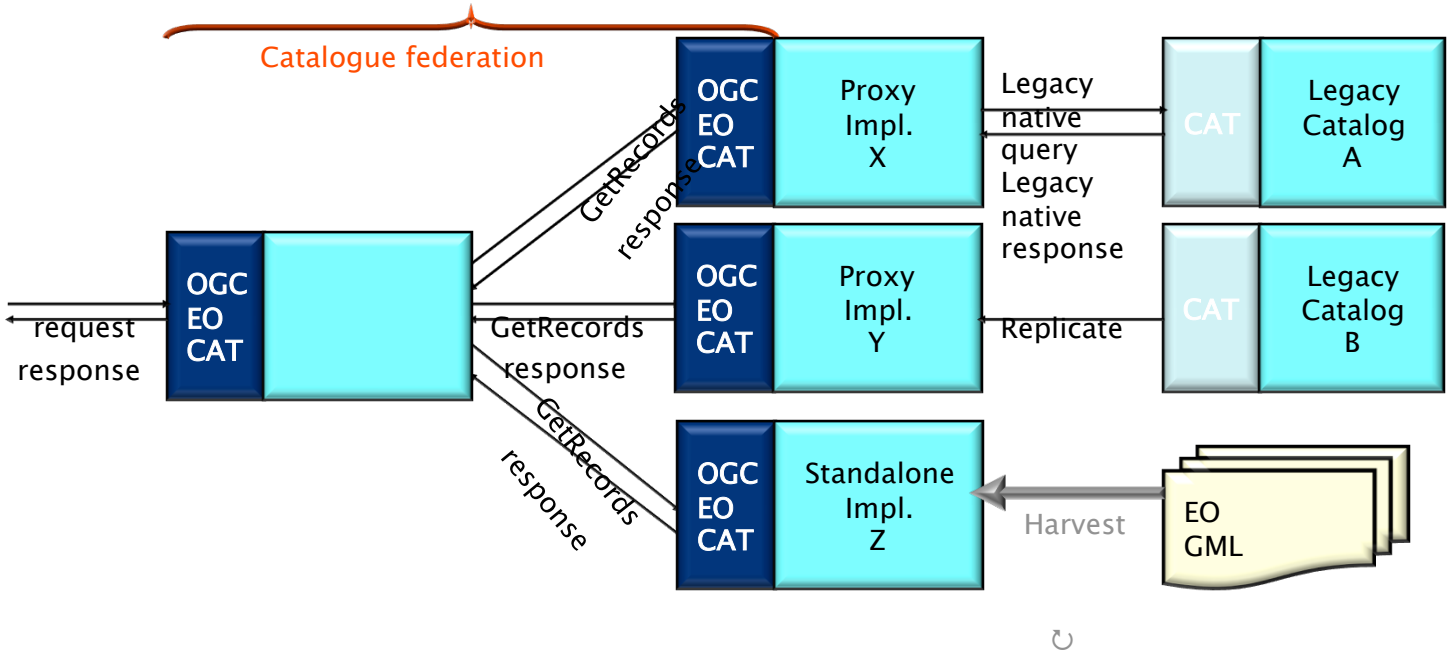






Catalog Service for Web (CS-W)

- GEO Grid でサポートする一連のOGC規格の一つ
  - REST(HTTP GET/PUT) & SOAP
    - OpenSearchの上位互換（次のCSW3.0）
  - 分散カタログの検索を支援→複数組織・横断検索に有利

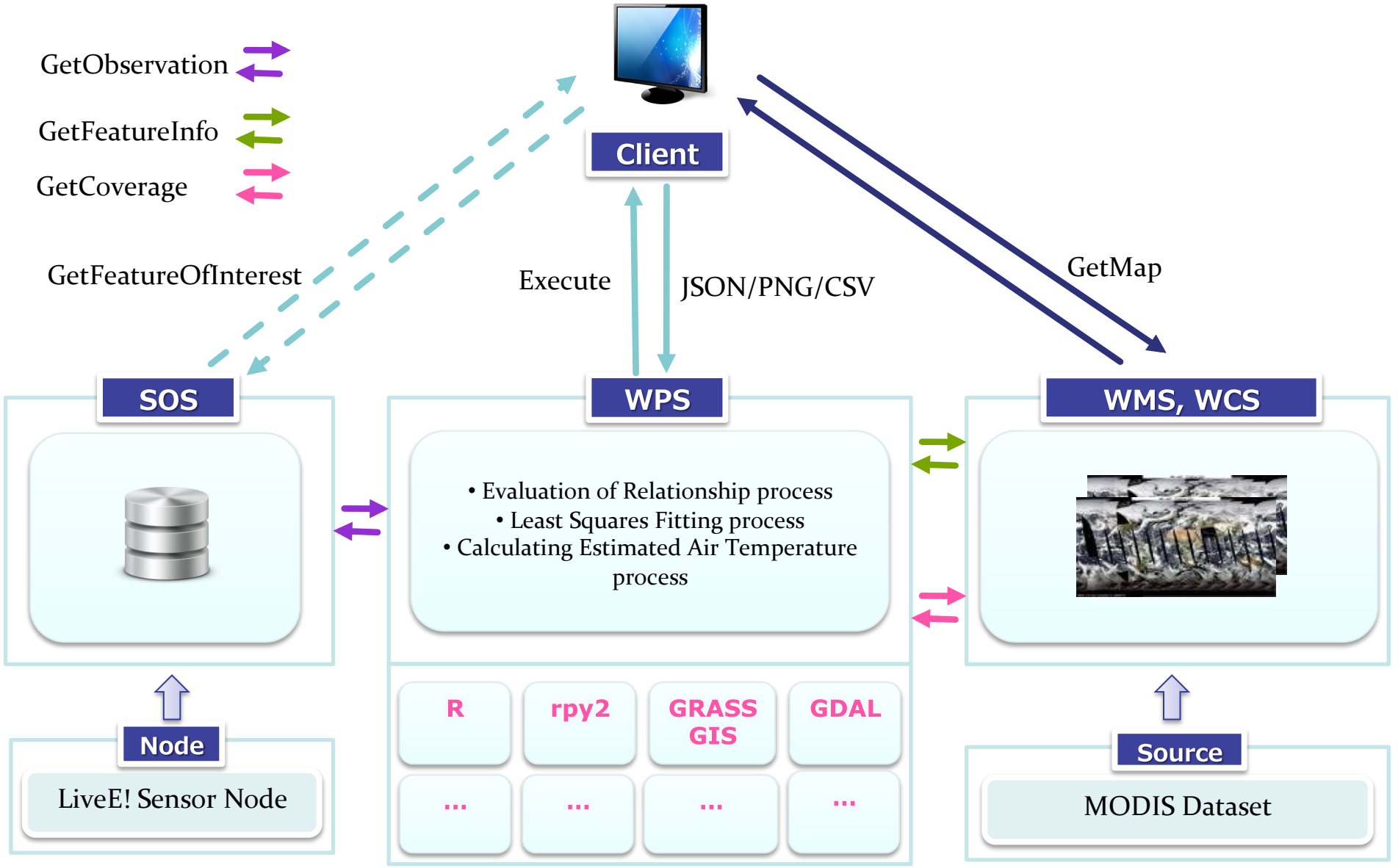


3）データ統合

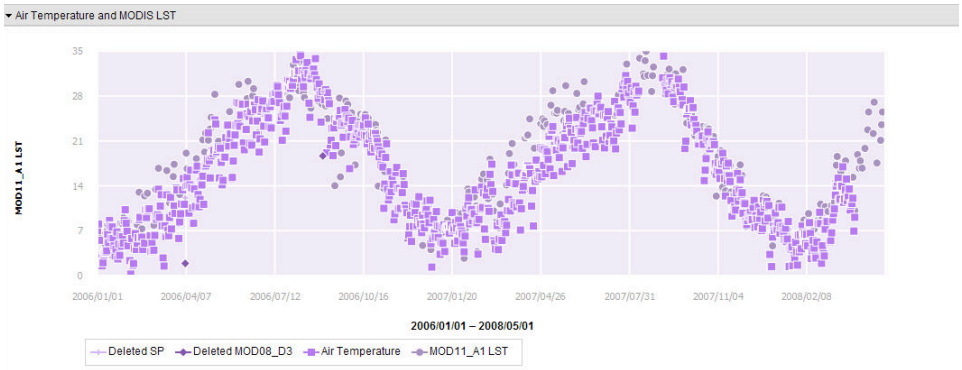
Satellite Field Integrator (SFI)

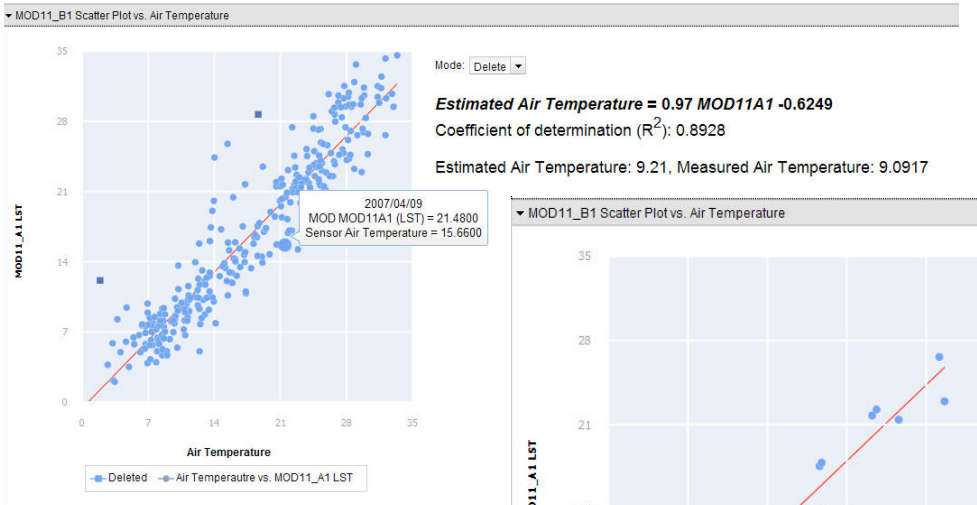
- データ収集、操作、処理の煩雑さを軽減するフレームワーク
  - 衛星データや地上観測データの非均質なデータ形式を処理可能
  - 新たなデータセットをインクリメンタルに導入可能
  - セキュアなデータアクセスを実現
- OGC標準プロトコルを用いて実装されている
  - Web Mapping Service (WMS) ← 画像
  - Web Coverage Service (WCS) ← 生データ
  - Sensor Observation Service (SOS) ← センサ情報
  - Web Processing Service (WPS) ← 処理実行

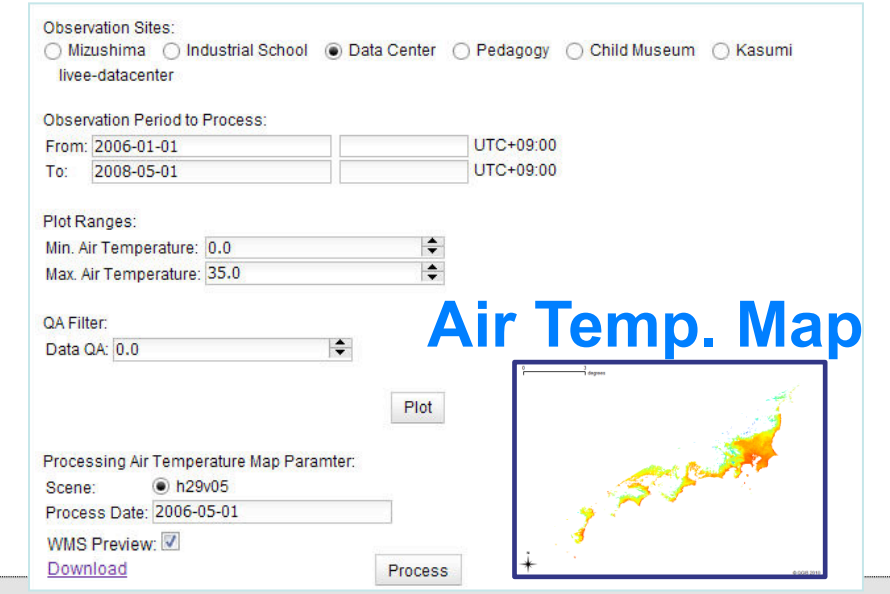
SFI Framework



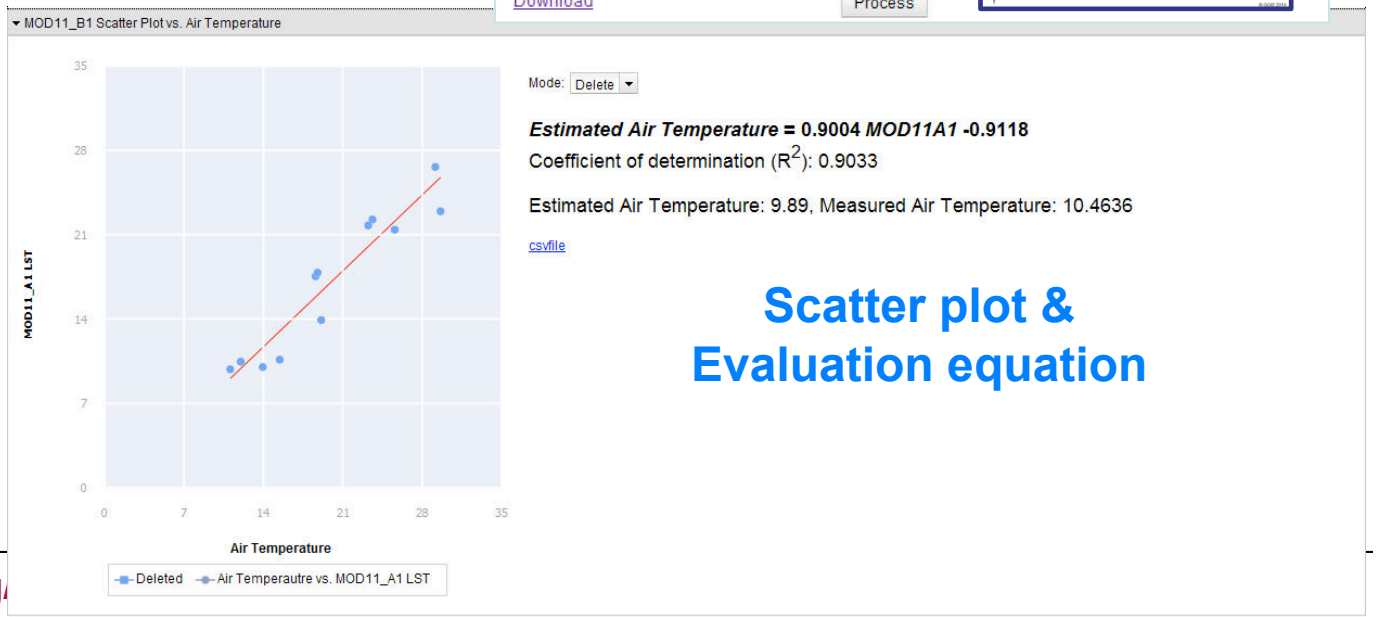
プロトタイプシステム





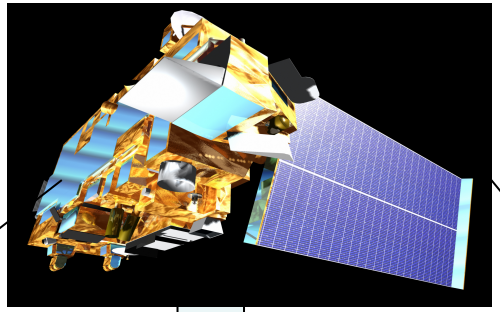


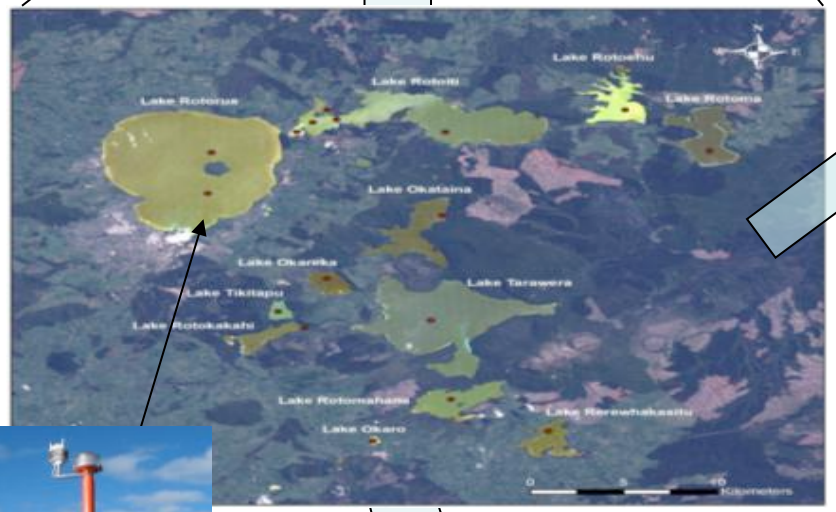
Air Temp. Map

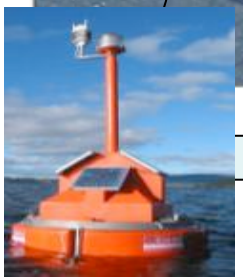


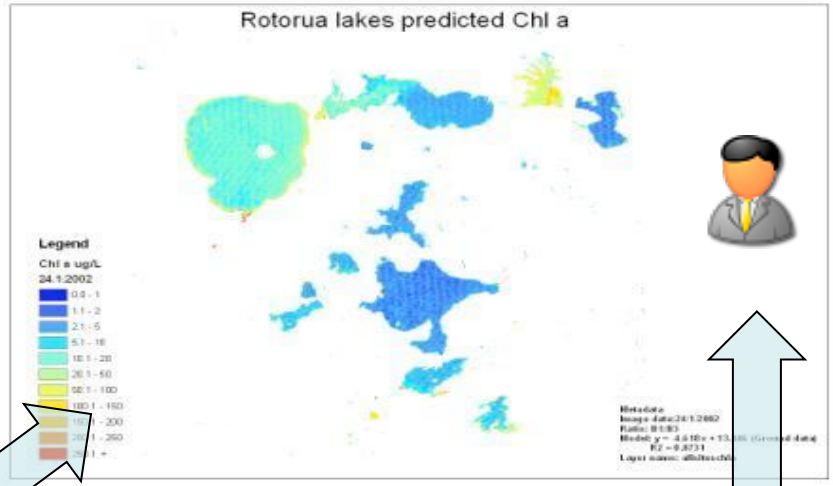
Scatter plot &
Evaluation equation

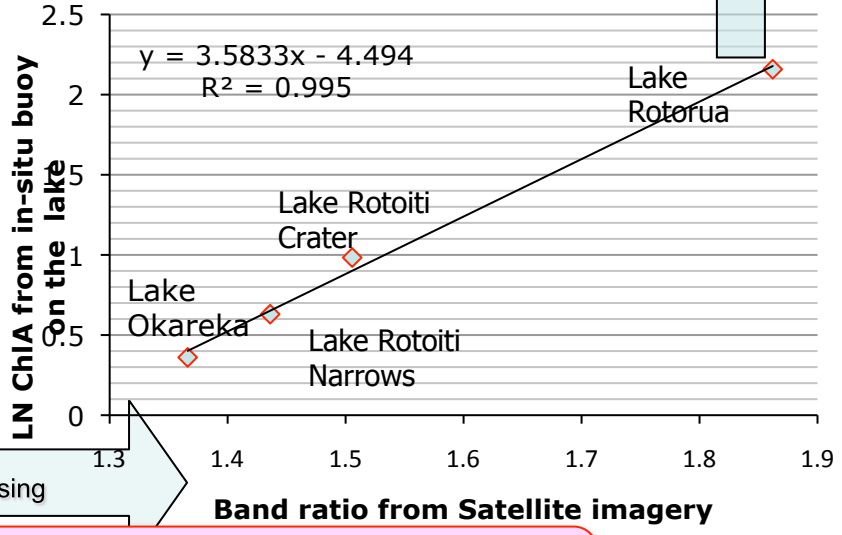
他にも様々な応用例がある(例:湖のクロロフィル量推定)











SOS=Sensor Net

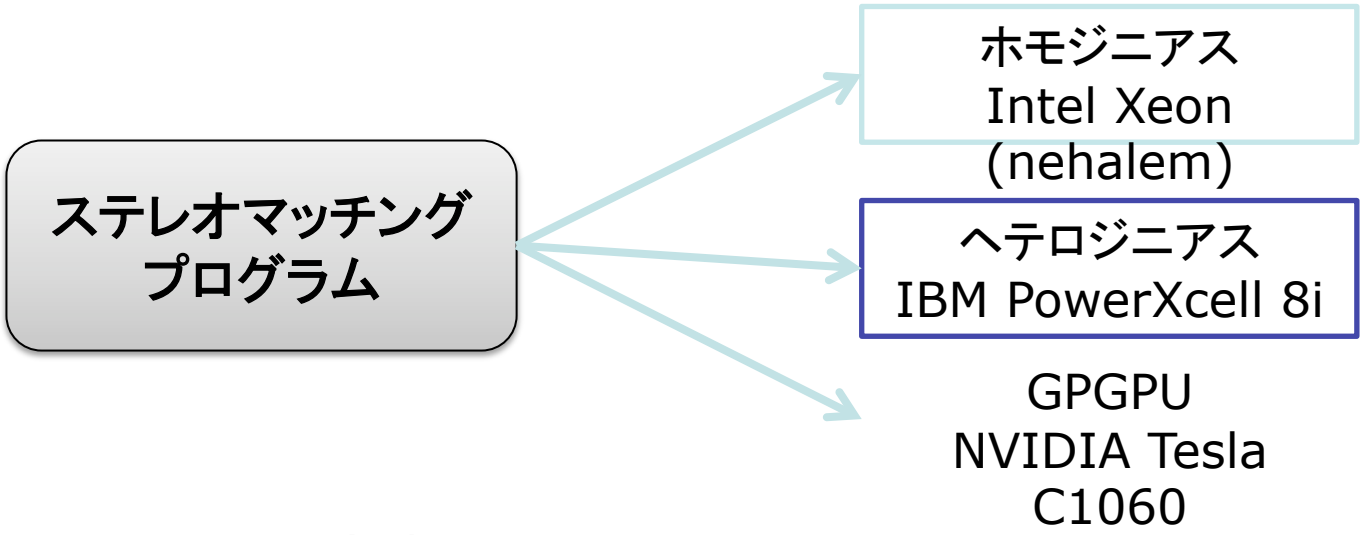
WPS= Satellite data processing

Accurate water quality map production with GLEON

4）データインテンシブなのか？

高性能地理情報システム

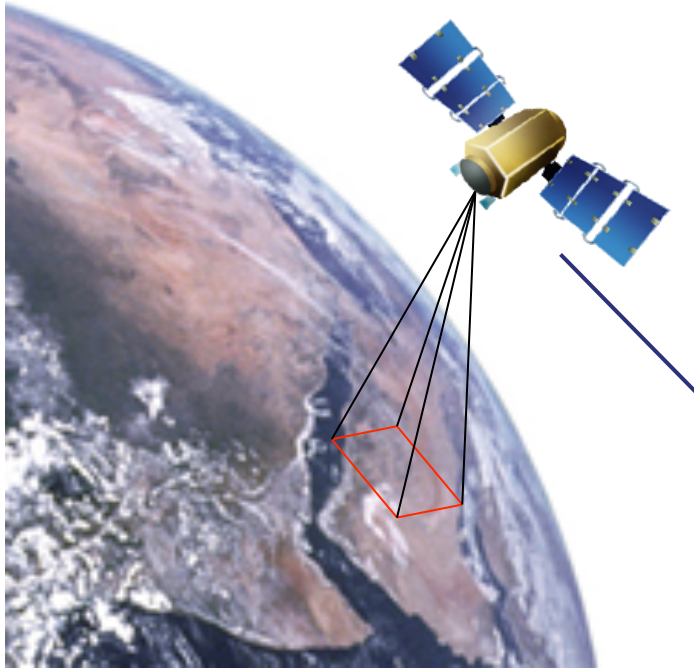
- 商用を含めた地理情報システムのソフトウェアはレガシーなものが多く、最新のマルチコアアーキテクチャを使いこなせていないのではないか？
- 最初のアプリケーションとして DEM 生成ソフトウェア（ステレオマッチング）の高速化を行なった.

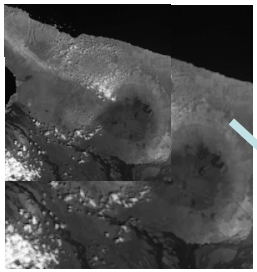


高速化手法とその効果を検証

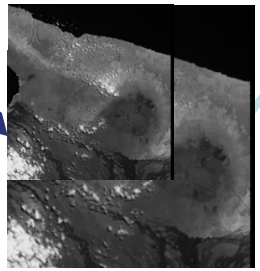
ステレオマッチングによる DEM生成

Digital Elevation Model

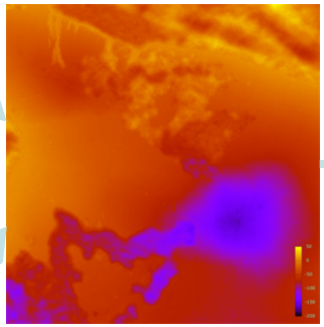




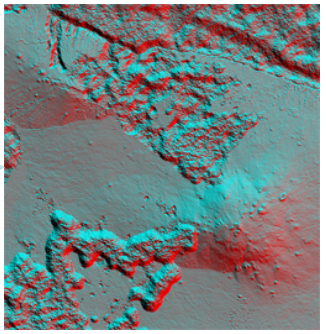
直下視



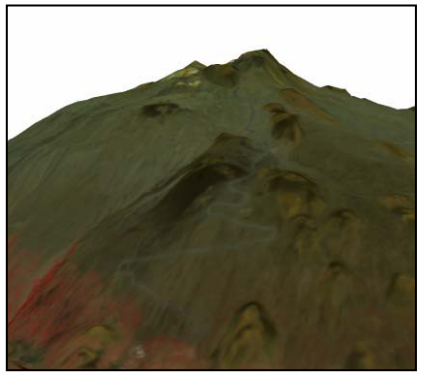
後方視



視差(Disparity)



DEM



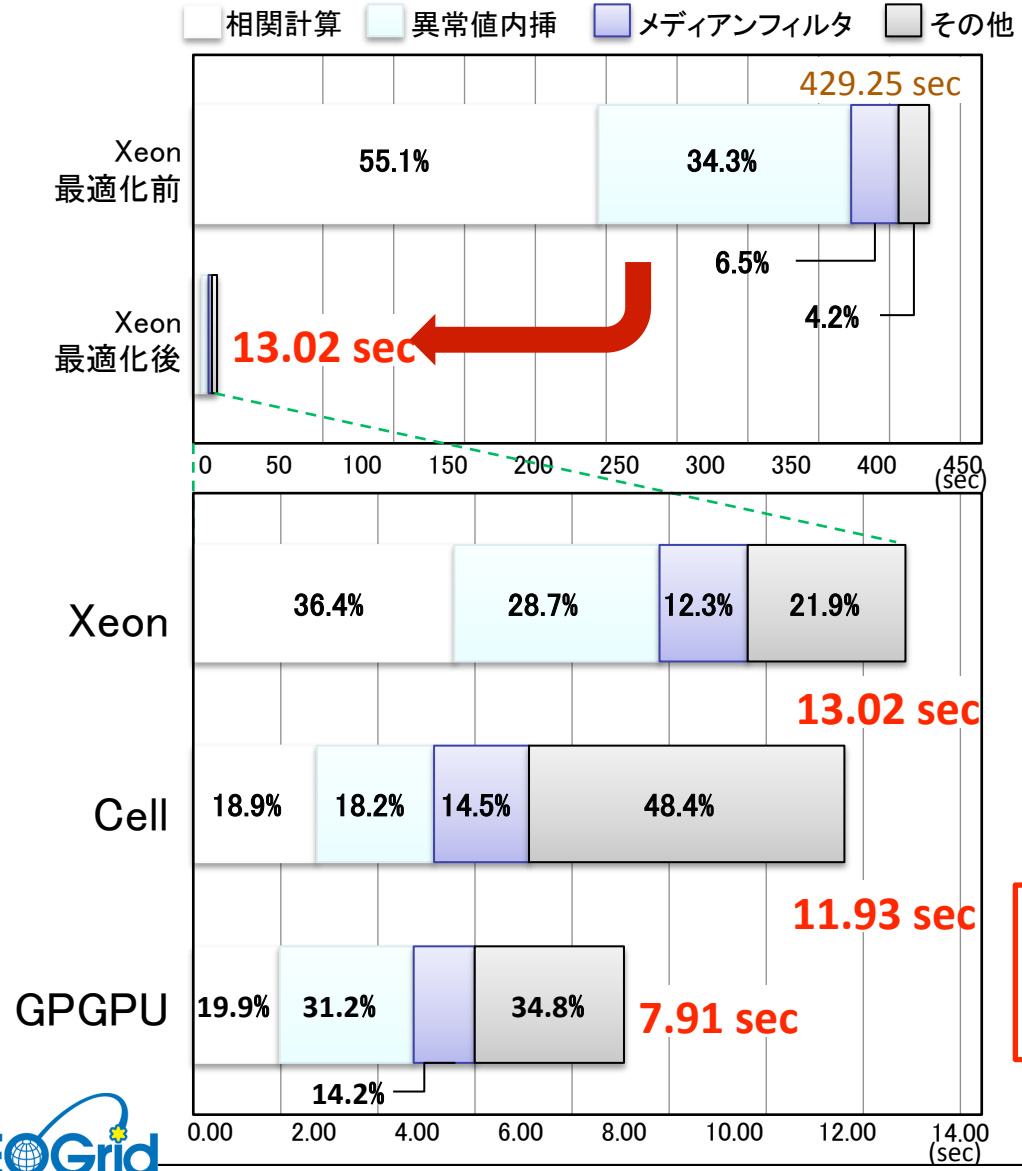
3D表示

- 応用分野
  - 地理情報システム、立体地図の作成、地形分析
- オンデマンドでの作成需要
  - フリーのDEM(GTOPO30,SRTM)で精度が十分でない時。
  - 地盤災害などで直近のデータが必要な時。

最適化前の処理時間

ASTER(4k x 4k) **7**分　PRISM(18k x 18k) **140**分（予想）

全体の実行時間



ファイル入出力
1秒程度

【処理に対して十分小さい】

高速化によって
I/O比率が上がる

別の懸念：
- チューニングはそれなりに大変
- 完全に同じ結果が得られるとは限らない

5）クラウドへ

グリッド v.s. クラウド

	グリッド	クラウド
生まれ	1996年？	2006年？
厳密な定義	ない	ない
曖昧な定義	グリッドは広域ネットワーク上の計算、データ、実験装置、センサー、人間などの資源を仮想化・統合し、必要に応じて仮想計算機や仮想組織を動的に形成するためのインフラ。 以下はIan foster氏によるチェックリスト。 ① 集中管理されていない分散した資源のコーディネート ② オープンスタンダードなプロトコルやインターフェースの利用 ③ 単純には得られない質の高いサービスの提供 （グリッド協議会HPより）	以下の条件を満たす情報処理、ネットワーク、ストレージを提供するハードウェアベースのサービス。 ① 利用者にとって、ハードウエアの取り扱い（管理）が、高度に抽象化されている ② 利用者は、インフラのコストを経費として支払う ③ インフラに、非常に柔軟性がある（スケーラビリティーがある） （McKinsey調査レポートより）
語源	電力網	雲

クラウド化の動機

- グリッドプロトコルは必要か？
  - GEOなユーザは、OGC標準な人たち。
- グリッドセキュリティは必要か？
  - 第三者転送にはdelegationが必要。
  - とりあえず残すが、代替技術を検討中。
- システムががっつり作ってあって、複雑だった。
  - データサーバと計算サーバが一体化
    - ファイルシステムをマウントしていた
  - 保守が煩雑
  - 拡張性がない
    - よそのリソースを簡単に使えない

現在の構成

WPS

別のサイト

CSW
WMS
WFS
WCS

ストレージ

別のサイト

CS-W
csw.geogrid.org

WMS WFS WCS
ows.geogrid.org

WPS
agcvsvc.geogrid.org

SOS

Apach Mapserver
Apach Mapserver
Apach Mapserver
Apach Mapserver

L0/L1 etc L3 geotiff
L0/L1 etc L3 geotiff
L0/L1 etc L3 geotiff
L0/L1 etc L3 geotiff

basemap QQ
basemap QQ
basemap QQ
basemap QQ

Metadata

R
rpy2
GDAL
GRASS GIS

52NorthS

Node

Sensor Node (e.g. Live E!, JapanFlux)

ASTER Data Archive

AIST Green Cloud

GEOGrid
http://www.geogrid.org/

技術的課題は何か？

- スパコンの非均質性や、複数台あることを意識せずに簡単に使いたい。
  - グリッドの実証実験では非均質性への対応が非常に手間だった。アプリケーションの配備を容易にする必要がある
- ベストエフォートなインターネットで良いのか？
  - ネットワークもクラウドリソースの対象に
- セキュリティは？
  - GSIは偉大な成果だと思うが、本当にPKIベースのGSIが必要か？
- 複数のスパコンからシームレスに見えるストレージの実現
  - データの入出力やVMイメージの共有に使いたい。

まとめ

- グリッドの応用として，GEO Gridは出来た
  - ユーザにも公開している
- スリムになる必要がある
  - 必要なミドルウェアはなるべく減らす
  - 提供するプロトコルも必要最低限に
- 他のサイエンスにもそのまま適用できるはず。
- サイエンスクラウドって、最初グリッドが目指していたものそのまま。
  - グリッドでうまく実現できなかった、同じ轍を踏まないように．．．
  - 重い「なんちゃらミドルウェア」なぞ作らずに、ライトウエイトにサクッと動くものを作らねば。

関連イベント

GEO Grid成果報告会2011

- 3月8日
- 秋葉原コンベンションホール2Fコンベンションホール
- http://docs.geogrid.org/2nd_GGOP

http://www.akibahall.jp/data/access.html より