ネットワークを利用した 惑星探査データの共有・利用技術の検討

○寺薗　淳也（日本宇宙フォーラム）
齋藤　潤、十亀　昭人（西松建設）

terakin@t3.rim.or.jp
http://www.t3.rim.or.jp/~terakin/

○TERAZONO Jun-ya (JSF)
SAITO Jun and SOGAME Akito (Nishimatsu Construction, Co. Ltd.)

The concept study of utilization of planetary exploration data through the network

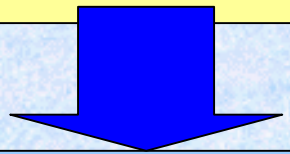
第44回宇宙科学技術連合講演会　2000.10　アクロス福岡

本日の発表内容

- はじめに
- DPLEX構想
  背景、概念…
- DPLEXに必要な要素の検討
  ネットワーク、データ圧縮技術、XML、データフォーマット他
- データ処理に今何が必要とされているのか？
- まとめ

はじめに

- 科学データの量が指数関数的に増大している。
  数年後には、年数十GB、全体で数TB程度のデータが「当たり前のように」生産されるようになる。(ALOSや「すばる」などの例がある)
- これらの大容量データを解析、管理するための体制整備はようやく始まったばかり。
  過去の科学観測データが解析されないまま、廃棄の瀬戸際にある事例もある。



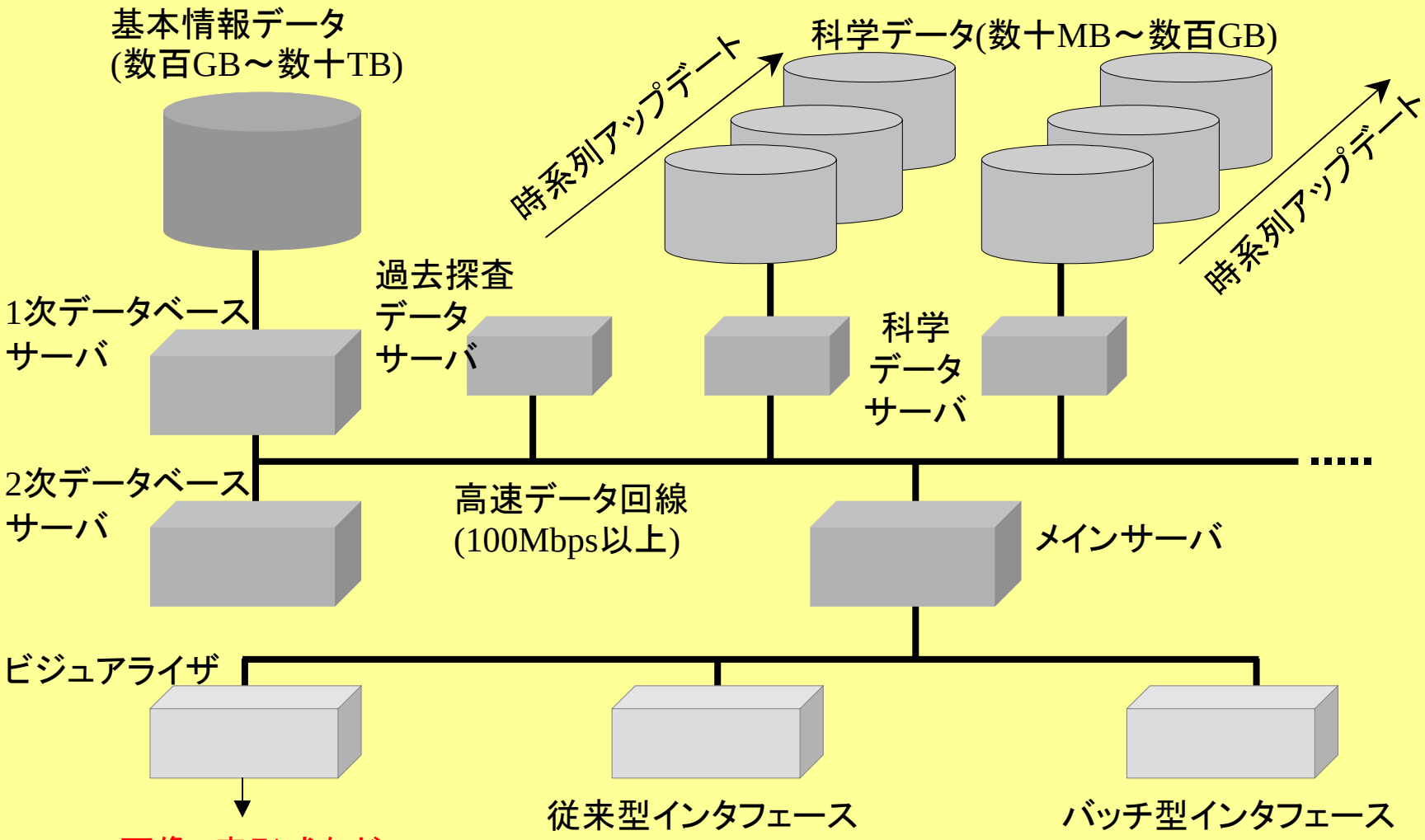
- 処理すべきデータに的確にたどりつけるようにするためのシステムの開発
- 大量のデータをいかにわかりやすく可視化するか

を考える必要がある!!

DPLEXという考え方
(Desktop Lunar and Planetary Exploration)

- ネットワークで結ばれた科学データサーバと、過去の探査データや科学情報などを蓄積したサーバで構成される。
- ネットワークで結ばれた個人や科学者が、自由に(自分のパソコンのデスクトップ上で)惑星探査データにアクセスし、解析を行うことができる。
- データを管理するサーバや、データを適切な形で可視化するビジュアライザなどからなる。
- ネットワーク上ではデータだけでなく、解析などのノウハウも共有することができる。

DPLEXの概念



寺園・齋藤(1999)より

DPLEXの要素

- ネットワーク回線
- データ圧縮技術
- XML
- データフォーマット
- 可視化技術
- サーバ
- データベース

DPLEXを実現させていくためにはどのような要素が必要か？

また、今後の技術革新などを考えて、どのような技術を今後開発していくべきか？

要素(1)　ネットワーク回線

回線	必要な時間
ISDN 回線 (64kbps)	341 分
ADSL 回線 (768kbps)	28 分
T1 回線 (1.5Mbps)	14 分
ATM (622Mbps)	2 秒

160MBのデータを送信するために必要となる時間

- 大容量データをネットワーク経由で解析する場合、実用となるのはT1回線以上のスピード
- 複数のデータセットを解析したいとなれば、より高速な回線を(安価に)使える環境が必須。
- バックボーン回線には最低でもギガビットクラスの回線を設置する必要がある。

要素(2) データ圧縮技術

- 圧縮技術には、可逆圧縮と非可逆圧縮があるが、科学データの圧縮では可逆圧縮をメインに考えるべき。
  ただ、ブラウズデータなどの場合には、非可逆圧縮を積極的に採用することも考えられる。
- 科学データの圧縮の場合、可逆データの圧縮率は平均して約50%。非可逆ならば10～30%までの圧縮が可能。
- 送受信の際のサーバプログラムに圧縮アルゴリズムを組み込んだり、データフォーマットが最初から圧縮をサポートするような仕組みが望ましい。
  画像のPNGフォーマットなどが代表的な例。
- ビニングや階調幅縮小など、データ圧縮以外のデータ逓減の手法も考慮すべきである。

要素(3) XML

XMLは、現在最も注目されている、汎用の情報流通基盤である。
多様なデータフォーマットなどを統一したやり方で扱うための方式として注目されている。

- データスキームを定義する言語であり、拡張性が最初から前提となっている。
- 標準化プロセスが進められていて、既に様々なXMLパーザが発表されている。
- 出力形態としてHTMLが前提となっていて、既存のインターネット技術と親和性が高い。
- データ可視化に関連の深い技術(3D、ベクトル表現、数式など)の開発、標準化が進んでいる。

XMLを利用した科学データの取り扱い

XMLでの科学データの取り扱いについては、その扱い方を定めたファイル(文書型定義、DTD)を作り、正式な規格として採用されるように働きかける必要がある。

- DTD開発の動きは天文分野が中心となって進んでいる。
- NASA GSFCのAstronomical Data Center XML、イリノイ大学のD. Guillaume氏によるAML (Astronomical Markup Language)などが代表的。

XML自体は標準規格だが、拡張が用意であるため、標準が「乱立」する事態も起こり得る。

規格の提案だけでなく、標準化プロセスや、分野を超えて共用できるようなDTDなどの開発を進めていく必要がある。

要素(4)　データフォーマット

現在、惑星科学分野で広く使われているフォーマットとしては、netCDF、PDS、FITSフォーマットがある。

- netCDFはUNIDATA(UCAR)で開発されたフォーマットで、自己記述性と、マシン独立のデータ収納方式が特徴。
- PDSはNASAの月・惑星データを収めるために使われているデータフォーマット。シンプルであるが拡張が容易で、ほとんどの惑星探査データがPDS形式で格納されている。
- FITSフォーマットは天文学で広く使われている共通フォーマット。データ構造が厳密に定義されているために互換性が極めて高い。

既存のデータフォーマットを DPLEXで扱うために

- XMLは、3D可視化やネットワーク上でのデータ流通などに、威力を発揮する。
- しかし、既存の膨大なデータフォーマットをすべてXMLなどに変換することはまず不可能。
- 既存のデータフォーマットをXMLに変換してネットワーク上に流通させる「変換ゲートウェイ」や、XMLをラッパーとして利用し、データ解析手法やメタデータなどを同梱してネットワーク上に流す技術などを開発する必要がある。
- 巨大なデータに対応するためには、データの一部を必要に応じて切り抜くなど、実際の要求に即した技術が必要。

要素(6)
データ可視化、サーバ、データベース

- DPLEXではデータを3次元で可視化することを前提としている。しかし、リアルタイムでのデータ3次元化はコンピュータの能力の面で難しい。あらかじめ構築されたDEMなどを必要に応じ切り出し、データを重ね合わせる技術の開発が必要。
- サーバはTBクラスのデータを扱うため、テープやCD/DVD-ROMライブラリをバックエンドに持ち、ネットワークでシームレスに結合されたサーバ群を構築する必要がある。
- 科学データベースでは、一般的に使われるデータ処理手法をあらかじめデータベースシステムに内包してしまうことにより、データ量の逓減や研究の効率化につながる。
- データベースの設計では、科学データの外側にあるメタデータの扱い方を考慮する必要がある。

何が必要か？

- 何を行えば研究活動を効率化できるのか？
  最も時間・手間がかかっているのはどこなのか？
  惑星科学データについていえば、**意外と基本的な部分である。**例えば画像データのオープンやデータの読み書きなど。基本的なことさえできてしまえば、あとは自然にノウハウが集まって、システムが構成されていく。その基本的な部分がないのが現在の問題点。
- 数年後はどのような技術が標準になっているのか？
  回線速度は飛躍的に高まるが(おそらくは数Mbps)、データ量の増大がそれを食い潰してしまう。また、特に天文分野などでは、**膨大なアマチュアの研究者を考慮する必要がある。**
- 研究スタイルを変えていく必要性
  研究室レベルだけでなく、学会、大学レベルでの「ノウハウの共有」を実現させていく必要がある。**惑星探査はもともと幅広い研究者層を結集させる必要がある**ため、DPLEXのような広い基盤をまとめていくインフラストラクチャが必須である。

まとめ

- 惑星科学の膨大な探査データをどうするか、組織的な動きを起こす必要がある。
  探査の本格化を控えて、待っている時間はない。とくに、これまでのノウハウを抱えた天文学分野から学ぶことは大きい。
- まずは要素技術からの実装、研究開発が必要
  探査データに適した圧縮技術、データのインテリジェントな切り出し技術、データの質についてのインデックス付けなど…。
- 既存技術がどこまで活用できるか
  今後数年間の技術の発展方向を見据えながら、使える技術は積極的に採用していく。
- そして、まずは作ってみること
  プロトタイプなどを作り、相互に競い合いながら、技術の標準化を進めていくべき。(司令塔の存在が前提)