

JVO 開発における 大規模天文データ処理

国立天文台 天文データセンター
白崎 裕治

急増する天文データ

✓ 天文データは毎年倍増

- CPU の計算性能は18カ月で2倍
- I/O 性能の向上率は 10%/年
- 並列計算処理は必須。
- データ移動を極力避ける解析システム

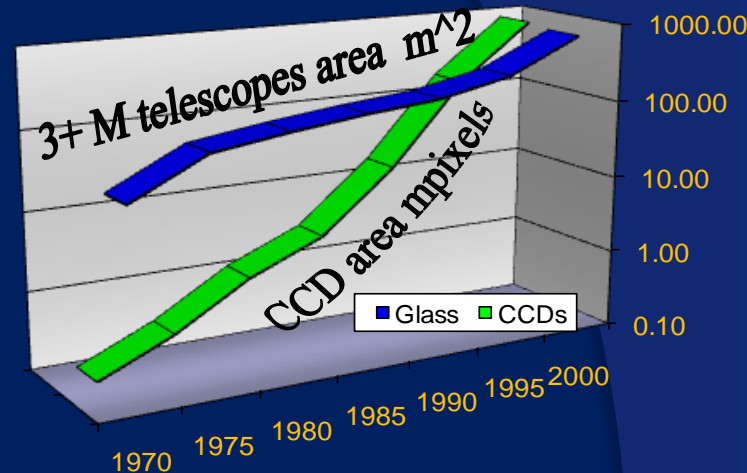
✓ 望遠鏡・観測装置の大型化、高機能化

- データ取得のコストの増大 → 科学成果の最大化
- 取得したデータを速やかに解析できる環境を構築し、国内研究者間で共有できる仕組みづくり

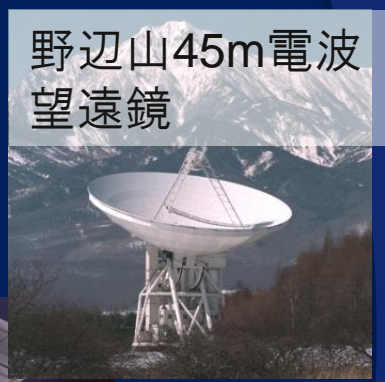
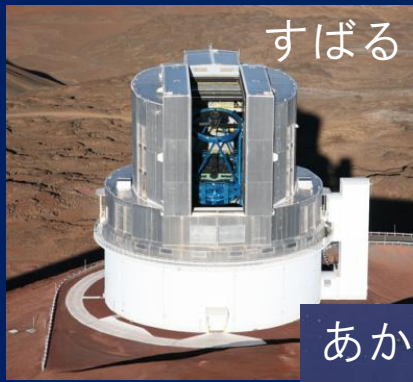
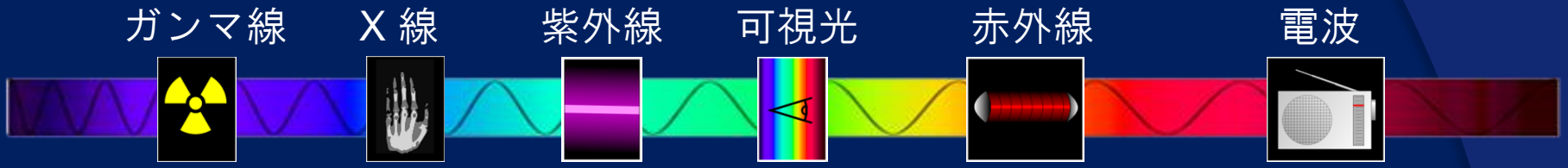
✓ 中小望遠鏡との連携観測の進展

- ガンマー線バースト、AGN の時間変動 (Time Critical な観測)
- Real time で解析し、結果を速報

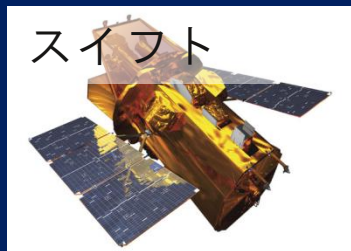
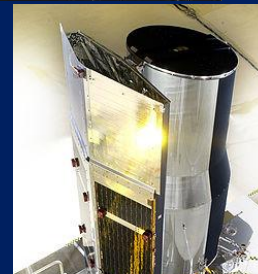
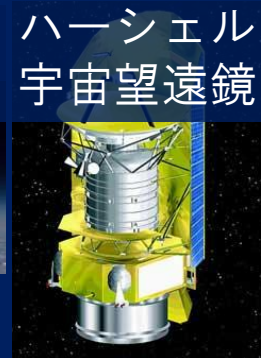
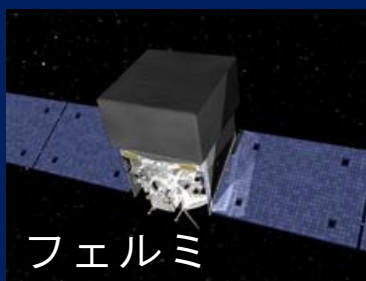
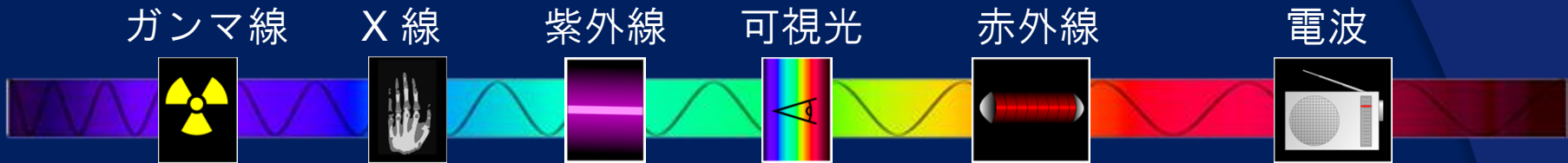
Courtesy of Alex Szalay



稼働中の主要な望遠鏡 (国内)

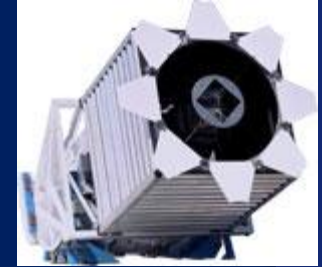


稼働中の主要な望遠鏡 (海外)



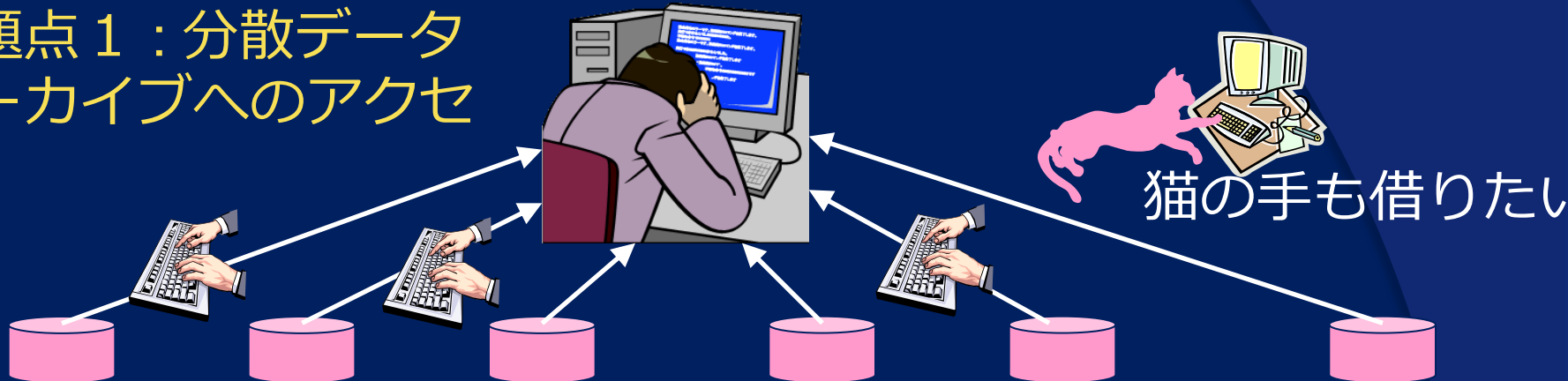
インターネット上の天文研究リソース

- <http://www.aanda.org/>
- <http://www.journals.uchicago.edu/ApJ/>
- <http://www.blackwellpublishing.com/journal.asp?ref=0035-8711&site=1>
- <http://www.asj.or.jp/pasj/>
- <http://adsabs.harvard.edu/>
- <http://www.arxiv.org/>
- <http://cxc.harvard.edu/>
- <http://heasarc.gsfc.nasa.gov/>
- <http://irsa.ipac.caltech.edu/>
- <http://lambda.gsfc.nasa.gov/>
- <http://archive.stsci.edu/>
- <http://nedwww.ipac.caltech.edu/>
- <http://nssdc.gsfc.nasa.gov/>
- <http://www.spitzer.caltech.edu/>
- <http://cdsweb.u-strasbg.fr/>
- <http://cfa-www.harvard.edu/iauc/SearchIAUC.html>
- <http://www1.cadc-ccda.hia-ihp.nrc-cnrc.gc.ca/cadc/>
- <http://skyview.gsfc.nasa.gov/>
- <http://archive.eso.org/>
- <http://dbc.nao.ac.jp/>
- <http://idlastro.gsfc.nasa.gov/>

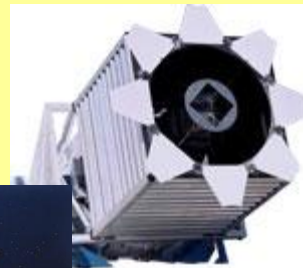


- <http://nrodb.nro.nao.ac.jp/>
- <http://www.darts.isas.jaxa.jp/>
- <http://www.sdss.org/>
- <http://www.ipac.caltech.edu/2mass/>
- <http://www.ukidss.org/>
- <http://www.astro-wise.org/>
- <http://terapix.iap.fr/>
- <http://www.roe.ac.uk/ifa/wfau/>
- <http://www.jach.hawaii.edu/UKIRT/>
- <http://www.cfht.hawaii.edu/Science/CFHTLS/>
- <http://swire.ipac.caltech.edu/swire/swire.html>
- http://www.oamp.fr/virmos/virmos_vvds.htm
- <http://deep.ucolick.org/>
- <http://www.eso.org/science/eis/>
- <http://www.galex.caltech.edu/>
- <http://www.stsci.edu/science/goods/>
- <http://www.ast.cam.ac.uk/~wfcsur/>
- <http://www.noao.edu/noao/noaodeep/>
- <http://www.esa.int/SPECIALS/ESAC/index.html>
- <http://www.eso.org/public/astronomy/archive.html>

問題点1：分散データ アーカイブへのアクセス

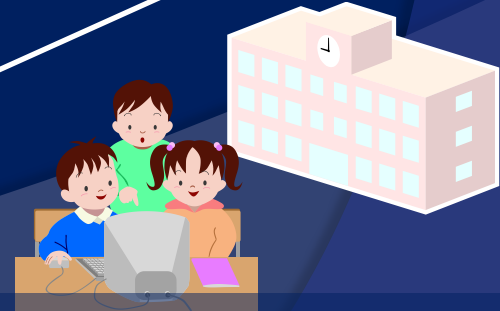


猫の手も借りたい



データベースアクセスインターフェースの共通化

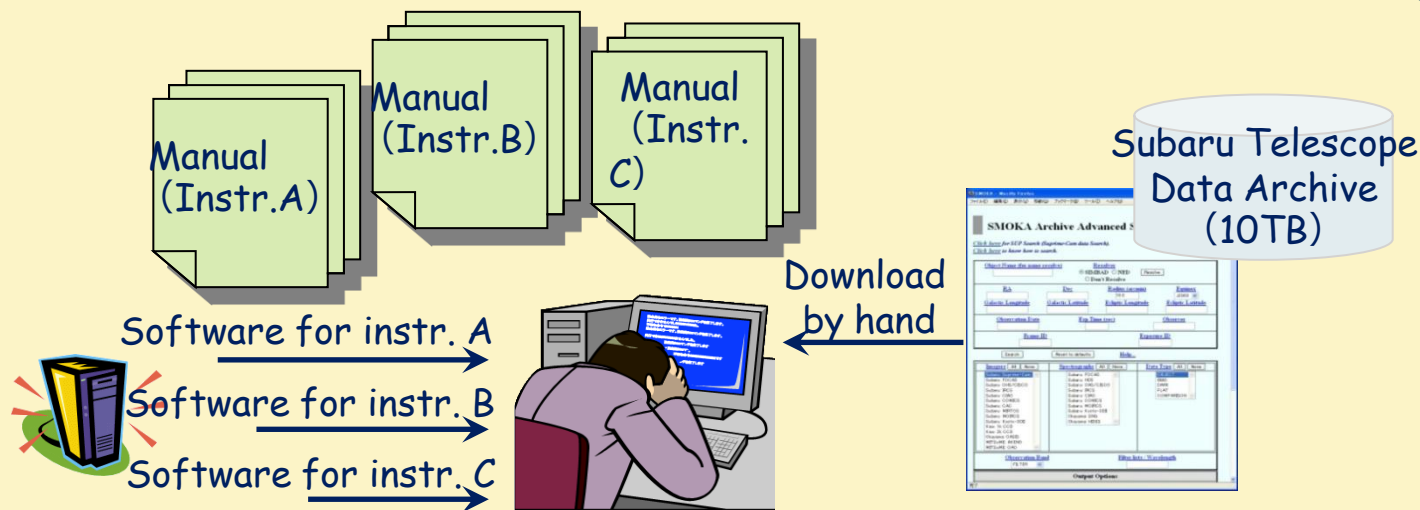
データ
取得を
自動化



教育用教材としても利用可

問題点2：巨大観測データの処理

- ◎ データアーカイブからのダウンロード困難
- ◎ 大容量ディスク＋並列計算システムが必要
- ◎ 生データのリダクションは装置毎に異なりラーニングコストが高い



解決策：データと計算資源の集約化

◎ 解析機能付きデータアーカイブ

- データ移動のコスト最小化。
- 並列計算機システムを独自に持つ必要がない。
- リダクションソフトの管理を一元化。過去バージョンのソフトによる再リダクションをサポート。
- 請求の多い処理済みデータはアーカイブに蓄積。同じリダクションを繰り返さない。

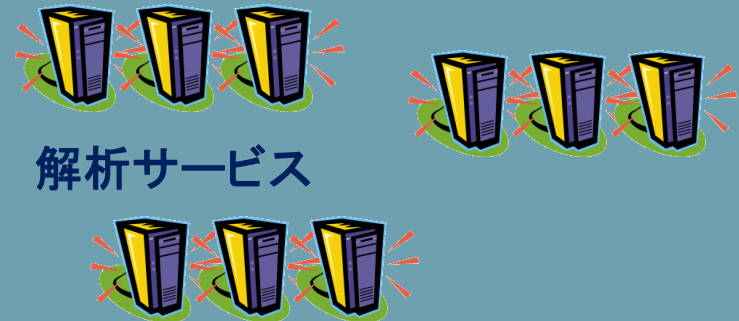


バーチャル天文台

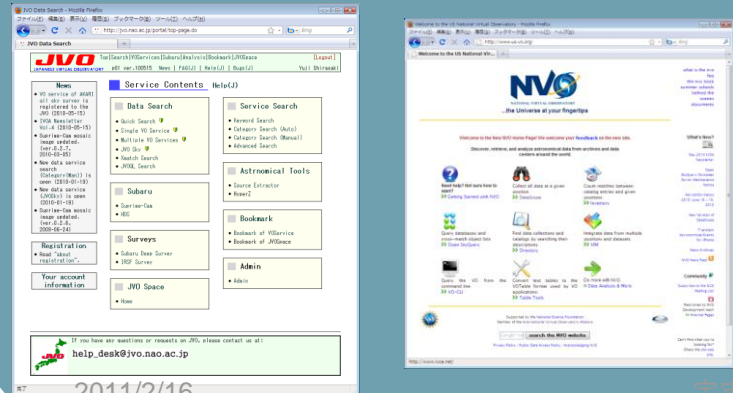
バーチャル天文台とは？

国際標準のインターフェイスを備えた天文データベースを利用したソフトウェアシステム。より高度な**データ共有**を可能にし、効率的な研究の支援をする。

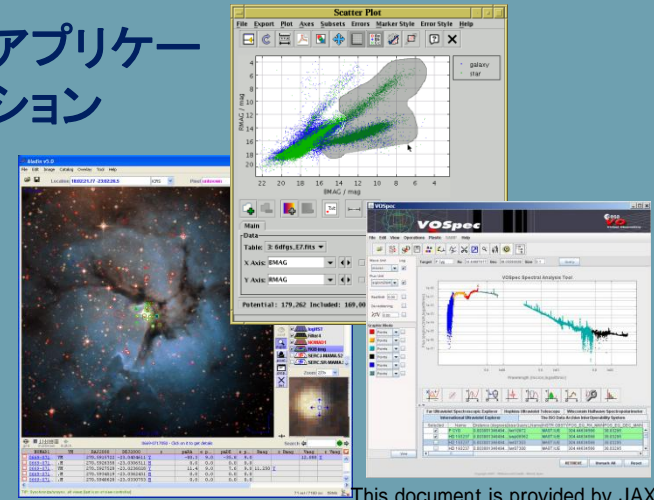
データサービス



Webポータル



アプリケーション



International Virtual Observatory Alliance

- 世界各国の 18 VO プロジェクトが参加
- 天文データの共有をより効率的に行うための標準仕様策定団体

<http://www.ivoa.net/>



- 策定された主な仕様 (全部で31の仕様)

IVOA Registry Interfaces – データサービスの公開方法

Resource Metadata for the Virtual Observatory – データサービスのメタデータ仕様

VO Table Format Specification – 検索結果等のデータフォーマット

Simple Cone Search

Simple Image Access

Simple Spectral Access

Table Access Protocol

データ検索のインターフェイス定義

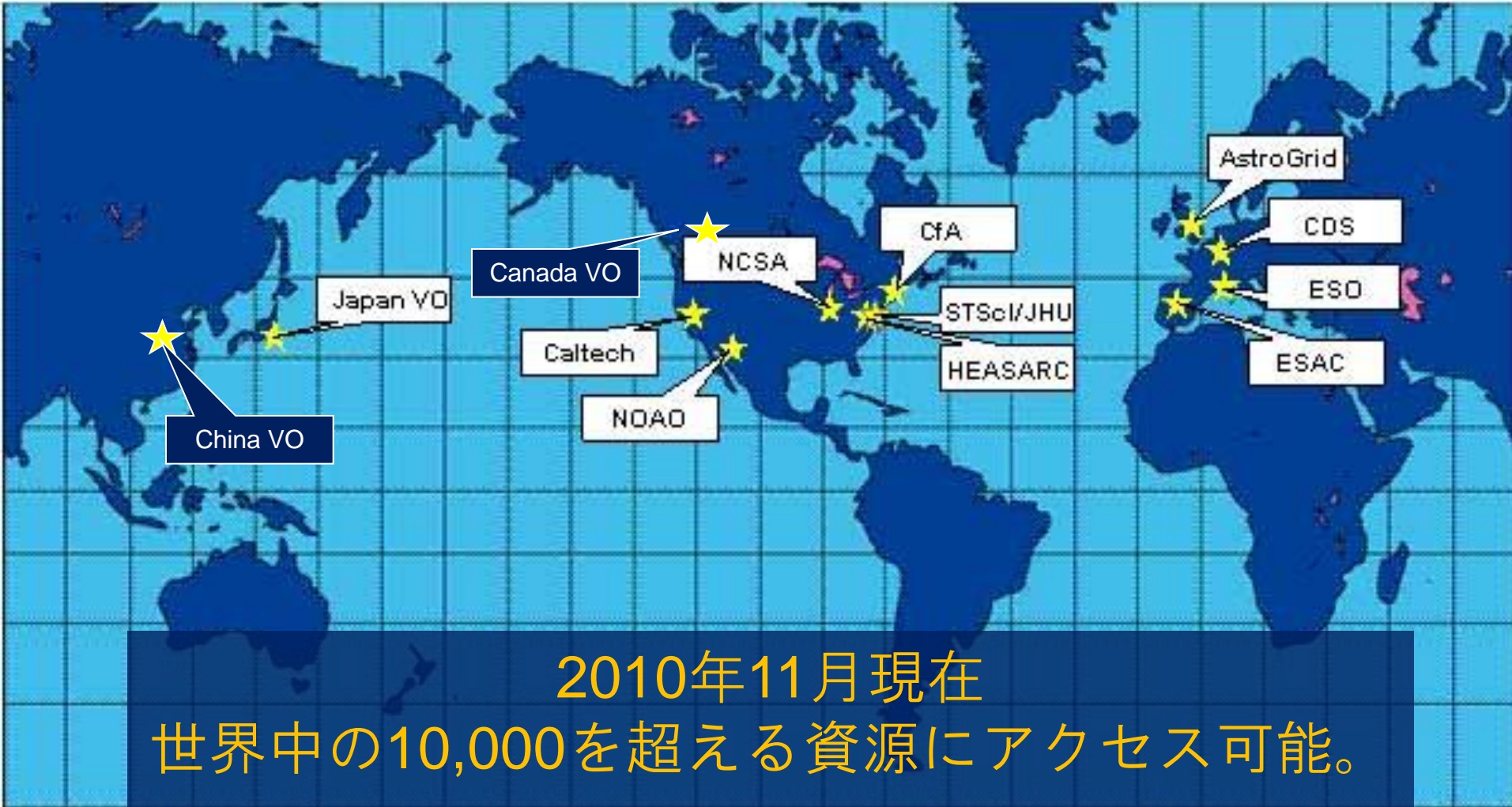
IVOA Astronomical Data Query Language – 検索言語仕様

VOSpace service specification – 分散ファイルシステムを実現する仕様

Data Model for Astronomical DataSet Characterisation – データモデル

Simple Application Messaging Protocol – アプリケーション間連携の仕様

バーチャル天文台サイト



2010年11月現在
世界中の10,000を超える資源にアクセス可能。

<http://jvo.nao.ac.jp/portal/>

JVO portal の開発

- ✓ Web ベースの天文データ検索・解析システム
- ✓ 開発体制 (設計・開発の実務者)
国立天文台 (2名)、富士通 (2名)、セック (1名)

開発ターゲット 二つの柱

1. だれでも簡単に利用できるシステム

汎用的でシンプルな機能。

大多数の利用者むけ。教育などでの利用も。

2. バーチャル天文台ならではの研究ができるシステム

少数のパワーユーザ向けに高度な機能を提供。

最先端の研究を可能にする。

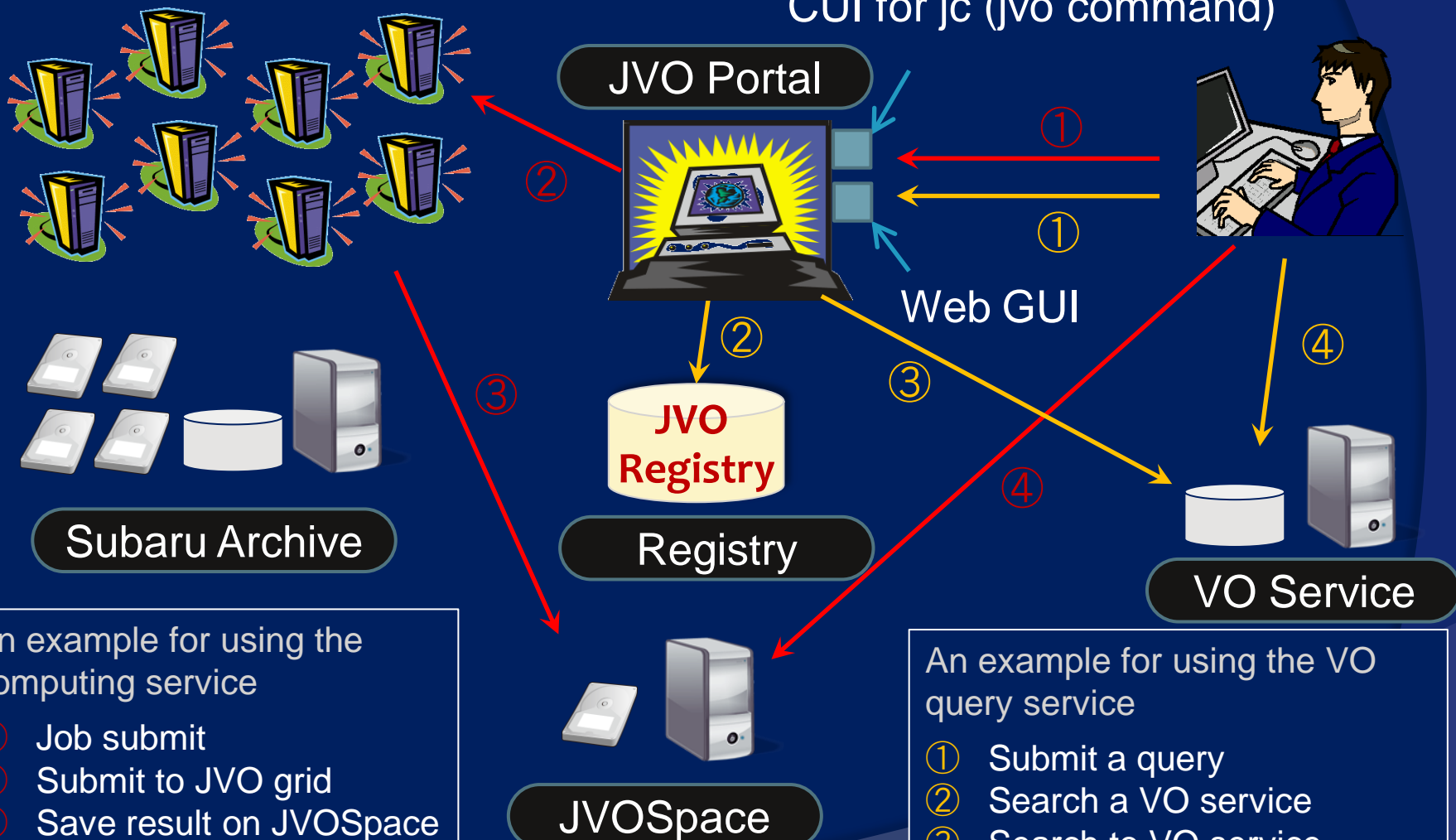
両者の均衡を考えながら、バランスのとれた開発を進める。

JVO system の概略

五つのサブシステムと分散データサービス (VO Service)

JVO Computing GRID

CUI for jc (jvo command)



An example for using the computing service

- ① Job submit
- ② Submit to JVO grid
- ③ Save result on JVOSpace
- ④ Retrieve the result

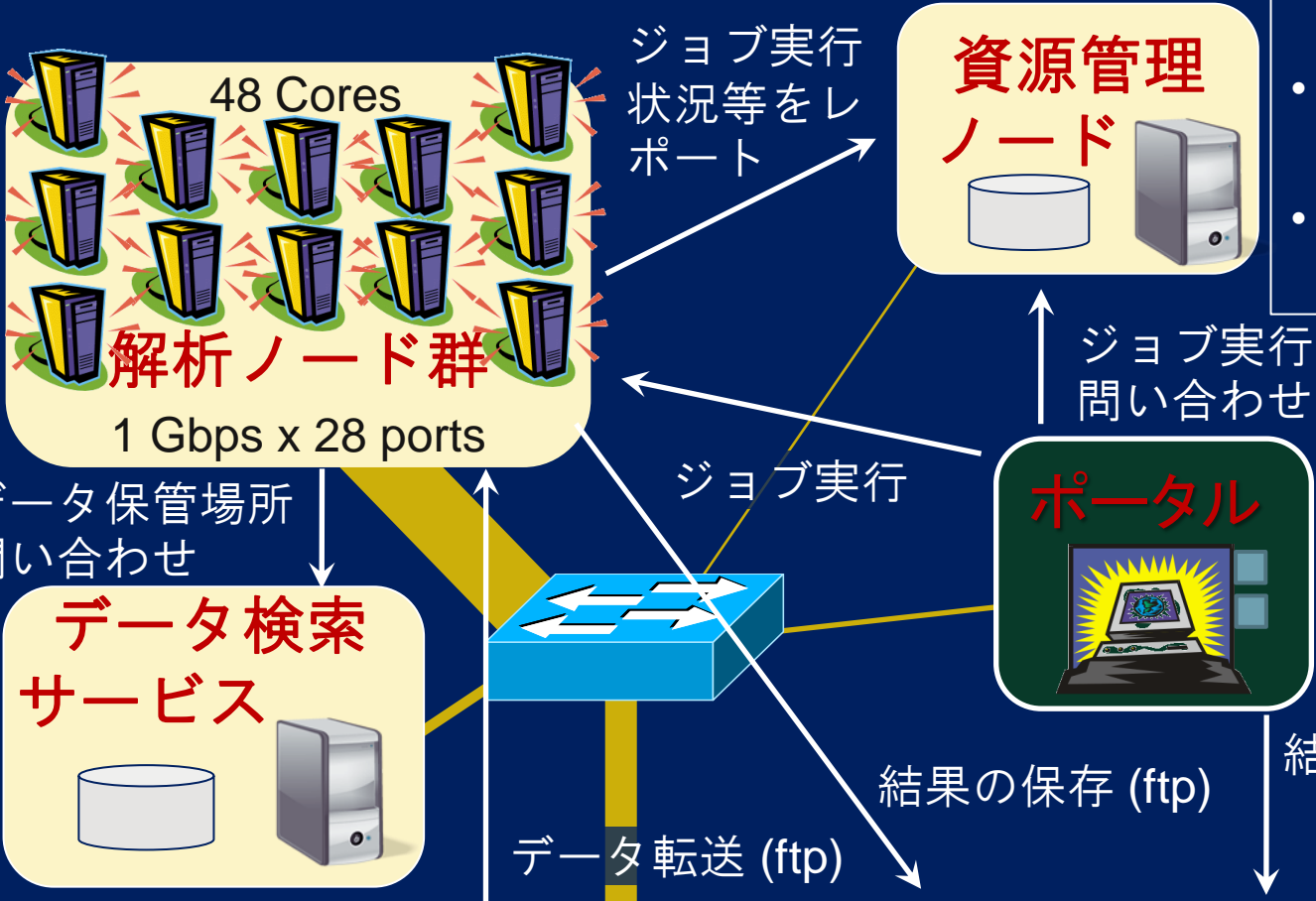
An example for using the VO query service

- ① Submit a query
- ② Search a VO service
- ③ Search to VO service
- ④ Retrieve a FITS image

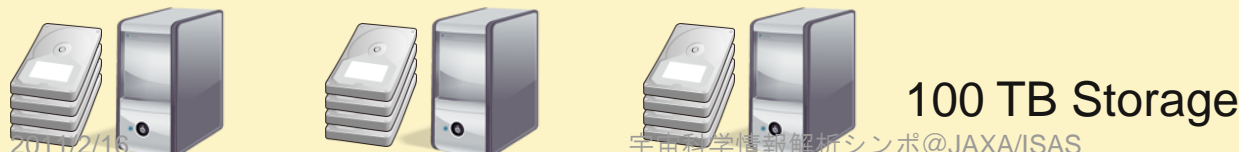
JVO Computing GRID

きわめて単純な仕組み

- ラウンドロビン方式でジョブ実行
- 解析データはローカルディスクに ftp 転送
- 結果はディレクトリ毎コピー



すばる望遠鏡
Suprime-Cam の
全データ (10TB)
を約 2 週間で解析
完了



JVO ポータル (GUI) の機能

データ検索

1. 高速検索

JVO 内部のデータベース
主要なカタログを登録

2. サービスを一つ指定して検索

サービスを検索し選択
テーブル選択
検索条件指定

3. 複数サービスへの同時検索

領域指定検索のみサポート

4. JVO Sky サービス

Google Sky IF を利用したデータ検索

データサービス検索

1. キーワード検索

2. カテゴリ検索

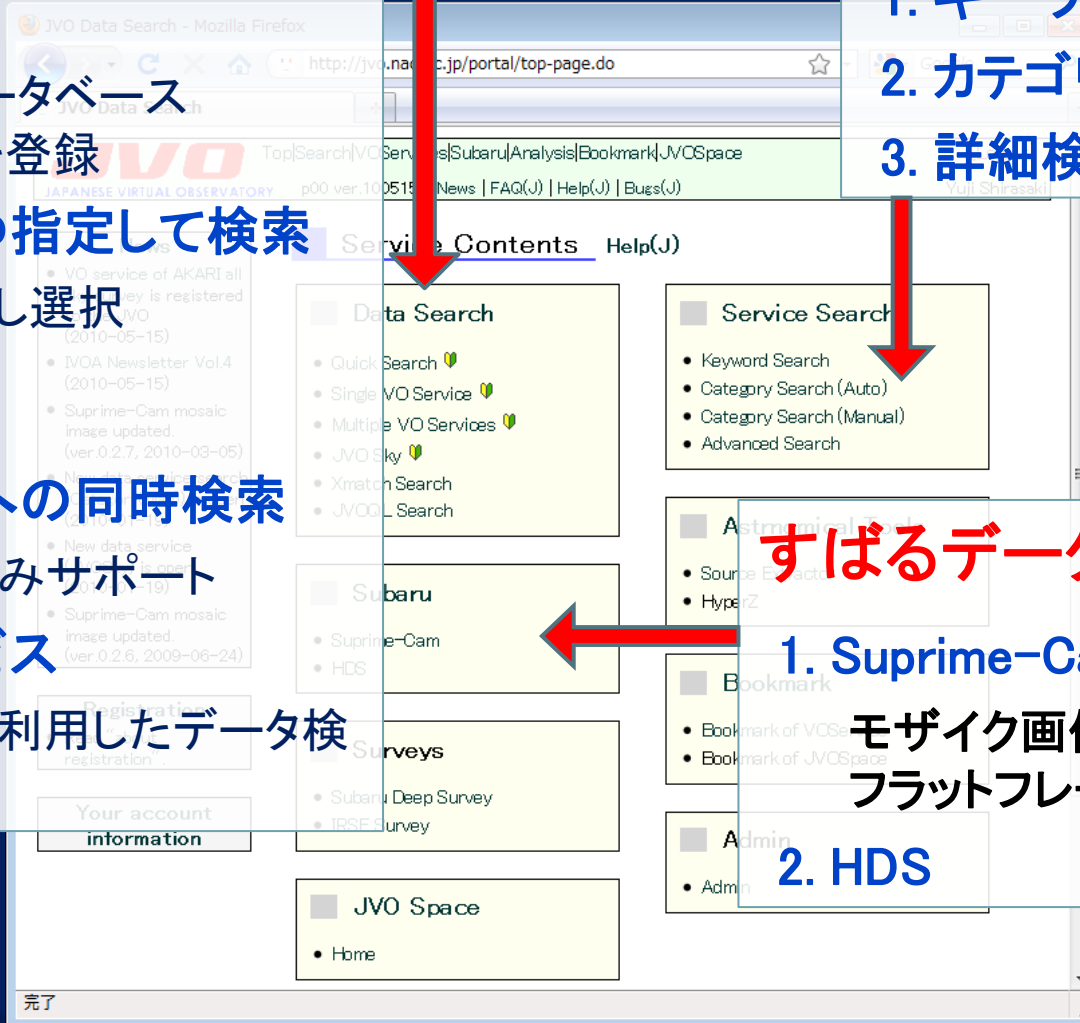
3. 詳細検索

すばるデータ検索・解析

1. Suprime-Cam

モザイク画像作成
フラットフレーム作成

2. HDS



JVO Sky

JVO Sky Help(J)

Object Name Suprime-Cam/Subaru HDS/Subaru Suzaku

Coordinate or ObjectName : Go (ra, dec)=(184.326, 12.899)=(12h17m18.21s, 12d53m54.82s)

OK

suzaku:9/9, hds:3/3, spcam:36/36, searchtime:40ms

Subaru:
SuprimeCAM, HDS

ISAS:
Suzaku

More to come

id : SUPM4A4E5EA95463300
title : 12h24m13+14d08m03 (VIRGO_FIELD_2)
url : Link
type : image
center : 186.05584593 14.13427855
band name : W-C-RC

Goto TOP

Mosaic frame: SUPM550826F900000 (field=0,type=coadd_all,subtype=all,totalExposureTime=29620[sec],totalNum

Download (zipped FITS) Download (uncompressed FITS) 1159.07 MB

Data Archives and Transmission System

DARTS

MAI AKARI SUZAKU HALCA ASCA IRTS GINGA TENMA JUDO

BACK

SUZAKU MASTER / Search Result

The number of display 1 / The number of hits 1

Output ALL hit records : select an output format >>

Download data : select checkboxes below and THEN push >>

CHECK ALL CLEAR ALL

NO	SELECT	FTP	NAME	RA	DEC	LII	BIJ	ROLL_ANGLE	TIME	STOP_TIME	OBSID	EXPOSU
1	<input checked="" type="checkbox"/>	<input type="checkbox"/>	NGC 4406	186.5483	12.9401	279.0885	74.6311	294.6659	2009-06-19 18:14:51	2009-06-22 11:30:14	803043010	10177

Google Map と同様の GUI で視覚的にデータを見つける
複数の装置で観測された領域が一目でわかる

JVO Command (jc) (開発中)

- ✓ コマンドラインから portal に検索・解析ジョブを投入
- ✓ JVOspace (portal 上の Storage) へのアクセス
- ✓ スクリプトから何度も実行することができ便利

Syntax of jc (jvo command):

```
jc <command> [<option>] [<argument>]...
```

Examples:

```
jc search -i <jvoql_file>
```

```
jc registry -k <keyword>
```

```
jc copy2l <source> <destination>
```

```
jc run <program_name> <arguments>
```

Other commands:

```
ls rsync passwd resume suspent abort ps
```

```
union join select
```

Hadoop の試験利用

Digital Universe @ JVO

◎ A big table :

- 主要天体カタログをまとめた 200億レコードテーブル
- 座標、明るさ、観測波長等の基本データのみ
- カタログ間の天体同定は行っていない。
- 単純に観測波長毎の明るさをレコードとして追加
- 狭い領域（半径1度）で検索

◎ カラー情報にもとづく全天検索を可能にしたい

- 同一天体のレコードをグルーピングする必要
- 座標でクロスマッチ
- 分散データ処理 → **Hadoop** を使ってみた

Hadoop

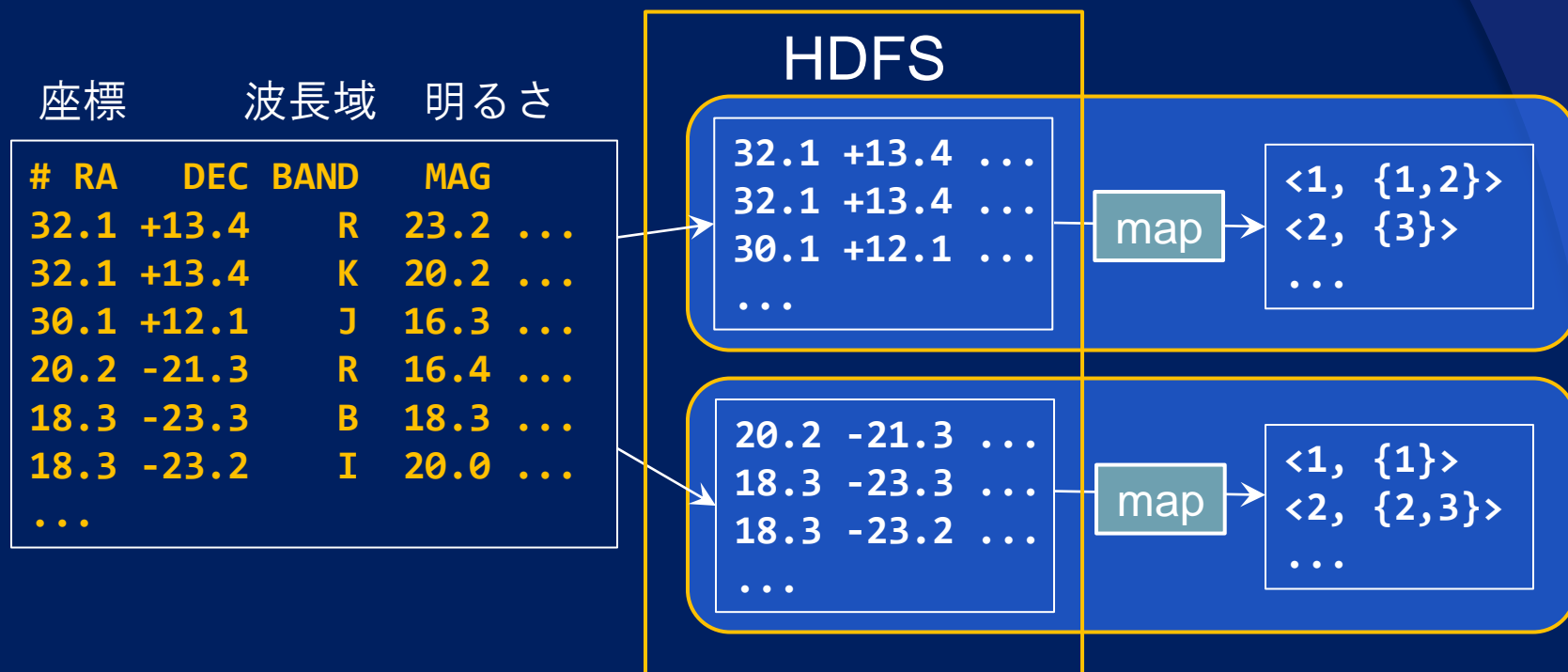
◎ Hadoop とは？

- 分散データ処理のための Java ソフトウェアフレームワーク
- “data is processed where the data resides”
- Apache トッププロジェクトの一つ。
- <http://hadoop.apache.org/>

◎ 適用例

- Facebook : ログ解析, machine learning
- The New York Times: 1000 万の記事を PDF 化
- Yahoo: ランキングの計算
- その他 多数の企業・大学で利用（試用）。

MapReduce の適用



- 全データを天球の領域毎（座標誤差のマージン込）に分割しファイル化する。32768 分割。
- Map 関数は一つのファイル进行处理し、座標でマッチするレコードのリストを出力する
- Reduce は実行しない。各ファイルは独立なので。

実験

- ◎ 10 億レコードで実験 (全体の 1/20)
- ◎ 6112 ファイルに分割。
- ◎ 各ファイルを gzip し (~3MB/file)、HDFS にコピー
- ◎ ハードウェア
 - サーバー 10台 (Quad Xeon 2.5 GHz x 2)
 - 各サーバーは 8 cores & 4 SATA HDDs (RAID5)
- ◎ 同時で実行されるタスクの最大数
 - 1, 40, 70, 100, 160

結果

- ✓ 並列度 1 で実行した場合
 - 10 億レコードの処理に 9 日かかる → 全データ (200億) 処理するのに 180 日 要する。
- ✓ 並列度 70 tasks (コア数とほぼ一致, HDDの数の2倍)
 - 3.7 時間 (10億) → 3 days (200億)

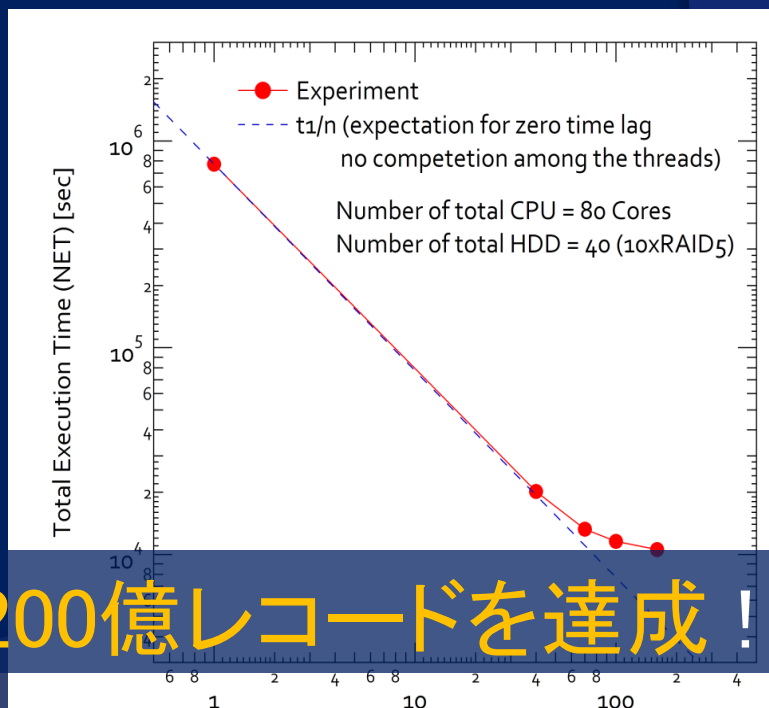
✓ 40並列でさちり始め

- ローカルファイルシステムへの書き込みオーバーヘッド

✓ 所要時間は2週間!

- 導入が簡単、応用が容易

34 nodes 163 並列で **4時間/200億レコードを達成!**



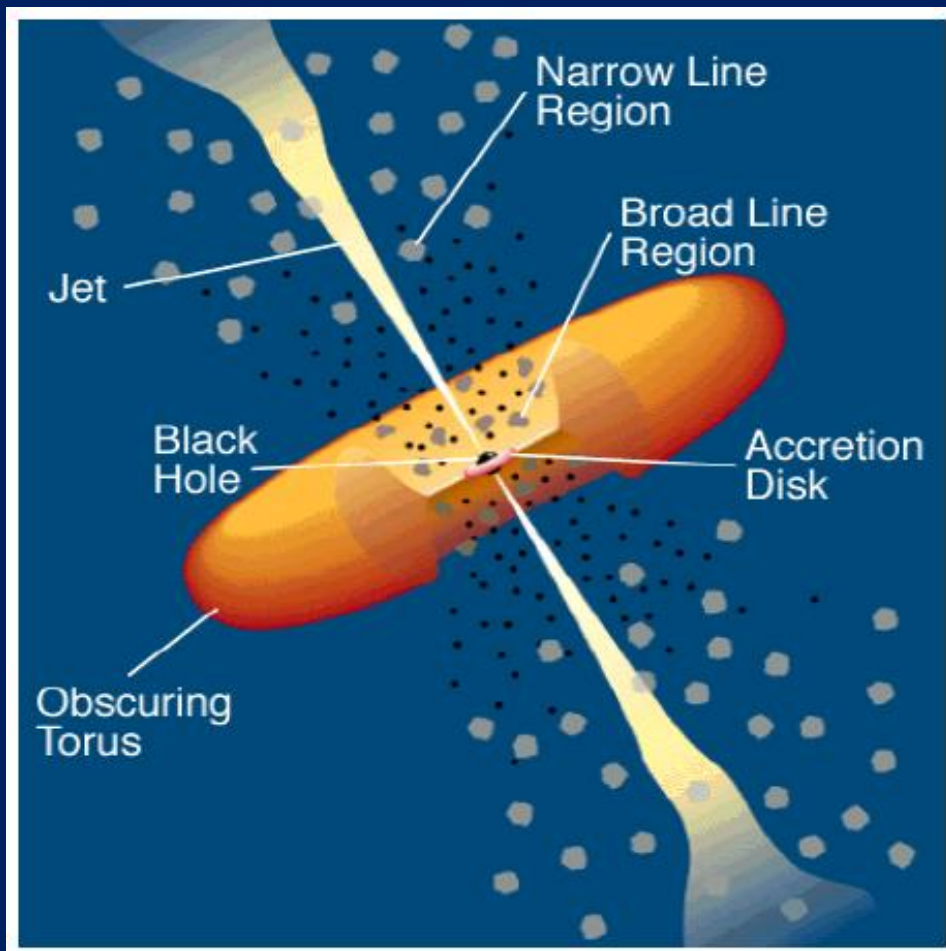
JVO を利用した天文学研究

JVO による研究成果

“Early Science Result from the Japanese Virtual Observatory: AGN and Galaxy Clustering at $z = 0.3$ to 3.0 ” Y. Shirasaki et al. 2011 accepted to PASJ (arXiv:0907.5380v2)

- ✓ AGN (Active Galactic Nucleus) の周りの銀河数密度を**多数のサンプル (~1809)** にもとづき測定。
- ✓ 遠方の AGN ほど、周辺銀河密度が高いことを観測的に確認した。
- ✓ AGN の起源に**銀河同士の衝突・合体**が大きく影響していることを示唆する結果。

AGN 想像図



Credit: NASA

- ✓ AGN (Active Galactic Nucleus) 活動銀河核
- ✓ 銀河中心に存在する巨大ブラックホールに物質降着することにより放射
- ✓ 太陽系程度の広さの領域から銀河一個分に匹敵する放射
- ✓ 中心の巨大ブラックホールはブラックホール同士が合体をくり返し、成長したと考えられている。

AGN 周辺画像データの取得

JVO Top|Search|VOServices|Subaru|Analysis|Bookmark|JVOSpace
JAPANESE VIRTUAL OBSERVATORY p01 ver.100925 News | FAQ(J) | Help(J) | Bugs(J)

=> Location: Top Page > Search > JVOQL Search

Input JVOQL

```
SELECT qso.*, img.*
FROM ivo://jvo/vizier/VII/235:qso_veron_2006 AS qso,
      ivo://jvo/subaru/spcam:image_cutout AS img
WHERE qso.z >= 1.0 and qso.z < 1.1
      AND img.region = Circle(qso.raj2000, qso.dej2000,
```

1. AGNの座標を検索
2. その座標を中心とする半径 0.14 度の領域の画像検索

Submit Generate JVOQL Clear

Service Table Region Criteria Samples

SQL を拡張した JVOQL により分散データベースへの検索を実行

AGN の画像を簡単に検索

JVO Top|Search|VOServices|Subaru|Analysis|Bookmark|JVOSpace
JAPANESE VIRTUAL OBSERVATORY p01 ver.100925 News | FAQ(J) | Help(J) | Bugs(J)

=> Location: Top Page > VOTable Viewer

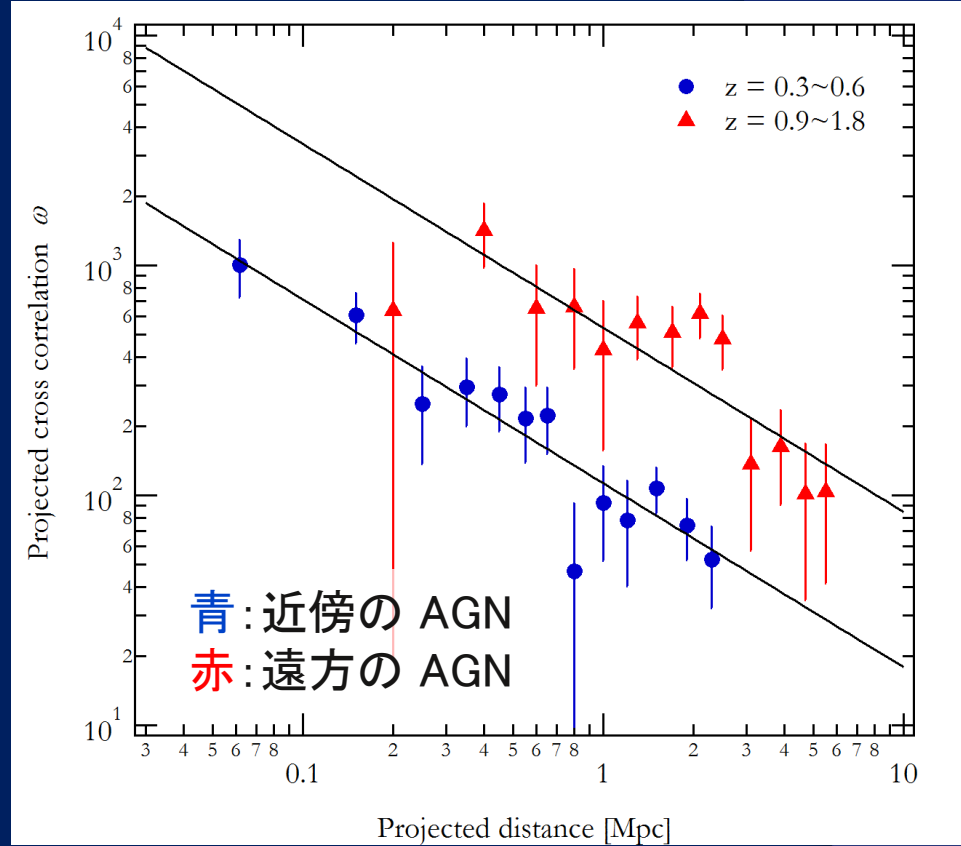
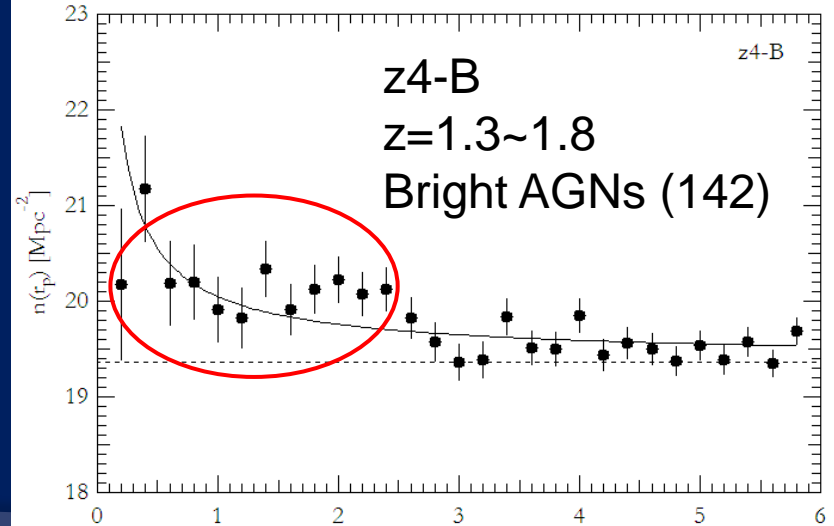
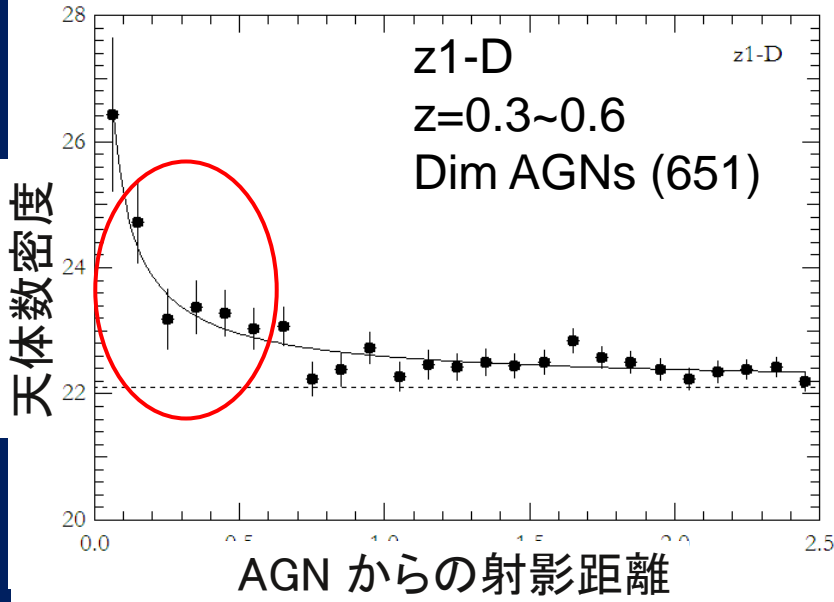
Save/Download Filter Metadata Graphic Add Column Appearance

Total 5390 records page: 1

Alias Name	C42	C31	C43	C23	C30	
check	download	QSO_NAME sort	QSO.RAJ2000 sort	QSO.DEJ2000 sort	IMG.IMAGE_TITLE sort	IMG.ACCESS_REF sort
<input type="checkbox"/>	Download	Q J02399-0134	02 39 56.6	-01 34 27	A370 (W-C-RC)	Link
<input type="checkbox"/>	Download	Q J02399-0134	02 39 56.6	-01 34 27	A370-new (W-S-Z+)	Link
<input type="checkbox"/>	Download	Q J02399-0134	02 39 56.6	-01 34 27	A370-wide (W-S-Z+)	Link
<input type="checkbox"/>	Download	Q J02399-0134	02 39 56.6	-01 34 27	A370 (W-S-Z+)	Link
<input type="checkbox"/>	Download	TEX 2152+172	21 54 39.9	+17 27 39	A2390 (W-S-I+)	Link
<input type="checkbox"/>	Download	SDSS J17110+6400	17 11 05.3	+64 00 14	A2255 (W-C-RC)	Link
<input type="checkbox"/>	Download	SDSS J14022+0308	14 02 14.4	+03 08 12	A1835 (W-S-I+)	Link
<input type="checkbox"/>	Download	SDSS J09570+0238	09 57 01.6	+02 38 57	COSMOS (W-J-B)	Link
<input type="checkbox"/>	Download	SDSS J09589+0213	09 58 57.3	+02 13 14	COSMOS (W-J-B)	Link
<input type="checkbox"/>	Download	SDSS J09597+0247	09 59 46.0	+02 47 43	COSMOS (W-J-B)	Link
check	download	QSO_NAME	QSO.RAJ2000	QSO.DEJ2000	IMG.IMAGE_TITLE	IMG.ACCESS_REF
<input type="checkbox"/>	Download	SDSS J09567+0205	09 56 42.3	+02 05 53	COSMOS (W-S-Z+)	Link
<input type="checkbox"/>	Download	SDSS J09589+0213	09 58 57.3	+02 13 14	COSMOS (W-S-Z+)	Link
<input type="checkbox"/>	Download	2QZ J095958+0108	09 59 58.2	+01 08 47	COSMOS (W-S-Z+)	Link
<input type="checkbox"/>	Download	SDSS J09589+0213	09 58 57.3	+02 13 14	COSMOS (W-S-Z+)	Link
<input type="checkbox"/>	Download	SDSS J09589+0213	09 58 57.3	+02 13 14	COSMOS (W-J-V)	Link

解析結果

$$\omega(r_p) = \frac{n(r_p) - n_{bg}}{\rho_0}$$



遠方における AGN 周辺銀河の密度超過の増加を世界初検出！

まとめ

- ◎ 天文学研究が直面している問題（データ面で）
 - 利用価値の高いデータアーカイブが多数
 - 観測データサイズが大きくダウンロード困難
 - 並列処理技術が高速化には不可欠
- ◎ 私たちグループ (JVO) の解決策
 - データアーカイブの公開方法の標準化 (IVOA)
 - データアーカイブと解析システムをセットで開発
 - ユーザはネットワーク越しに遠隔解析システムを使う（クラウド！）
 - こうしたクラウド環境とローカル解析環境のシームレスな連携を実現したい。