
NGVS

CADC and CANFAR:

A platform to support data-intensive science

OR

Radical evolution in the core business of the CADC

David Schade
Canadian Astronomy Data Centre
Canadian Advanced Network for Astronomy Research

TMT data management

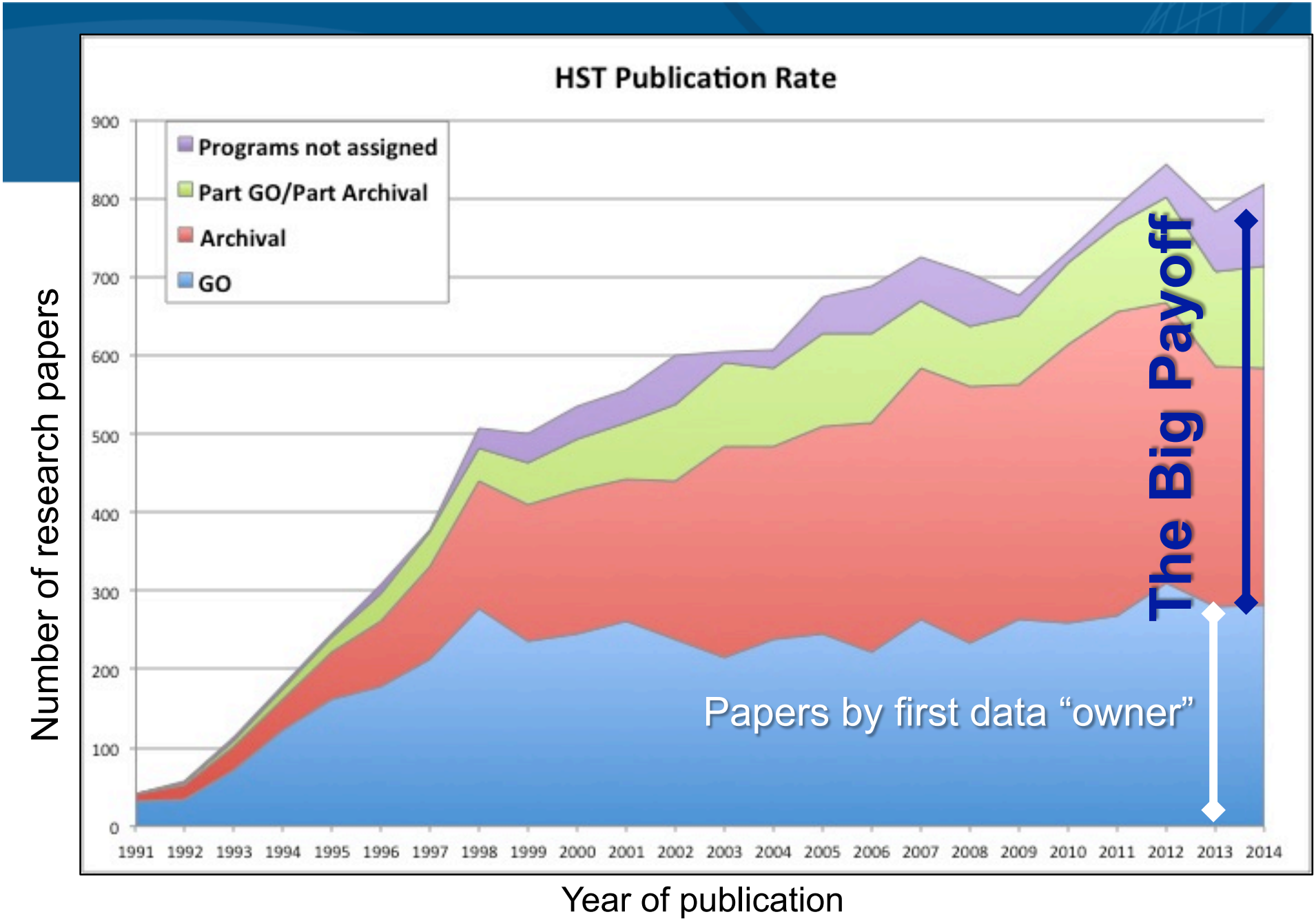
At this meeting I've heard a lot of things that I agree with:

- Science productivity is increased by good data management
- An archive-centric view of TMT from the beginning will save effort later
- Build upon existing expertise, software, infrastructure
- Key Projects are coupled to pipelines and are coupled to good data management

I will not be preaching to the choir

BUT:

- Not all data have equal potential for archival research
- Data management work needs to be subject to cost-benefit analysis
- Be selective


HST Publication Rate
Programs not assigned
Part GO/Part Archival
Archival
GO
900
800
700
600
500
400
300
200
100
0
1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
The Big Payoff
Papers by first data "owner"
Number of research papers
Year of publication

TMT data management

The “Core Business” of an observatory is to produce data

- Data Management choices for TMT will be driven by the science that the TMT communities choose to do

SDSS and LSST are not traditional observatories

"Massive Survey Data Collections"?

Cities versus Villages

- Facilities like SDSS/LSST are in the business of constructing entire cities with all of the infrastructure needed to make them run
- Cities support a broad range of cultural and economic activities
- Cities are dynamic and connected strongly to the broader world

Farms or villages

- Have a single purpose (produce milk, dig the coal)
- Can subsist with a hand pump and an outhouse

Culture: Canadian astronomy

Canadian astronomy builds data “cities”

- Large collaborative teams
- Pool telescope time to support large programs
- International collaboration
- Need good data management

Canadian astronomy is highly effective

- Citation rate slightly higher than the US (2005)
- Consistently ranked among the best in the world

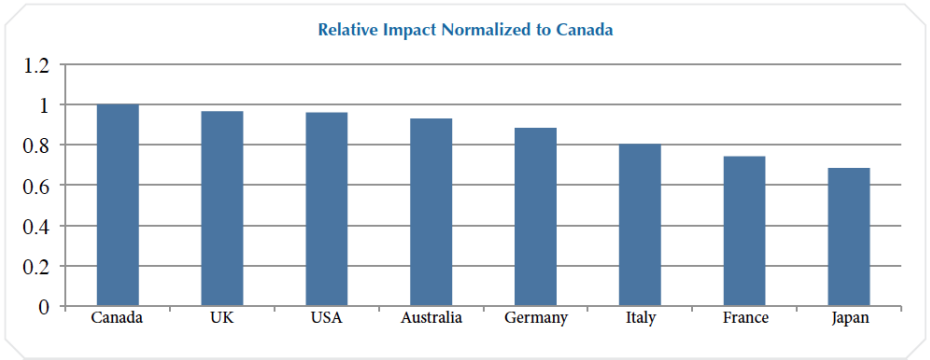


Figure 2-11. Relative impact of the astronomy communities (G7 countries plus Australia) as measured by CIF statistics 2005-2009.

The CADC view of data-intensive science

My bias

- Large homogeneous samples with well-defined statistical completeness
- Multi-wavelength, multi-facility data collections
- Massive surveys
- Homogeneous data processing
- Good data discovery and access tools

Integration of astronomy resources world-wide

- A philosophy driven by HST,CFHT,JCMT
- A philosophy driven by science practice

TMT data management

The “Core Business” of an observatory is to produce data

TMT data management

The “Core Business” of an observatory is to produce data

BUT ……..

TMT data management

The “Core Business” of an observatory is to produce data

“Core Business” includes a coherent and effective approach to:

- Metadata
- Calibration data
- Association of data
- Pipeline processing (at some acceptable level)

TMT will not be an island

TMT needs to be integrated into the broader science world

TMT is not Big Data but will live an a Big Data ecosystem

TMT data management

Don’t re-invent

- There exists a large body of experience and good practice in the astronomy data management community
- There is an obligation for TMT to fit into the global science data management world
- There is also a benefit and cost-savings to using established approaches

- Example:
- Common Archive Observation Model (CADC, STScI, other NASA centers, ALMA?)
- International Virtual Observatory Alliance (IVOA) standards

Common Archive Observation Model (CAOM)

- Inspired by International Virtual Observatory Alliance work
- Integrates all instruments into a common data model
- All downstream software is unified
- Efficiency and cost savings
- For science (not engineering)
- IVOA standards supported

CAOM supports all VO protocols and supports a unified single interface to all CADC collections

- 22 collections
- 113 instruments
- 4 levels of calibration

CANFAR (Canadian Advanced Network for Astronomy Research)

- A consortium of 15 university astronomers who form the Science Management Committee
- CADC is part of CANFAR and responds to CANFAR management
- CANFAR is a new model for science input into the development process for data management
- CANFAR is driving a move of all CADC data centre hardware assets onto Compute Canada (the national infrastructure for Advanced Computing)
- CANFAR is driving a radical change in CADC's core business

CANFAR is a new model for science input

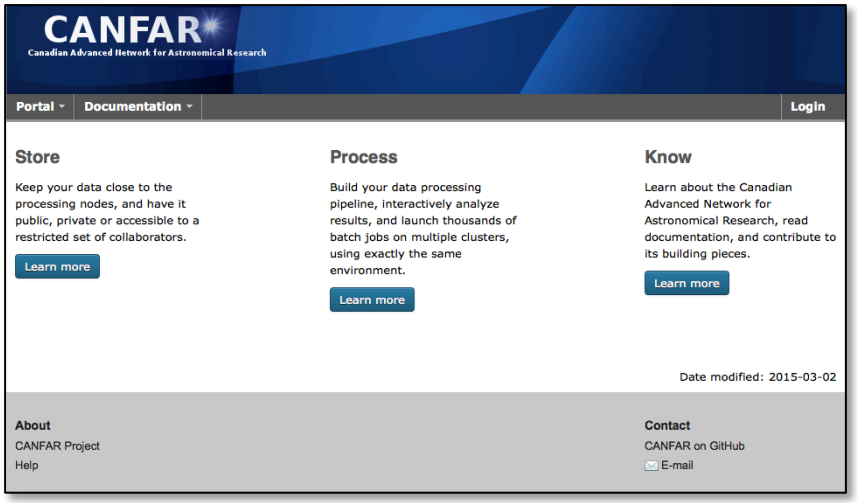
- CANFAR partners with Survey Science Teams to iteratively develop the capabilities that they need
- CANFAR funds post-docs that are doing the science
- CANFAR funds software developers and support staff to work with science teams

We believe:

- Survey science is of primary importance
- Survey science presents the sharpest challenges
- Input from a broad range of surveys will yield a “general” result
- Small project astronomers will be served adequately

Canadian Advanced Network for Astronomical Research

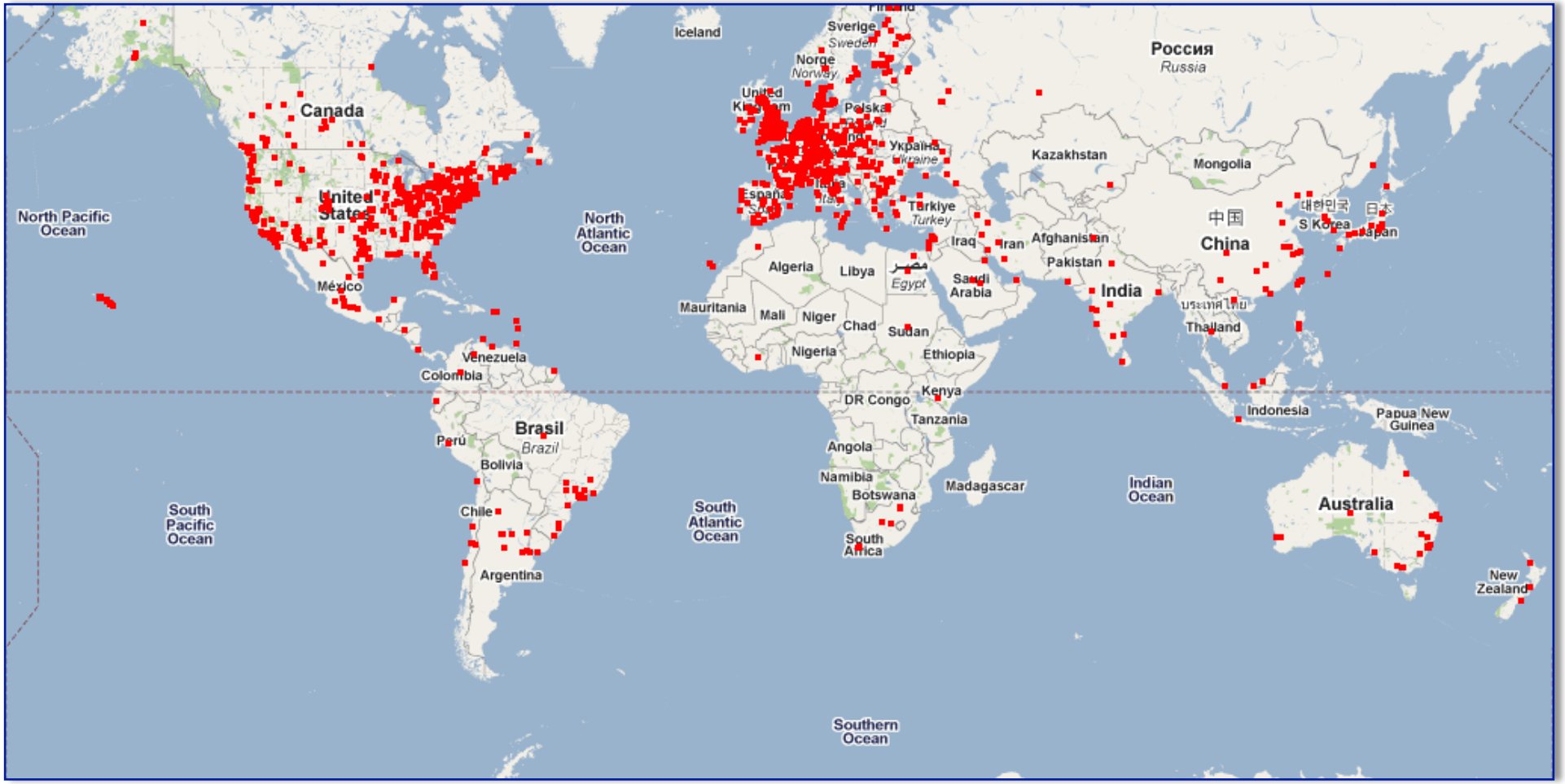
- A cloud ecosystem for data intensive astronomy
- User services
  - Store and share data
  - Create and configure VMs
  - Run interactive VMs
  - Run persistent VMs
  - Batch processing with VMs
- Using research cloud resources
  - Compute Canada
  - CADC
- Integrated authentication and authorization



CANFAR/CADC 2014

- Size:
  - 932M files
  - 2.3 PiB
- Users
  - Authenticated access: 762
  - Anonymous access: 7,544
  - Registered: 7,018
- Data handled in the last year
  - TiB: 1,106
  - Files: 91M

CANFAR/CADC data delivery

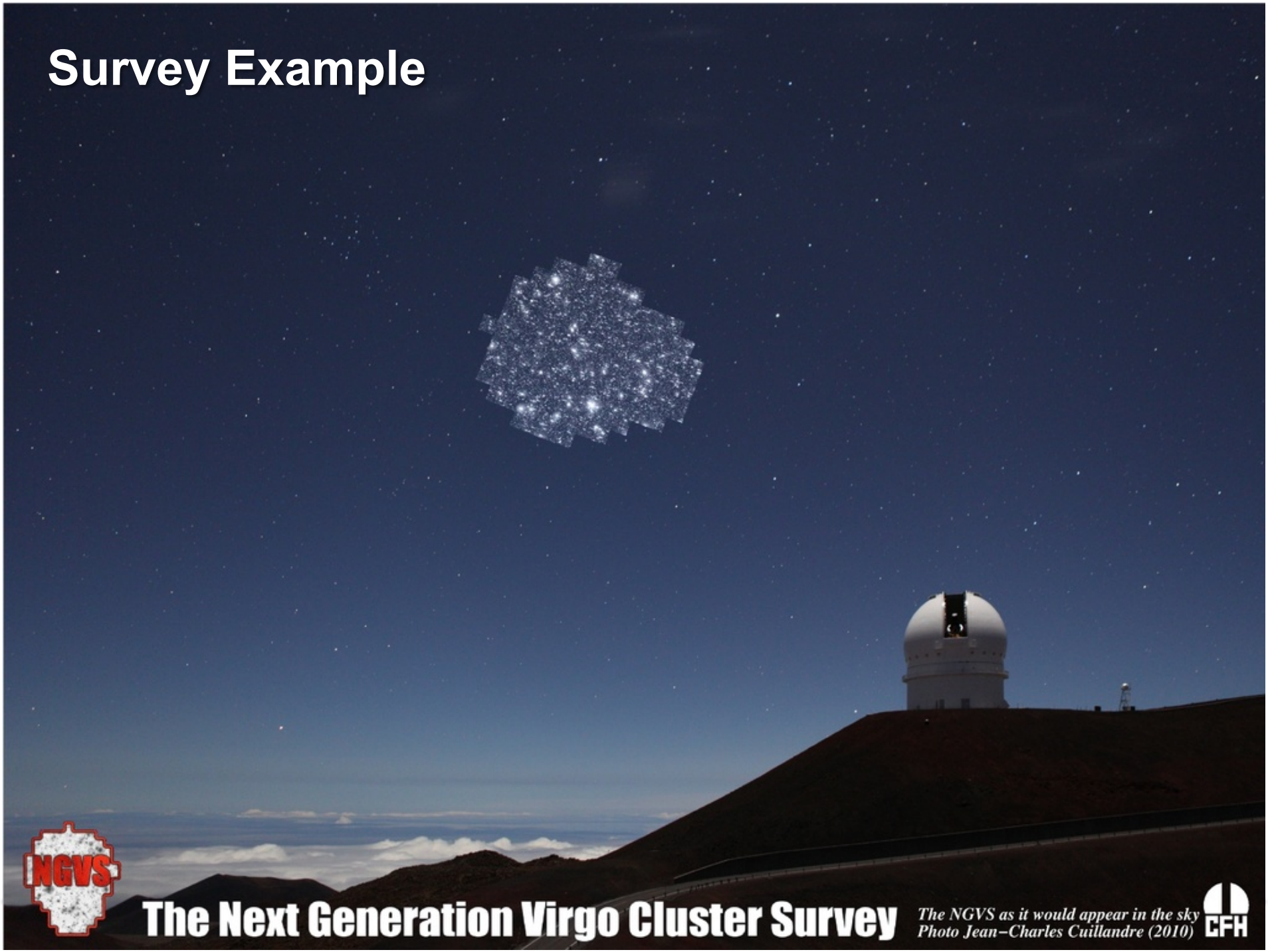


CANFAR Services

- **Store and share data**
  - (VOSpace = DropBox)
- **Create and configure VMs**
  - Scientists replicate their personal environments
- **Run interactive VMs**
  - Access powerful compute environments from laptops
- **Batch processing with VMs**
  - Scalable processing (1000's of cores, large RAM)
- **Run persistent VMs**
  - Hosting Database, SaaS, services

- **Integrated authentication and authorization**

Survey Example



Survey Example

200 nights of CFHT time
103 sq. deg

- **VOSpace is 40 TB**
  - **Optical & IR imaging,spectroscopy, databases, advanced data products**
- **50 users in 6 countries**
- **VOSpace Network traffic ~ 1 PB --- 10 million files**
- **Many 1000's of core-years of batch processing**
  - **Cloud used for everything**
- **VM's:** python, iraf, fortran, perl, ds9, SuperMongo, Galfit, Gnu Data Language (GDL), SourceExtractor, THELI, Le Phare, galapagos, swarp

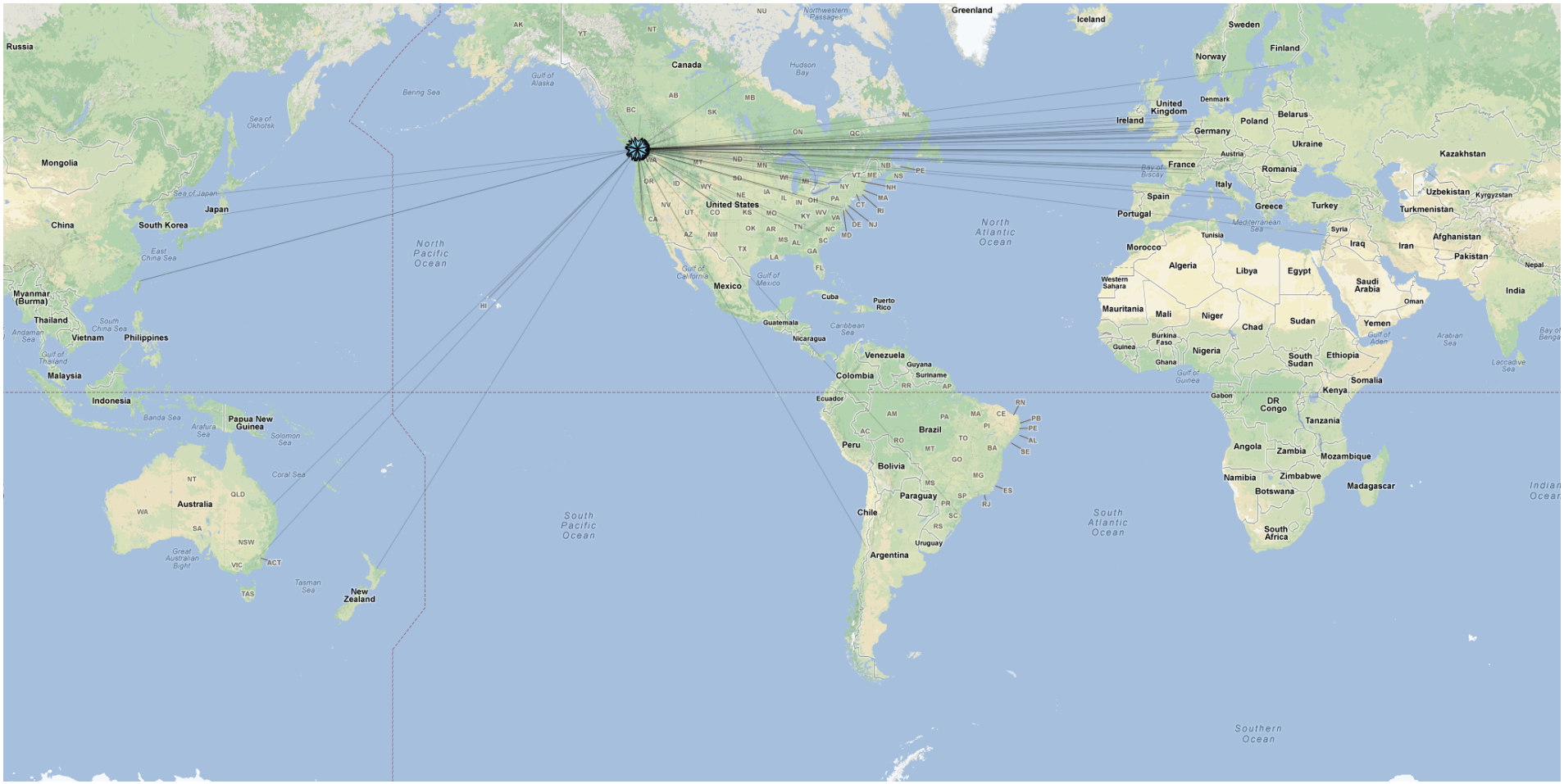
NGVS

The Next Generation Virgo Cluster Survey

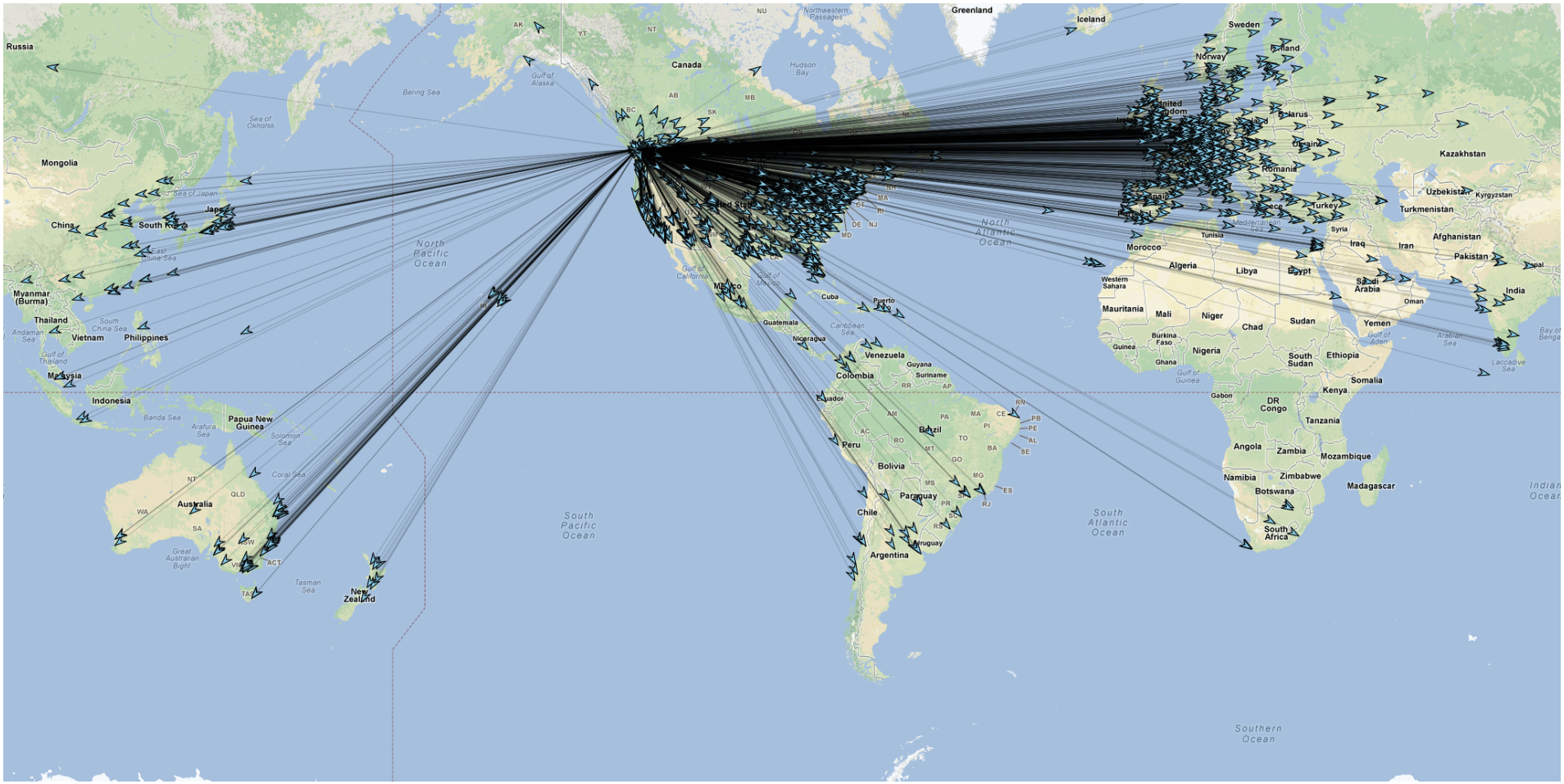
The NGVS as it would appear in the sky
Photo Jean-Charles Cuillandre (2010)

CFH

Geography of VOSpace PUTs

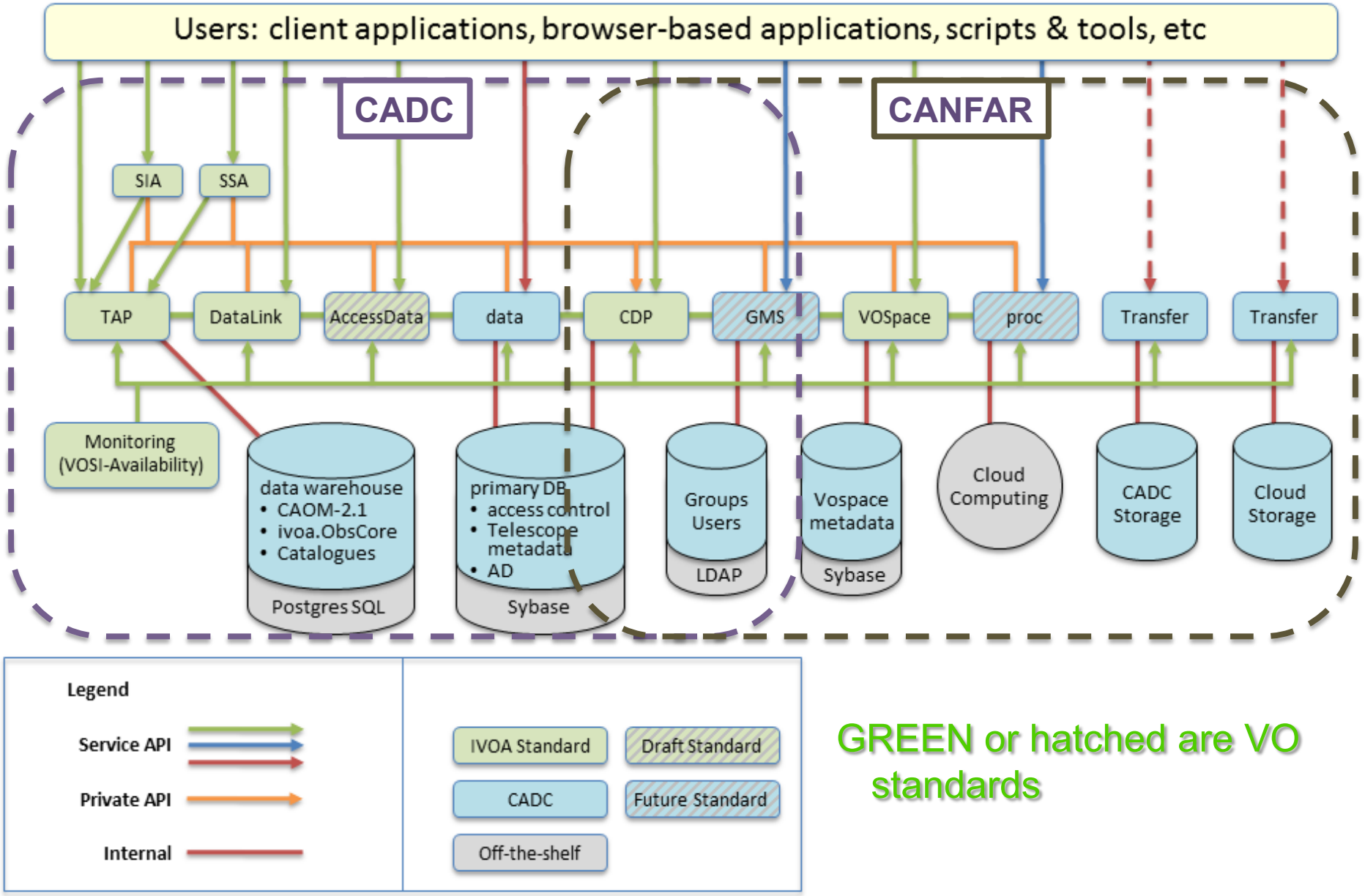


Geography of VOSpace GETs



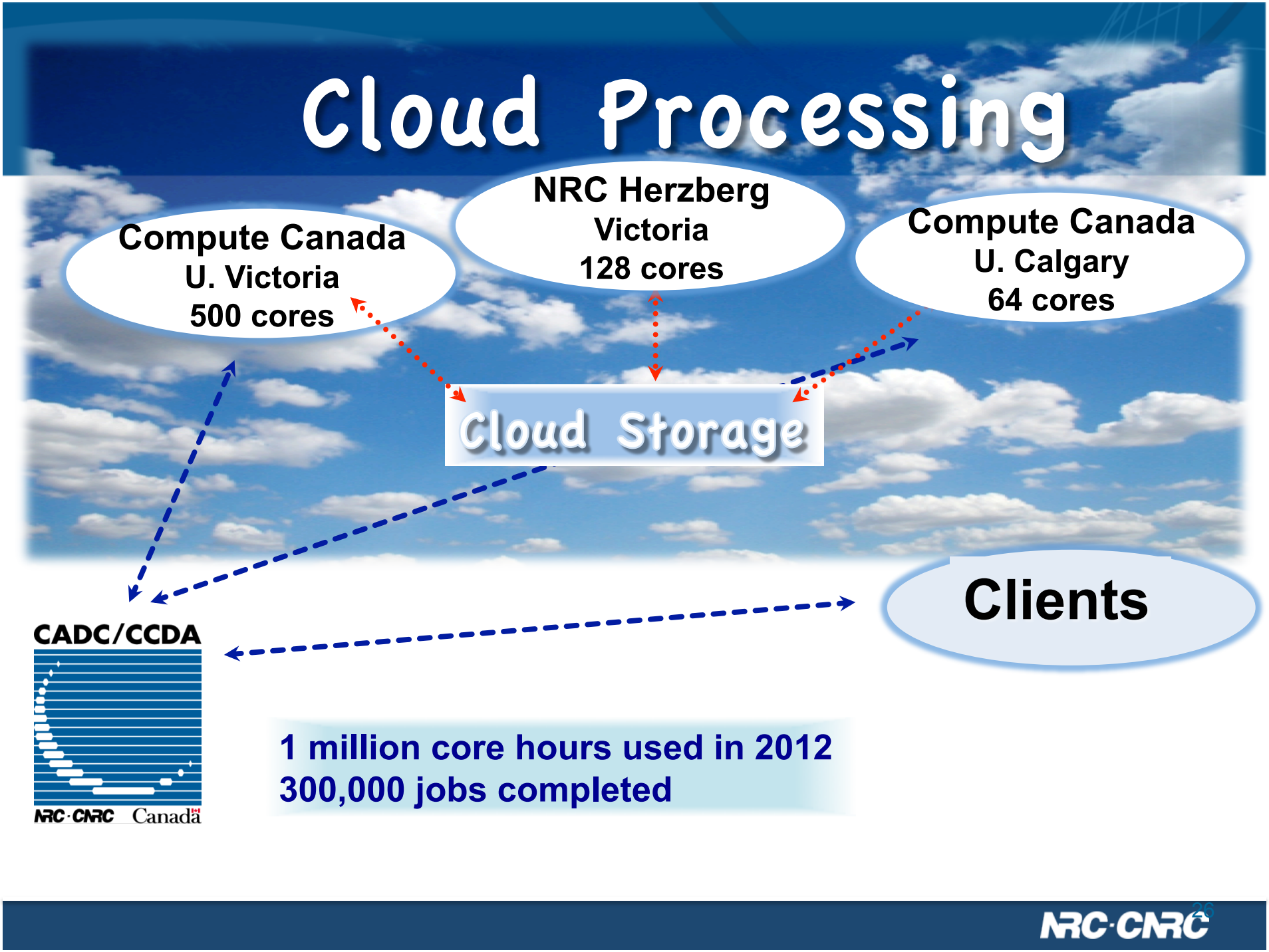
CANFAR integrates Virtual Observatory standards

- CADC is a leader in the International Virtual Observatory Alliance
- CADC/CANFAR has a higher degree of VO compliance than any other data center
- CADC/CANFAR has integrated VO into its internal architecture



GREEN or hatched are VO standards

Cloud Processing



1 million core hours used in 2012
300,000 jobs completed

CANFAR is a distributed cloud system

European Grid Initiative: CANFAR/INAF/EGI

PROPOSAL – Technical Annex

Sections 1-3: Excellence, Impact & Implementation

Proposal full title: Engaging the Research Community towards an Open Science Commons

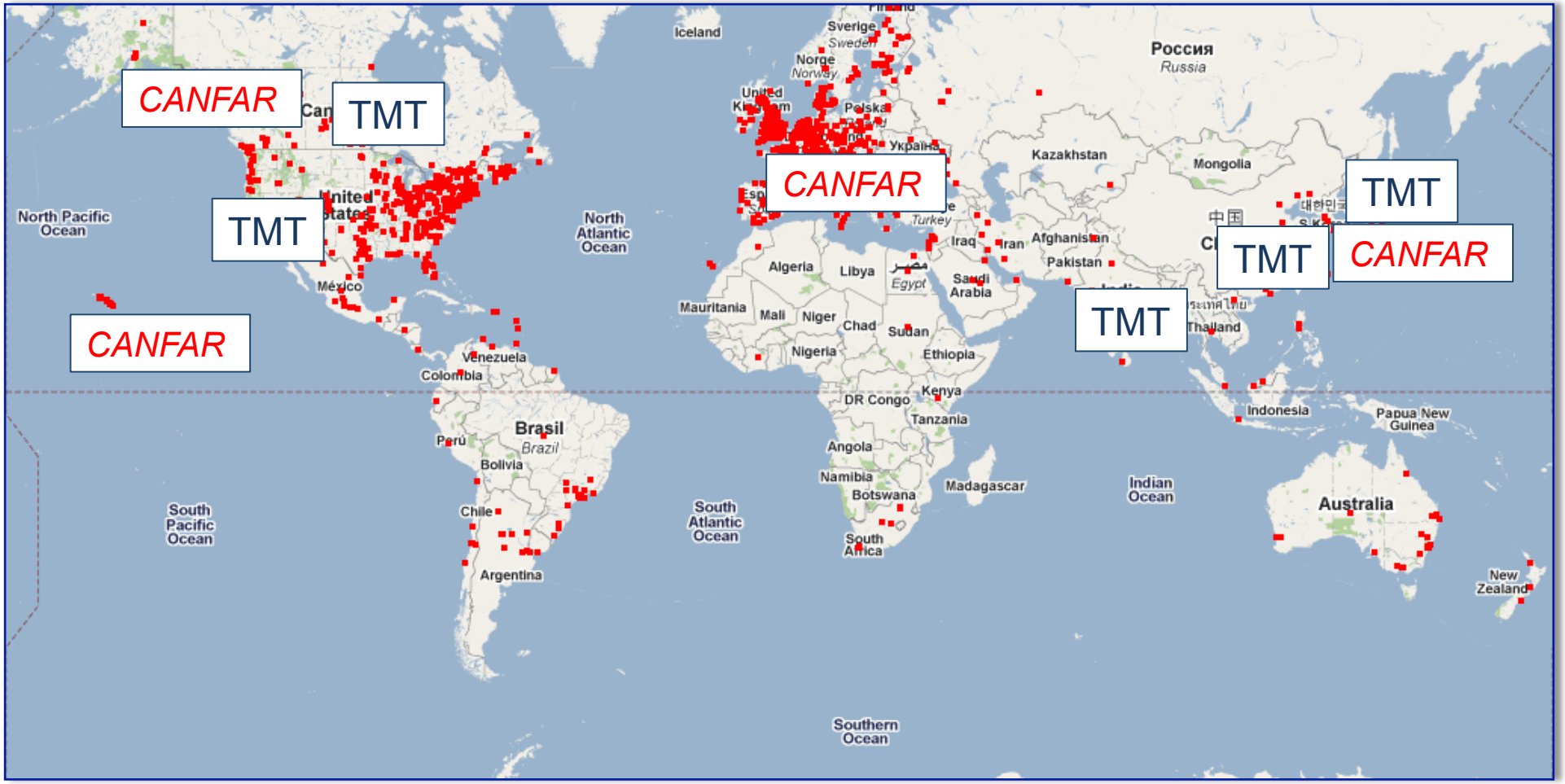
Proposal acronym: EGI-Engage

Canadian Advanced Network for Astronomical Research (Lead: INFN) (M6 – M30)

The Canadian Advanced Network for Astronomical Research (CANFAR)[20] is a computing infrastructure for astronomers in Canada. International collaboration in the Astronomy discipline will be supported both by the Canadian Astronomy Data Centre (CADC) and EGI. CANFAR and EGI will work together to integrate both e-Infrastructures towards a seamless and uniform platform for international astronomy research collaboration. Community services will be provided on top of the federated cloud of EGI using open source solutions and re-using the CANFAR experience.

TMT is distributed: cloud makes sense

Supports international collaboration



Funding for Research Data Management (cyber-infrastructure) in Canada & Europe (US?)

- Research Data Management is a highly visible national priority
- ***Generic*** shared cyber-infrastructure is the fashion
- Research Data Management will not be funded within domain silos like astronomy
- CANFAR is, in part, a response to this new reality
- There is a place for shared generic infrastructure
- There is also a place for highly-efficient special-purpose infrastructure

Scalability

- Scalability is the primary motivation for using cloud
- If CADC/CANFAR could get scalability without using shared infrastructure …….

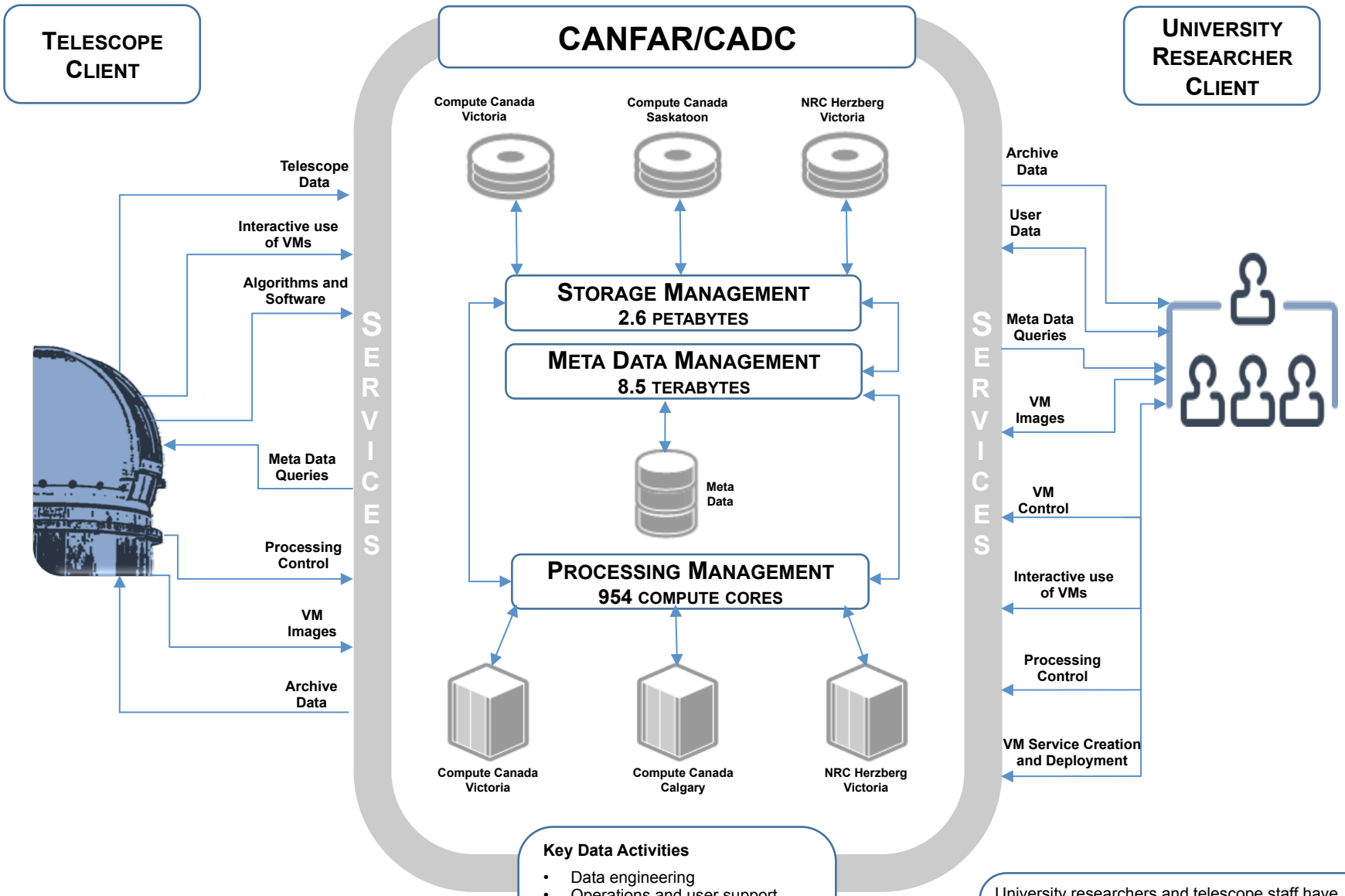
What is the relevance to TMT?

- A broad range of options exist for TMT Data Management
- TMT needs to properly manage the telescope side of the data management system
- TMT should leverage established practices to reduce development effort (CAOM,VO)
- TMT should outsource its data management needs to an appropriate partner
- Outsourcing is more effective and less expensive than in-house
- The science community should make major contributions to pipeline processing
- Cloud infrastructure can be used very effectively for pipelines, databases, and many other applications
- There may be funding from non-astronomy sources for data management

CADC's core business has changed

We have evolved to better serve our science community

- Not only have we evolved a new model to support data-intensive science
- We have evolved a model of interaction with the most technologically progressive elements of our science community that will sustain ongoing evolution in the right direction
- The present challenge is to position software and data in the right way to support massive processing, visualization, and analytics
- The looming challenge is to create a global network of "islands of capability" that are dynamically linked to support a new "meta-analytics"



	Data In		Data Out	
	# of files	Terabytes	# of files	Terabytes
Peak per day	2,169,190	8.0	648,093	16.8
Avg per day	130,952	0.4	99,253	2.6

Key Data Activities

- Data engineering
- Operations and user support
- Software development
- Software integration
- Data processing
- Data management
- User web services
- User web interfaces

University researchers and telescope staff have privileges to upload data, create VMs and install science applications, run interactive VM sessions, submit batch processing jobs to VMs, share their VMs, control the life-cycle for their VMs, offer software-as-a-service applications in their VMs.

Definition: VM – Virtual Machine