CyberSKA

Cyberinfrastructure for Radio Astronomy

(http://www.cyberska.org)

Cameron Kiddle
Research Fellow, Grid Research Centre, University of Calgary
Technical Coordinator, CyberSKA

What is CyberSKA?

- Initiative to develop a scalable and distributed cyberinfrastructure platform to meet evolving science needs of the SKA
- Led by the University of Calgary with several partner institutions from North America currently
- Canadian funding for CyberSKA provide by CANARIE, as part of their Network Enabled Platforms (NEP) program, and Cybera
- Starting by establishing cyberinfrastructure to support current large-scale astrophysical data needs generated by GALFACTS, PALFA and other high data volume SKA Pathfinder projects

Project Team

CyberSKA

University of Calgary

- Russ Taylor (Professor, Lead PI)
- Eric Donovan (Associate Professor)
- Robert A. Este (Project Manager)
- Cameron Kiddle (Technical Coordinator)
- Mircea Andrecut (Developer)
- Roger Curry (Developer - Grid Research Centre)
- Pavol Federl (Developer)
- Arne Grimstrup (Developer)
- Sukhpreet Guram (PhD Student)
- Paolo Pragides (Developer)
- Dina Said (PhD Student)
- Christian Smith (System Administrator)
- Tingxi Tan (Developer – Grid Research Centre)

McGill University

- Victoria Kaspi (Professor)
- Rafal Klodzinski (Developer – Sequence Factory)
- Patrick Lazarus (MSc Student)
- Shibl Mourad (President – Sequence Factory)
- Alex Samoilov (Developer – Sequence Factory)

University of British Columbia

- Ingrid Stairs (Associate Professor)
- Bryan Fong (Developer)
- Mark Tan (Developer)

University of British Columbia, Okanagan

- Erik Rosolowsky (Assistant Professor)
- Venkat Mahadevan (Developer)

Cornell University

- Jim Cordes (Professor)
- Adam Brazier (Research Associate)
- Shami Chatterjee (Research Associate)
- Eric Chen (Analyst Consultant)

IBM Canada

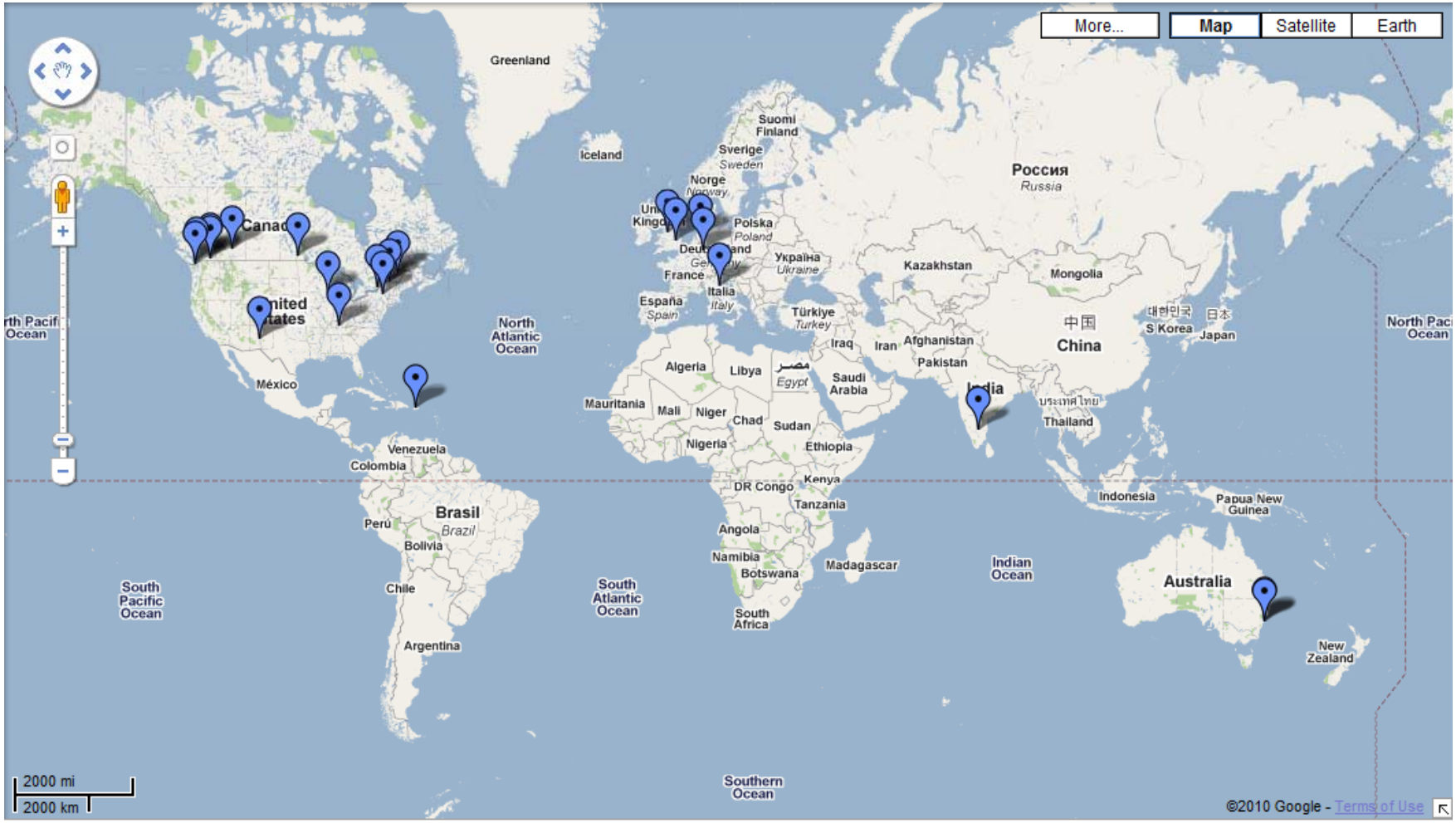
- Don Aldridge (General Manager, Research & Life Sciences)
- Olivier Eymere (IT Architect)

National Research Council Canada

- Tom Landecker (Principal Research Officer)
- Tony Willis (Senior Research Council Officer)

Current Usage

- ~100 members from around the world
- 10+ groups (GALFACTS, PALFA, EVLA, GMRT, CASA Users…)



Requirements

- **Distributed**
  - provide access to distributed data, computing resources and services
- **Scalable**
  - must be able to scale to support increasing data and processing needs
- **Deployable**
  - different sites should be able to deploy developed tools and participate in CyberSKA relatively easily
- **Heterogenous**
  - provide a framework to enable interaction with different types of data, computing resources and services and to add/execute different processing algorithms and workflows
- **Automated**
  - Automation and dynamic reconfiguration of services and data workflows in response to user demand, changing user objectives, available data and resource availability

Requirements - II

- **Transparent**
  - provide users with transparent access to data, computing resources and services
- **Web-enabled**
  - a Web-based platform that users can access from anywhere with Internet access
- **Collaborative**
  - enable international/distributed teams to collaborate and communicate effectively
- **Interactive**
  - enable on-line interactive visualization of data
- **Auditable**
  - be able to track where data has come from and processes applied to it (data provenance)
- **Interoperable**
  - compliant with existing standards such as Virtual Observatory (VO)

Experience/Background

CyberSKA

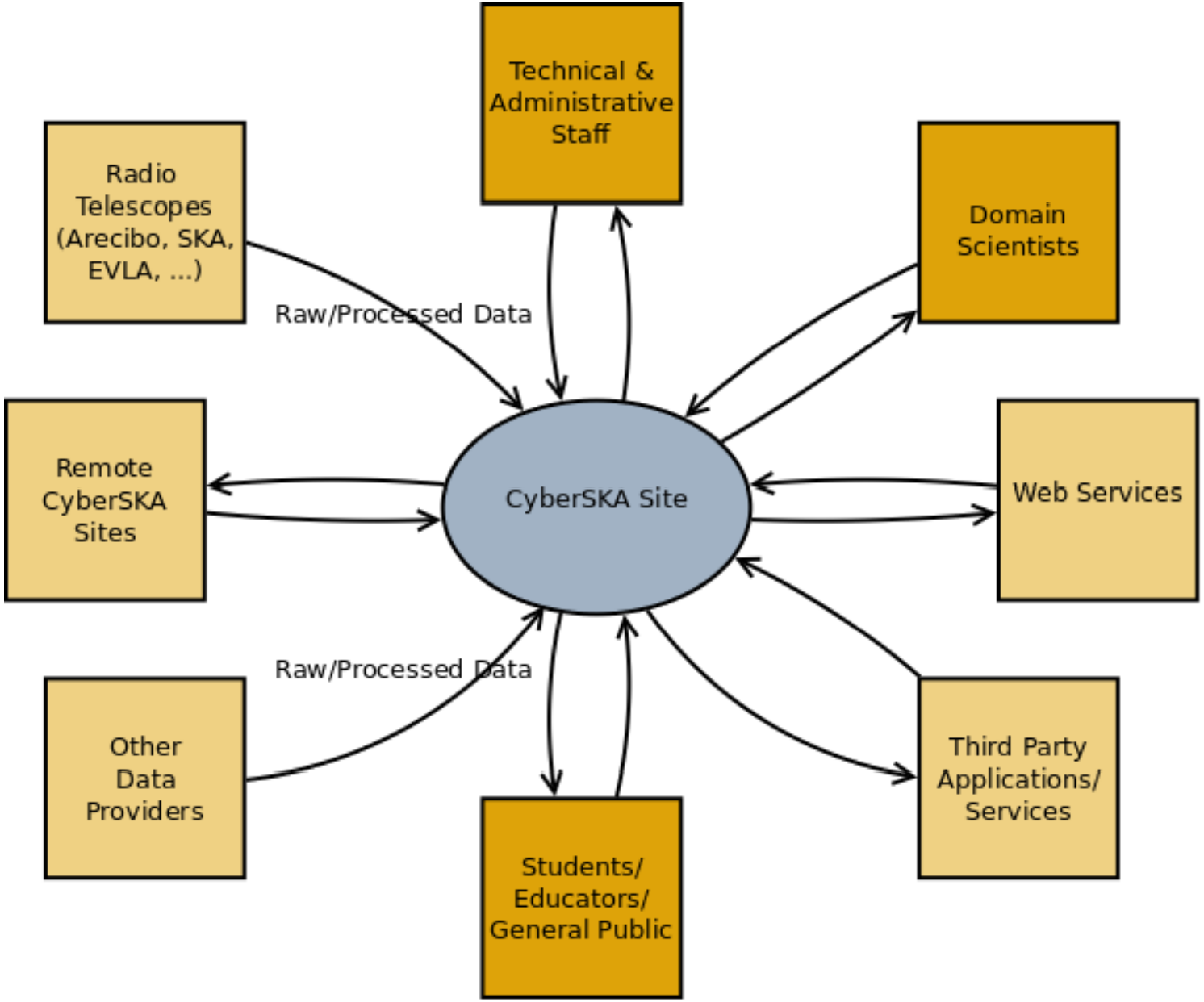
- Leveraging knowledge and experience of the Grid Research Centre at the University of Calgary, IBM, and a large technical team
- Adapting, customizing and extending technologies used by GeoChronos (http://geochronos.org)
  - a platform developed by the Grid Research Centre
  - enables Earth observation scientists to access and share data and applications and collaborate more effectively
  - employs social networking, cloud computing and data management technologies
- Making use of other existing tools and technologies where possible

Motivation

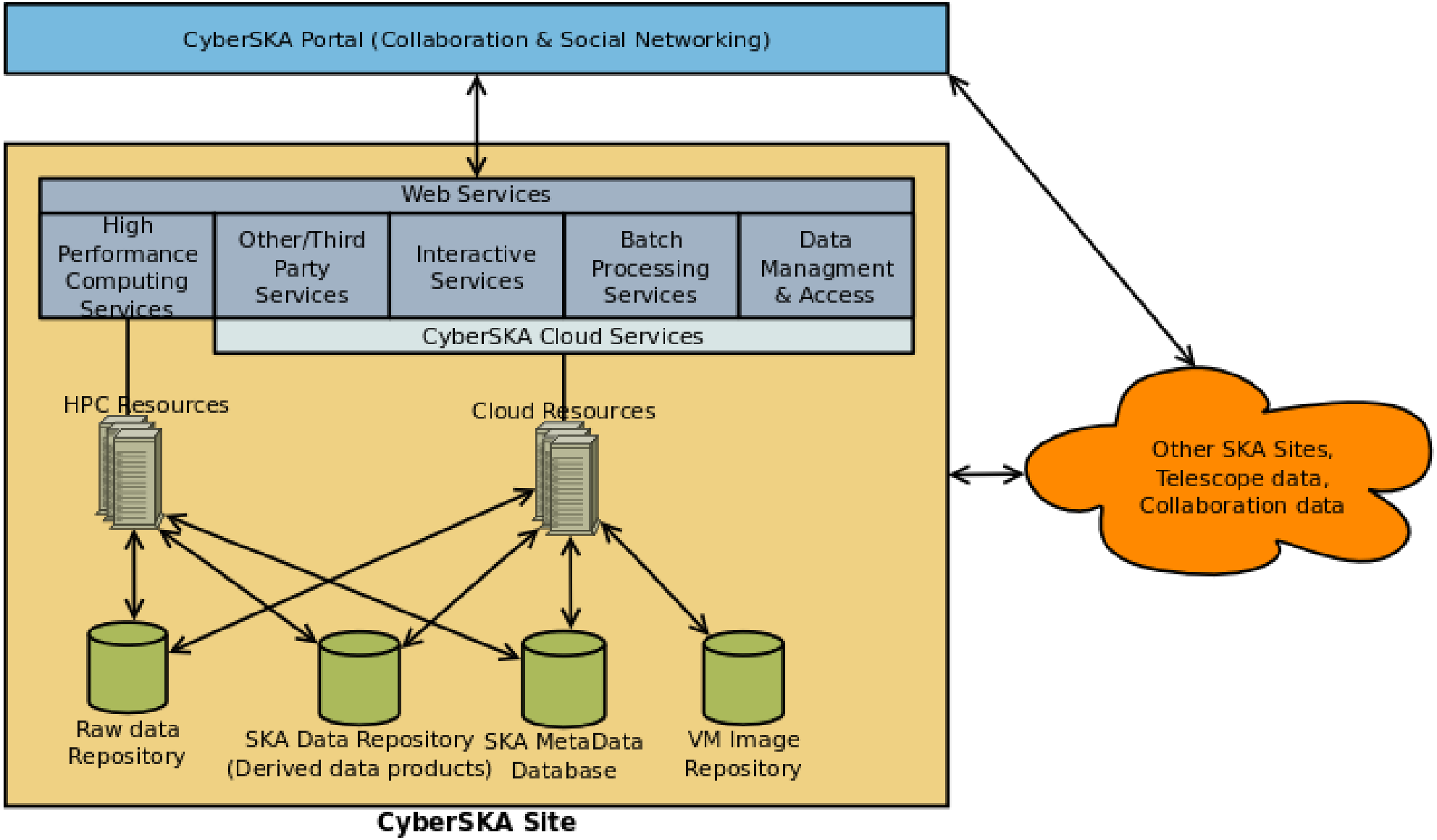
- **Why social networking?**
  - can enhance collaboration capabilities around data and applications - "*Facebook for Scientists*"

- **Facebook analogy**
  - a platform dealing with large scale in terms of users, data and applications
    - > 500 million users (50% log on to Facebook on any given day)
    - > 30 billion pieces of content shared each month
    - > 550 thousand applications on Facebook Platform

System Context Model

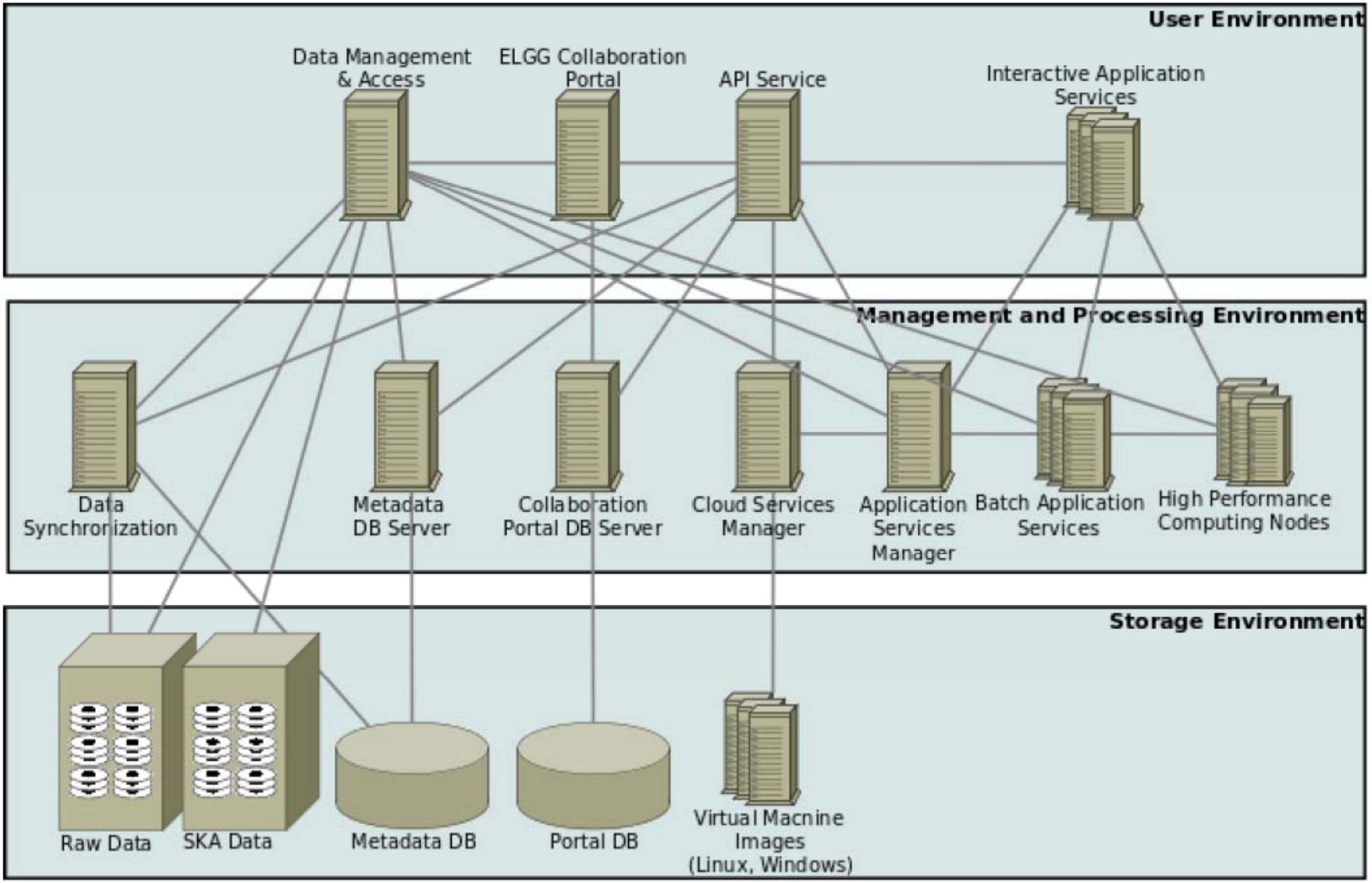


High Level Architecture



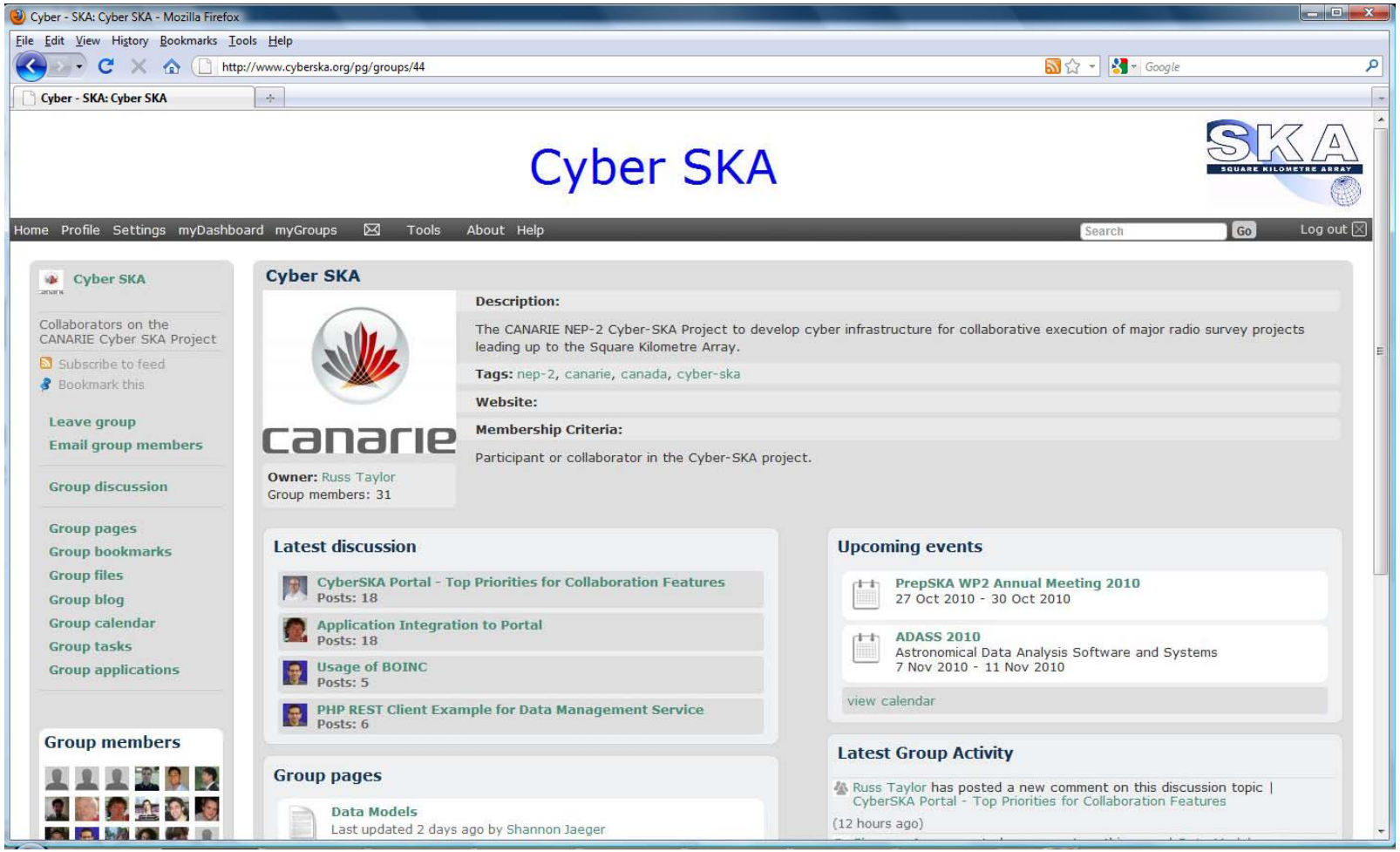
Logical Operational Model

CyberSKA



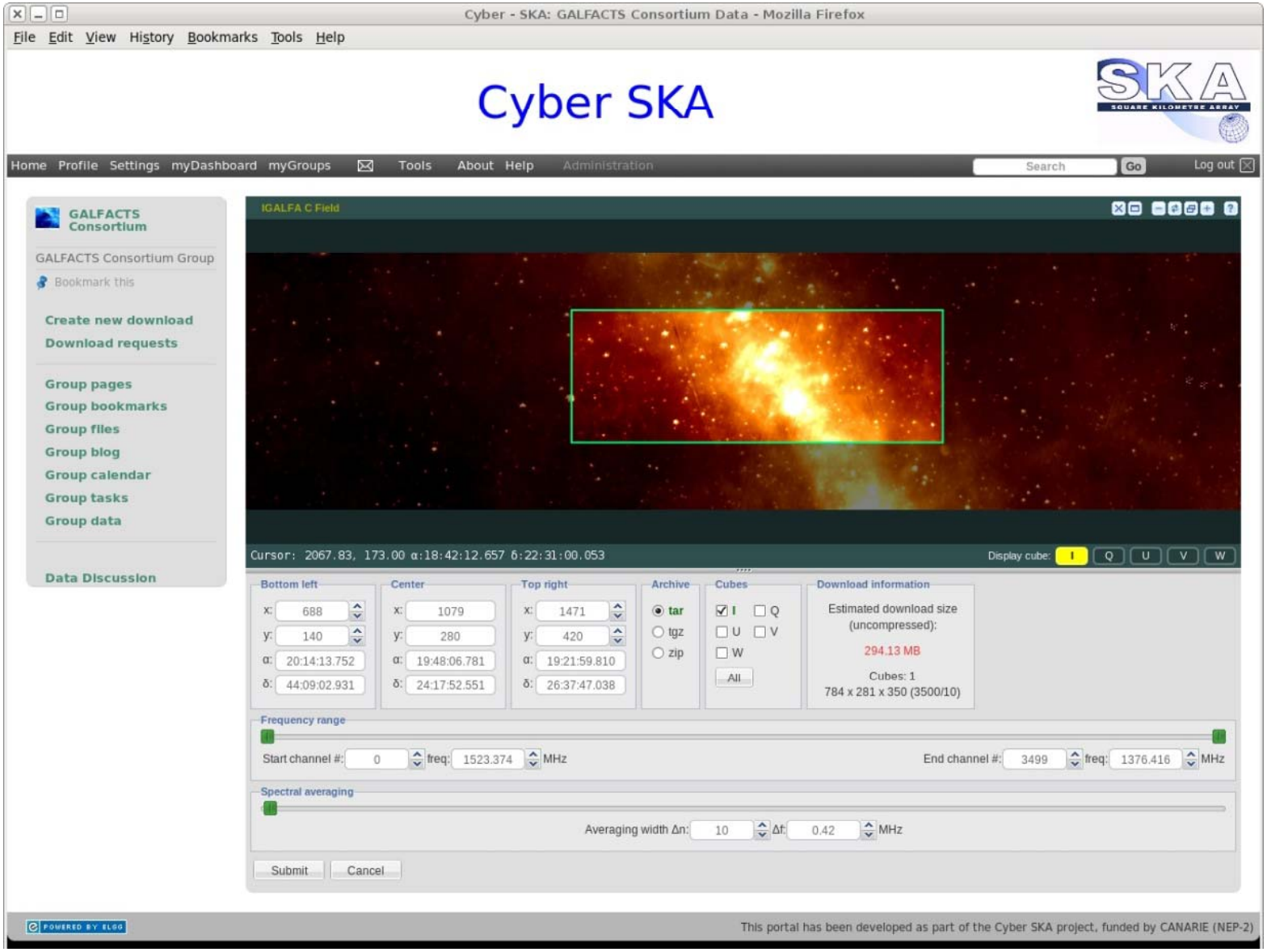
Solutions - Collaboration

- Portal (http://www.cyberska.org) built on top of the Elgg open source social networking platform
  - provides many Facebook-like features including: tags, bookmarks, profiles, blogs, wikis, friends/contacts, groups, media/document sharing, discussions, message boards, calendars, status, activity feeds



Solutions - Data

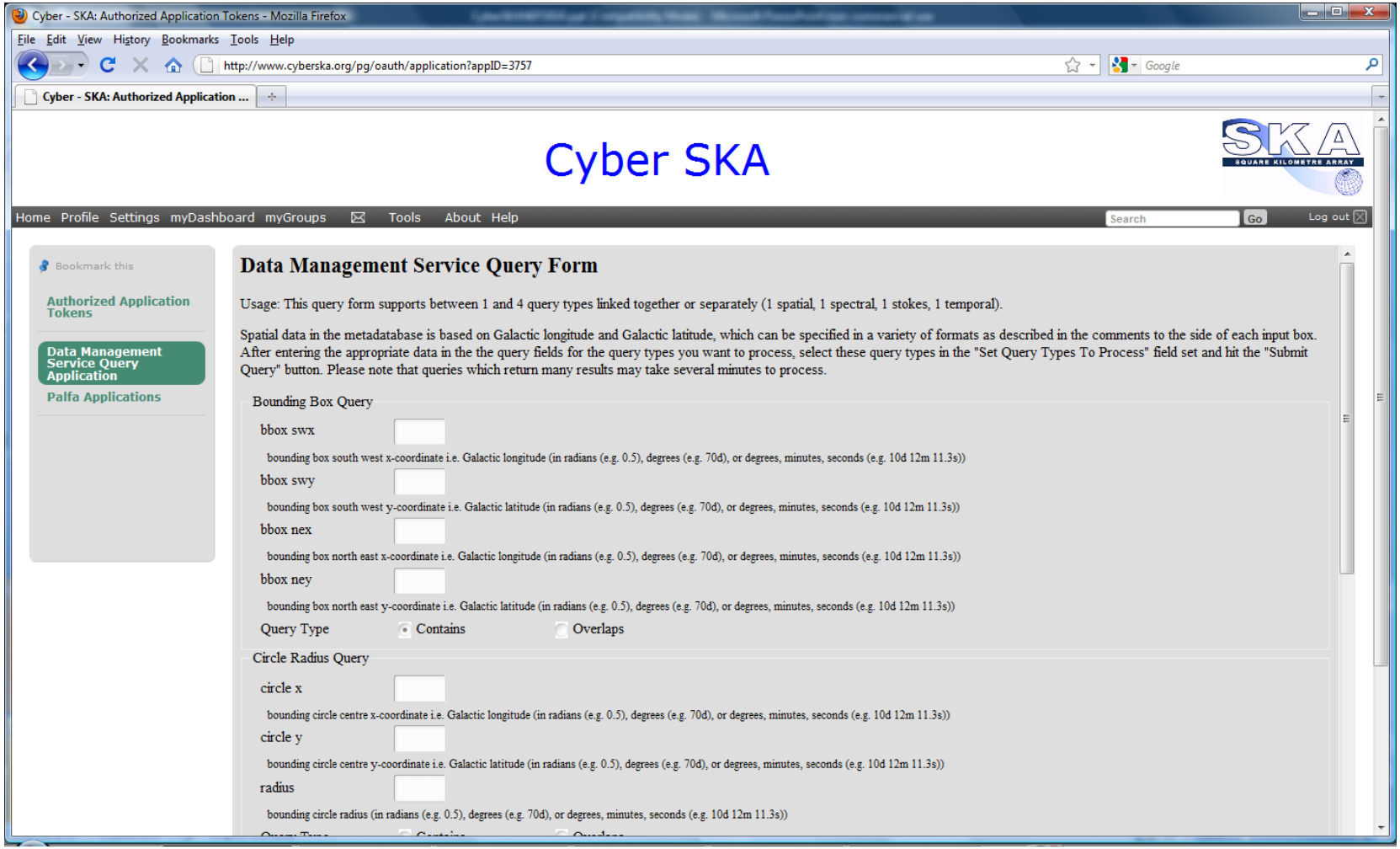
- Access/download data for selected parameters and region of interest
- Requested data generated in virtualized Condor pool on server side



Solutions – Data II

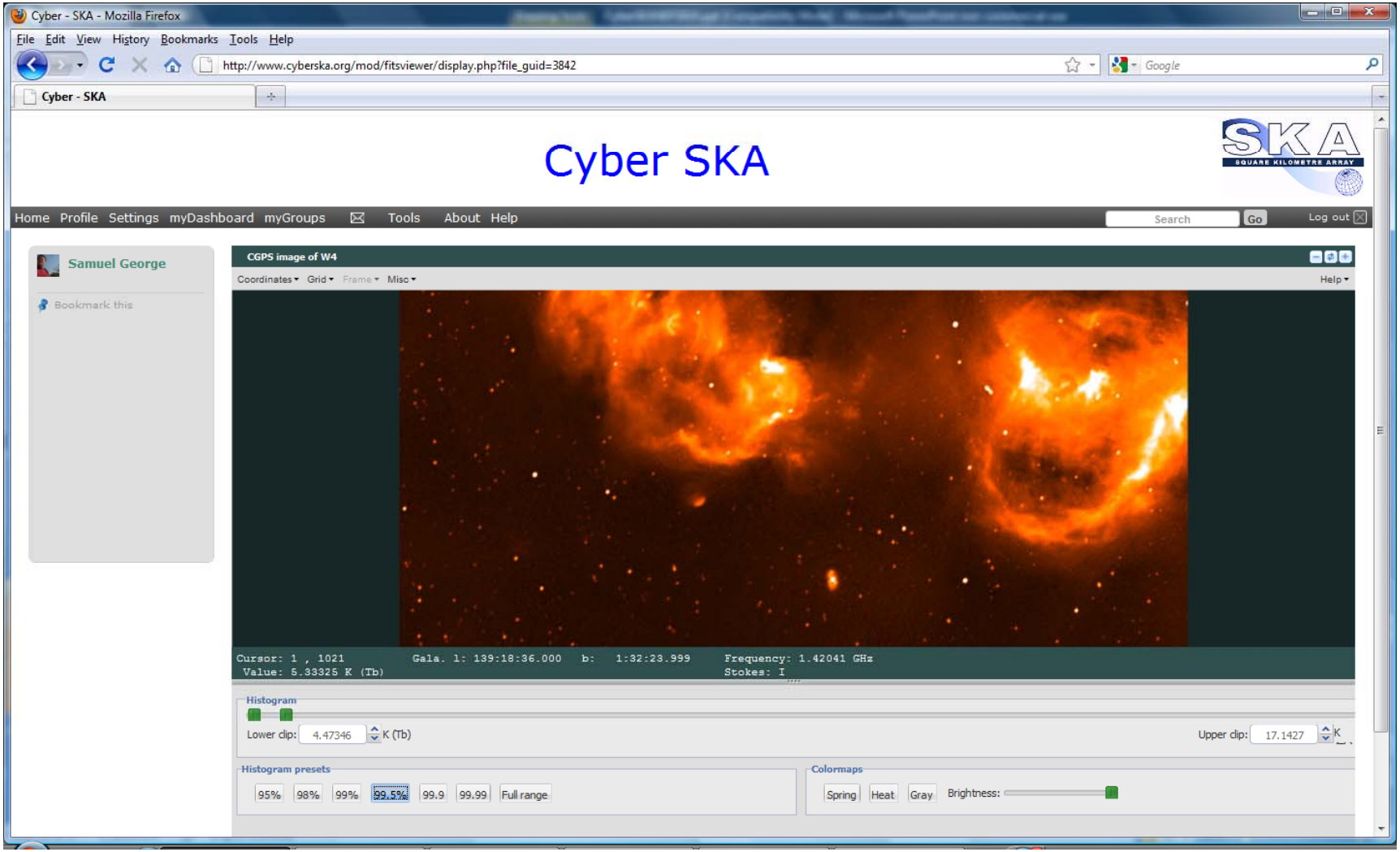
CyberSKA

- **Distributed data management service**
  - -built on iRODS (Integrated Rule-Oriented Data System)
  - -running at two sites currently (University of British Columbia Okanagan & University of Calgary)
  - -PostgreSQL database for image metadata (Adherent to VO metadata standards)
  - -query service with RESTful API (spatial, temporal and spectral queries supported)
  - -supports mosaicing of images returned by query



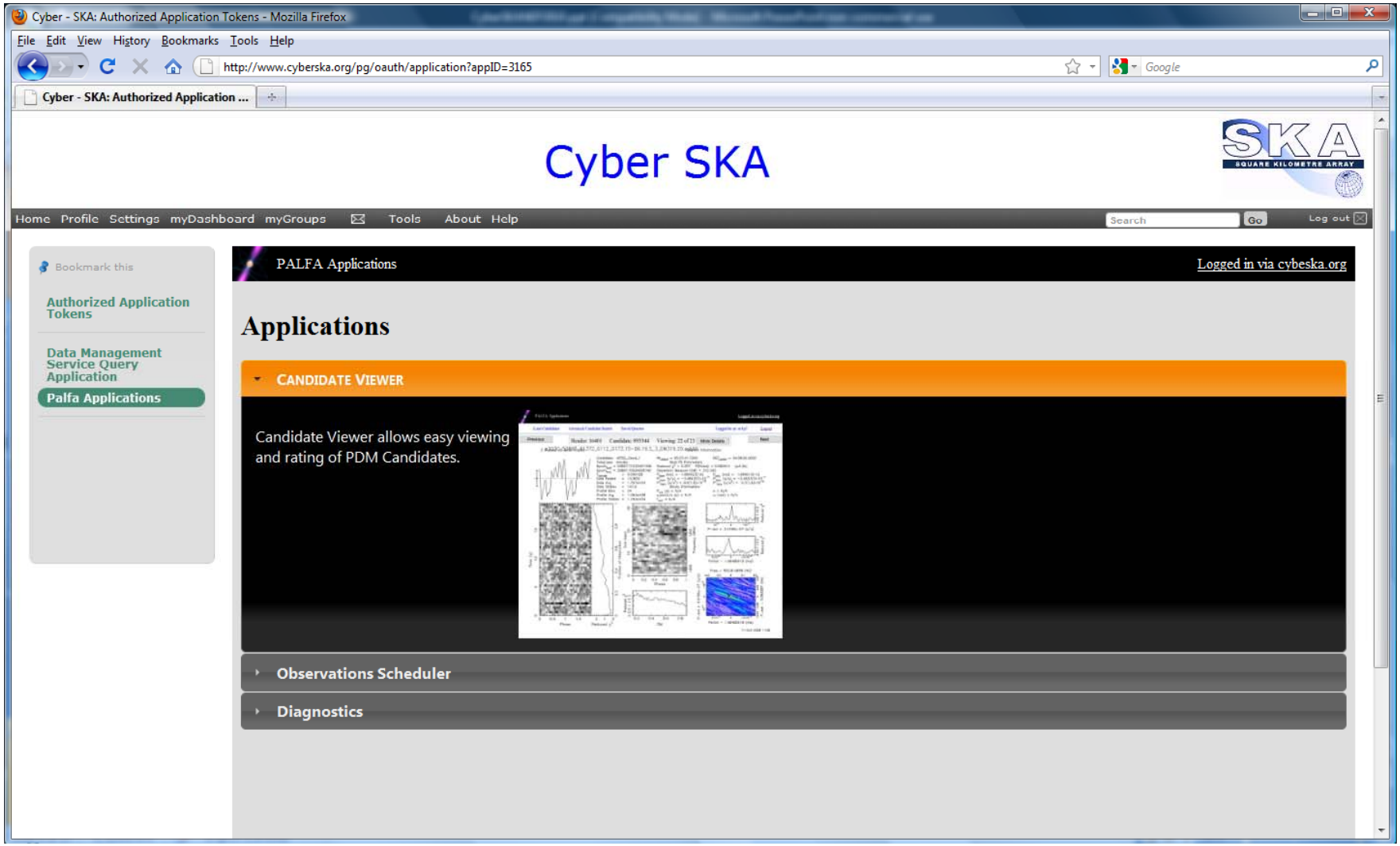
Solutions - Visualization

- **On-line visualization of multidimensional FITS files:**
  - Supporting: interactive panning & zooming, histogram correction, color map adjustments, display pixel data value, multiple coordinated systems, grids, selection of frame for multi-dimensional images



Solutions - Applications

- API for integrating third party / remotely hosted applications
- Single sign-on to applications enabled using OAuth



Next Steps

- **Infrastructure**
  - Acquisition of hardware at participating sites to establish prototype testbed
  - Set up cloud computing environments and key services at each site
    - Cloud platforms under consideration include ASPEN, Eucalyptus, Nimbus, OpenStack
- **Collaboration**
  - Refinement and development of collaboration features based on user feedback
- **Data Management**
  - Expansion of distributed data management system to other sites
  - Better integration of data management system with other CyberSKA tools and services
- **Data Visualization**
  - Provide server side support and improve scalability
- **Data Processing**
  - Establish dynamic batch based processing and interactive service environments on cloud platform
  - Establish framework for adding and integrating different processing algorithms and workflows
- **Applications**
  - Extension of third party application API to enable two way interaction between portal and applications (i.e. pull data/information from portal, push news feeds to portal based on application activities)

Contact Information

CyberSKA

Portal: http://www.cyberska.org/ E-mail: info@cyberska.org