ENVRI Common Operations of Environmental Research Infrastructures

Towards the Big Data Strategies for EISCAT 3D

Yin Chen
Y.Chen@cs.cardiff.ac.uk

Opportunities for new Research

- **EISCAT 3D New Measurement Capabilities**
  - Instantaneous, adaptive control of beam positions
  - Simultaneous multiple beams/interlaced beams
  - High-resolution coding of polarisation, phase and amplitude
  - Aperture synthesis imaging – small-scale 3D imaging(sub-beam-width)
  - Multi-beam volume imaging – large-scale 3D imaging
  - Full-profile vector measurements – large/small-scale 3D vector imaging
  - High-speed object tracking
    - Estimated for 3 MW Tx: improvement at least x 10 better

Opportunities for new Research

- **EISCAT 3D e-Infrastructure capabilities**
  - Real-time data access
  - Virtual research environment
  - Support long-tail scientists
    - Intelligent filter
    - Advanced discovery by signatures/patterns
    - User specific analysis/mining/processing
    - Support discovery of “unknowns”
    - Integration of external resources/global data sharing
  - New Applications, e.g.,
    - Space weather
    - Visualisation

The Big Data Challenges (3+1Vs)

- **Volume.**
  - 5PB/year in 2018, 40PB/year in 2023
  - Operate for 30 years, data products to be stored for > 10 years
- **Velocity.**
  - Each antenna : 120MB/s
  - 160 * antenna group (100 antennas): 2 Gbit/s/group
  - 5* Ringbuffer: each 125 TB/h
- **Variety.**
  - Measurements: different versions, formats, replicas, external sources ...
  - System information: configuration, monitoring, logs/provenance ...
  - Users' metadata/data: experiments, analysis, sharing, communications ...
- **Value.**
  - How to discover meaningful insights from low-value-density data
  - Needs new approaches to the deep, complex analysis e.g., machine learning, statistical modelling, graph algorithms etc.
  - Go beyond traditional approaches to the space physics

Towards The Big Data Strategies

- Interoperability between ESFRI ENV RIs
- Common solutions to common problems

The ENVRI Project

Started 11/2011 — Now 08/2013 — Ends 10/2014

Common Requirements

Common Services/Components

ENVRI Reference Model

- Evaluation/adoption of ENVRI RM
- Experiments with e-Science technology
- Evaluation of EGI/EUDAT Services

- Community Standard
- Common language
- Unified framework

The Pilot Study

Started 02/2013 — Ends 02/2014

- Support EISCAT Science Community
- Common services for the Big Data
- Training of Data Scientists

Horizon 2020 Proposal

Preparation 08/2013 — Submission 02/2014 — Starts 11/2014

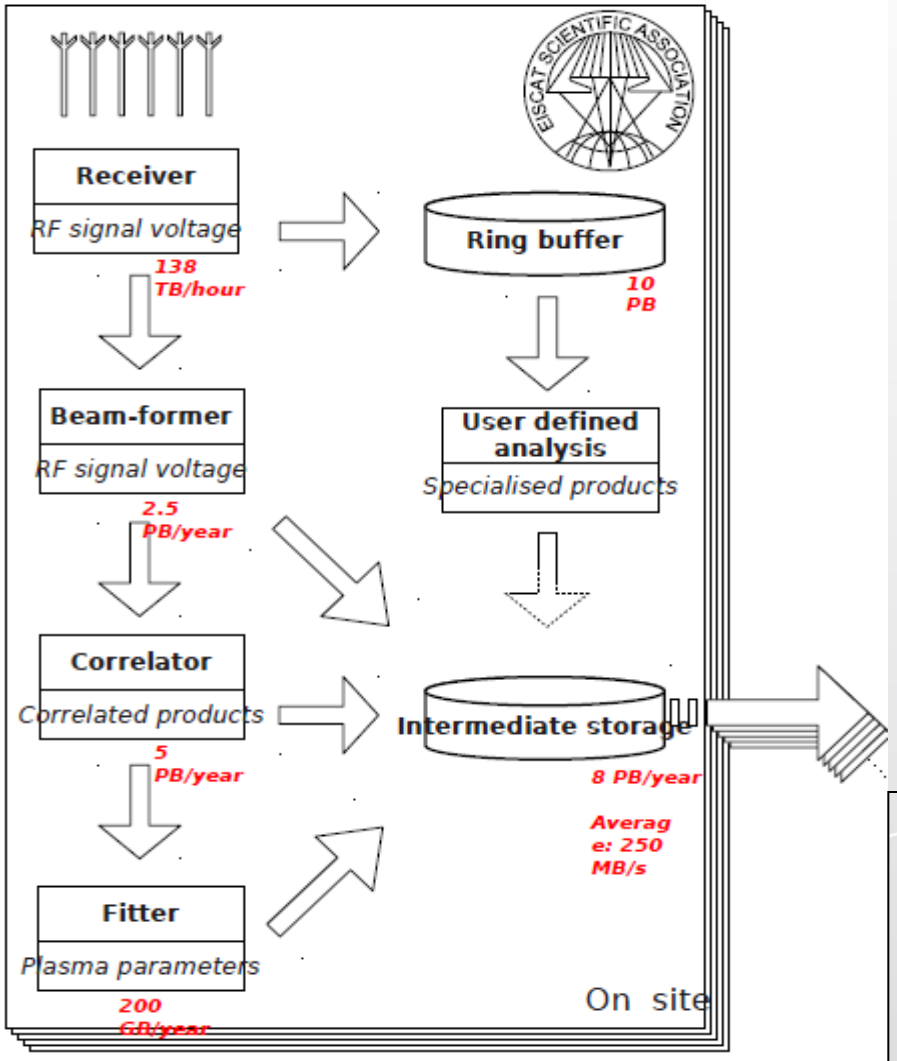
Goals for the Pilot Study

- Early adoption of the ENVRI Reference Model
  - Analysis and architecture design
  - Organising collaborative design activates
- Experiments with e-Science approaches
  - Distributed data archive
  - High throughput computing for processing
- Evaluation of the usability of EGI/EUDAT services
  - Within the EISCAT 3D e-Infrastructure
  - In supporting the EISCAT science community

E-Science Technologies

- **EGI**
  - A Europe-wide federation of national computing & storage resources.
  - ~350 resource centres from the NGIs, across 55 countries
  - > 370,000 logical CPUs, 248 PB disk, 176 PB of disk capacity
- **EUDAT**
  - European collaborative data infrastructure for e-Science
  - 23 partners from 13 countries 15 user communities
  - Offers storage, computing, metadata services to large research communities (e.g., ESFRI)

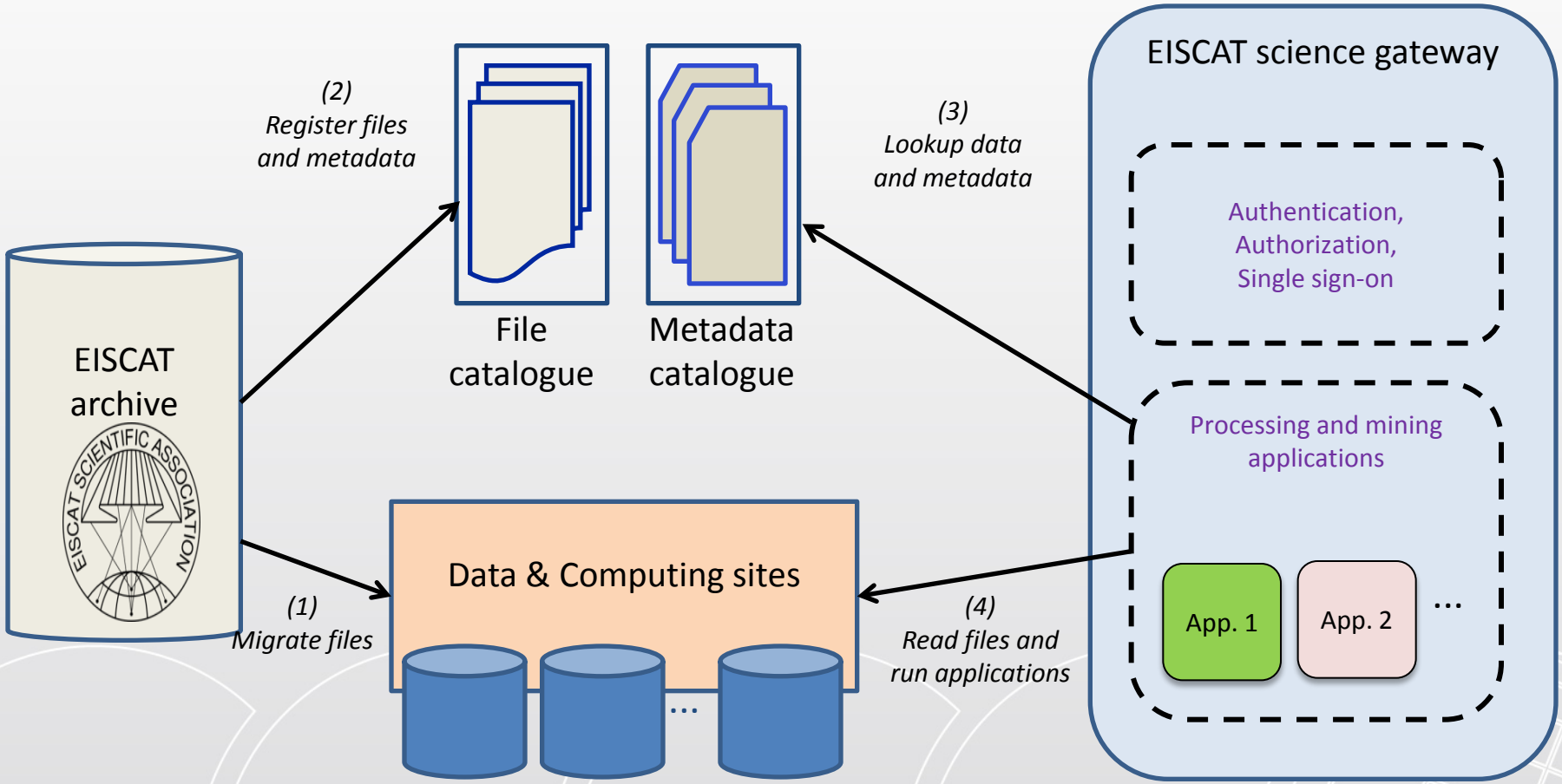
EISCAT-3D Data Acquisition



- **5 Types of data**
  - Raw antenna (group) data (138 TB/h)
  - Voltage beam formed data (2.5 PB/year)
  - Correlated products (5 PB/year)
  - Fitted data (200GB/year)
  - (User) Specialised Products

- Numbers are per site, 5 sites in total
- The yearly rates are based on 24/7operation
- 10% of the time with full power
- 90% of the time with 10% power,
- In total 20% of average maximum rates

EISCAT-3D Data Curation



EISCAT-3D Data Curation

- EGI (Grid) Services, e.g.,
  - Metadata catalogue -- AMGA
  - File catalogue -- LFC
  - Storage element
  - File Transfer Service
  - Portal for application development & hosting (e.g. SCI-BUS)
  - Access control
- EUDAT Services
  - Safe Replication
  - Data Staging (moving large data)
  - Simple Store (uploading and sharing data)
  - Metadata (including a portal for the service)
  - To come: Dynamic Data, Annotating Data etc.
- Usable solution with compromises

Data Access & Processing

- **Unlock the hidden value of the big data**
  - **Discovery & Access**
    - **Search through all levels of data, e.g.,**
      - **Find specific signatures**
      - **Plasma features, meteors, space debris, astronomical features**
      - **Automatic switching between high and low power modes**
    - **Search for other ISRs data resources**
  - **Processing**
    - **User specified analysis/correlation process**
    - **Visualisation**

Requirements Collection

- **For Scientists -- What applications do you use/need?**
  - **Search**
  - **Analysis/process,**
  - **Visualise**
  - **In any other way manage/ interpret the data.**
- **For Data manager -- How data are managed ?**
  - **Structure**
  - **Used file formats**
  - **Metadata structure**
  - **Needs for replication**
  - **User base,**
  - **Usage patterns**
  - **Accessibility/availability/security**

Towards Horizon 2020

- **Objective1: e-Infrastructure support to EISCAT Community**
  - Real-time data access
  - Community driven co-design
  - Virtual research environments
  - Support of long tail of scientist
  - Global data sharing and integration
- **Objective2: Common Services for Big Data**
  - Identify common requirements, challenging issues, state-of-the-art design experiences, e.g., LOFAR, LHC, SKA, etc.
  - Proof of concepts of data infrastructure-enabling software
- **Objective3: Training of Data Scientists**
  - A new data-centric way of organising research activities
  - New approaches to solve problems
  - Lead to significant scientific breakthroughs

Next steps

- **Widen Discussions**
  - **16.20pm: COOPEUS/ENVRI/EGI open forum & splinter group meetings**

- **Requirements Collection**
  - **From scientists: http://tinyurl.com/EISCAT-Sci**
  - **From data managers: http://tinyurl.com/EISCAT-DM**

Involved Organisations

- Cardiff University, UK
- CNRS, France
- CSC, Finland
- EGI.eu, The Netherlands
- EISCAT, Sweden
- EUDAT (via its partners)
- University of Edinburgh, UK