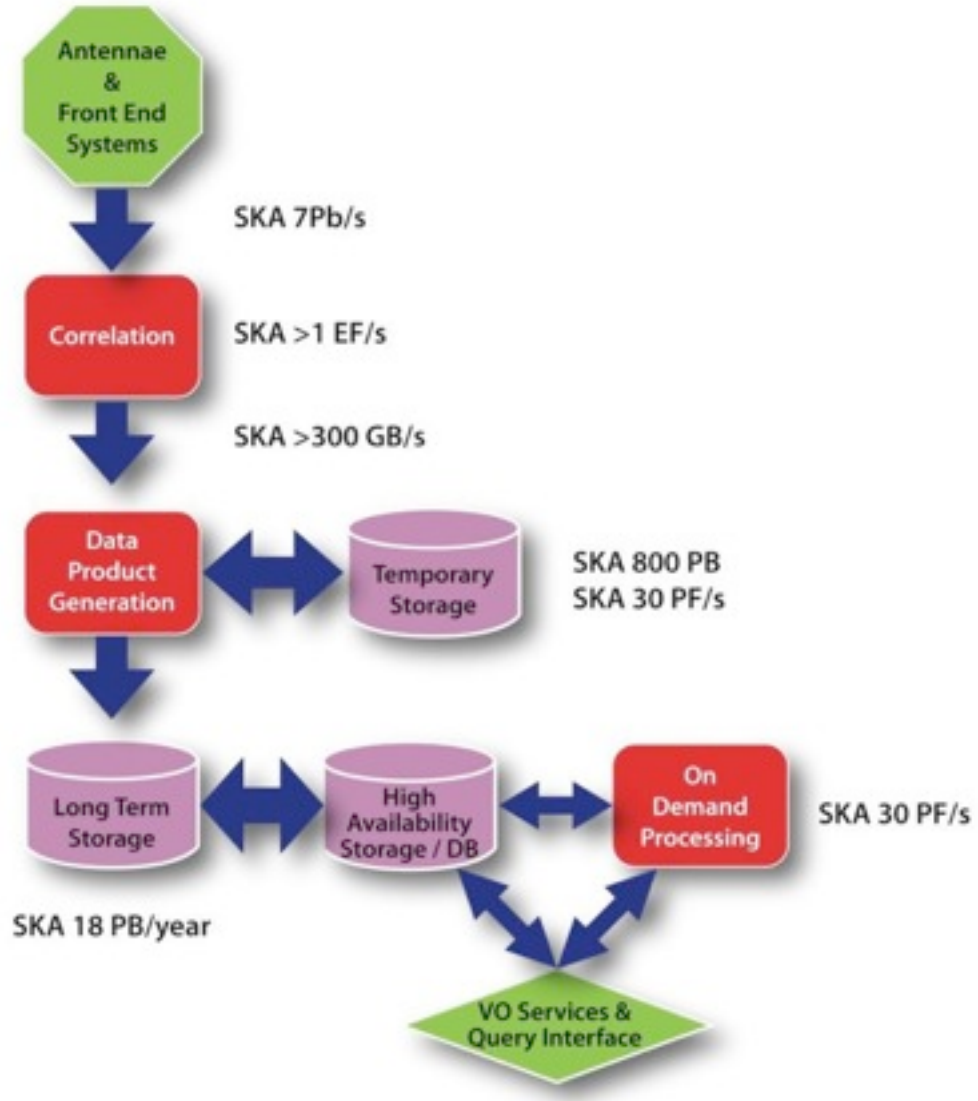
SKA data-deluge: e-Science solutions

Lourdes Verdes-Montenegro
Instituto de Astrofísica de Andalucía (CSIC)

Challenges in SKA and pre-SKA era



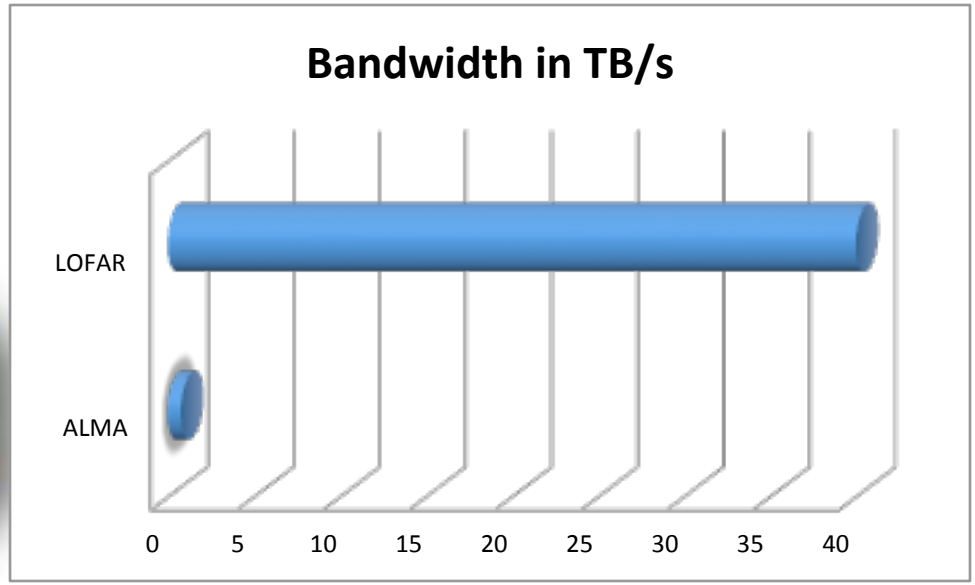
Processing

Storage

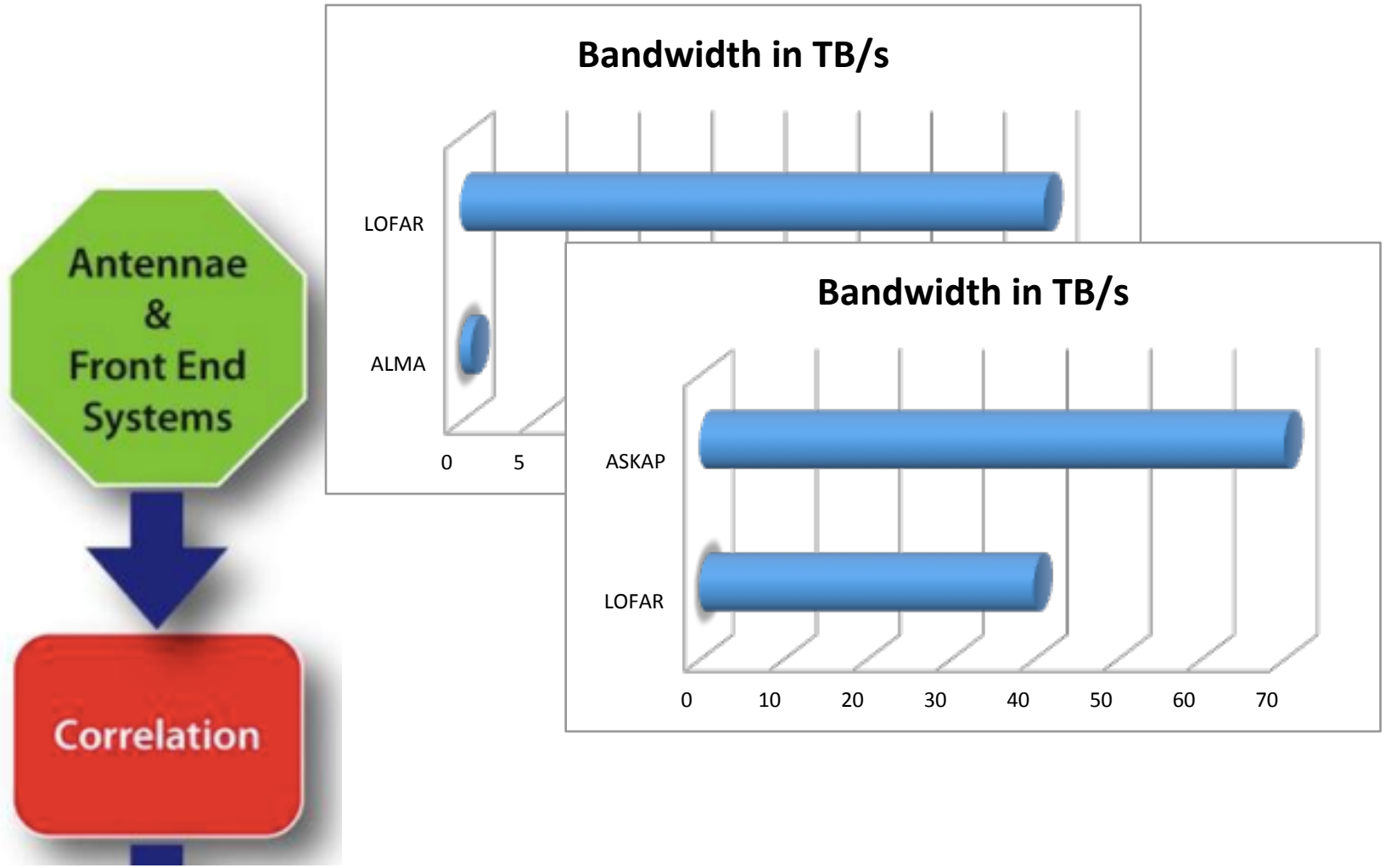
Bandwidth

Power

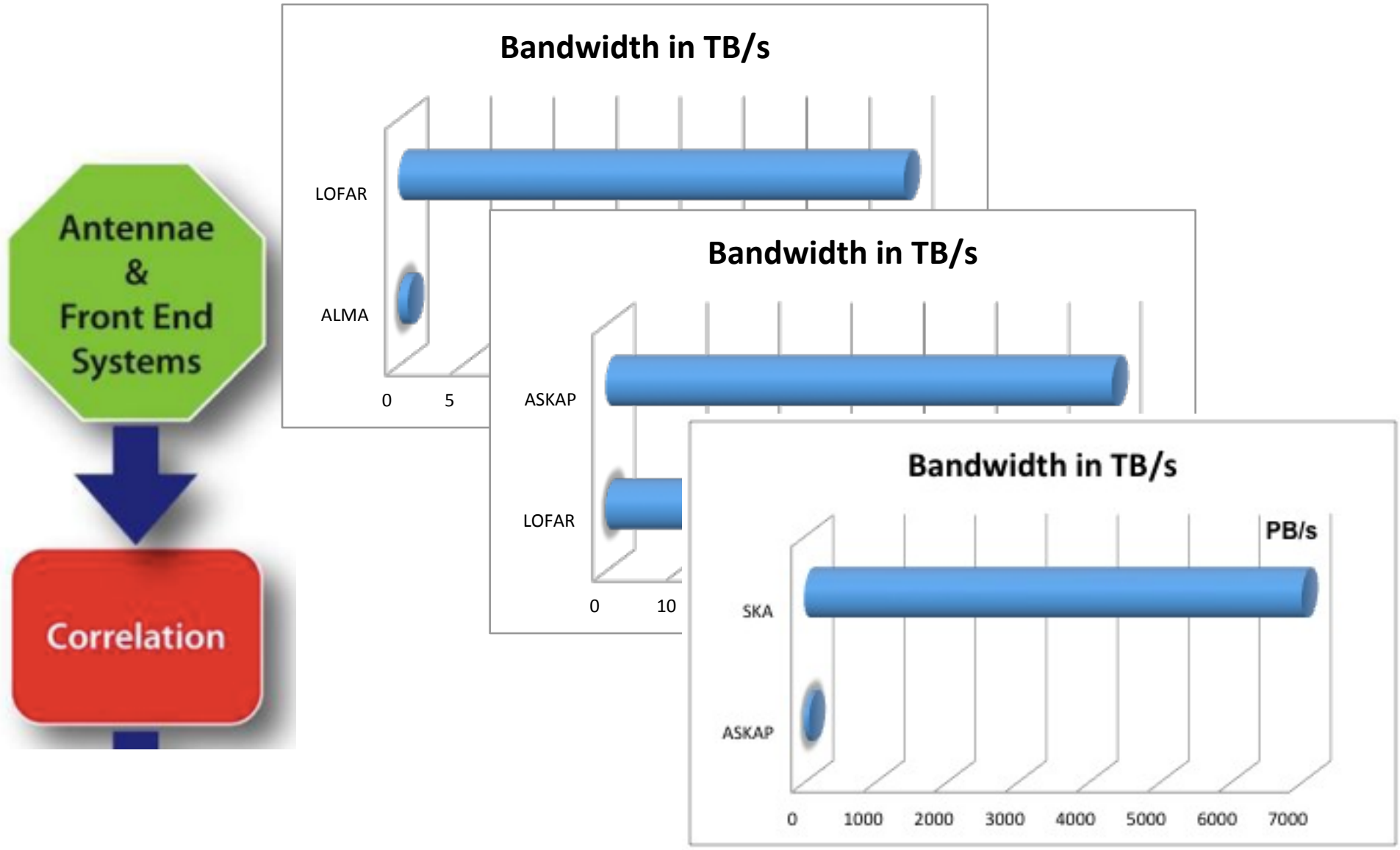
Challenges in SKA and pre-SKA era



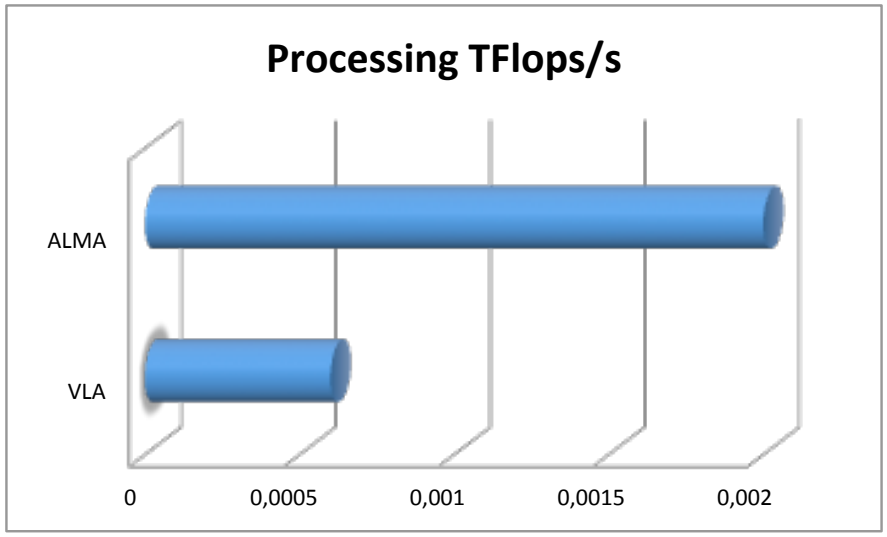
Challenges in SKA and pre-SKA era



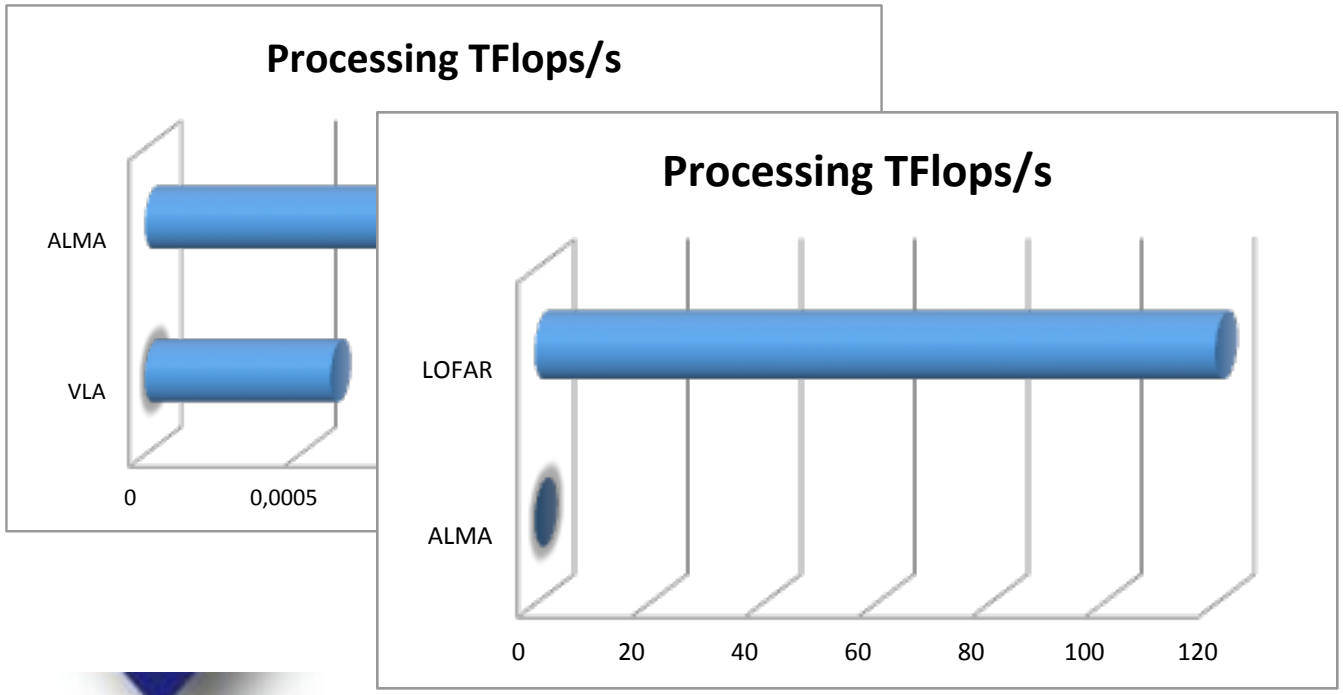
Challenges in SKA and pre-SKA era



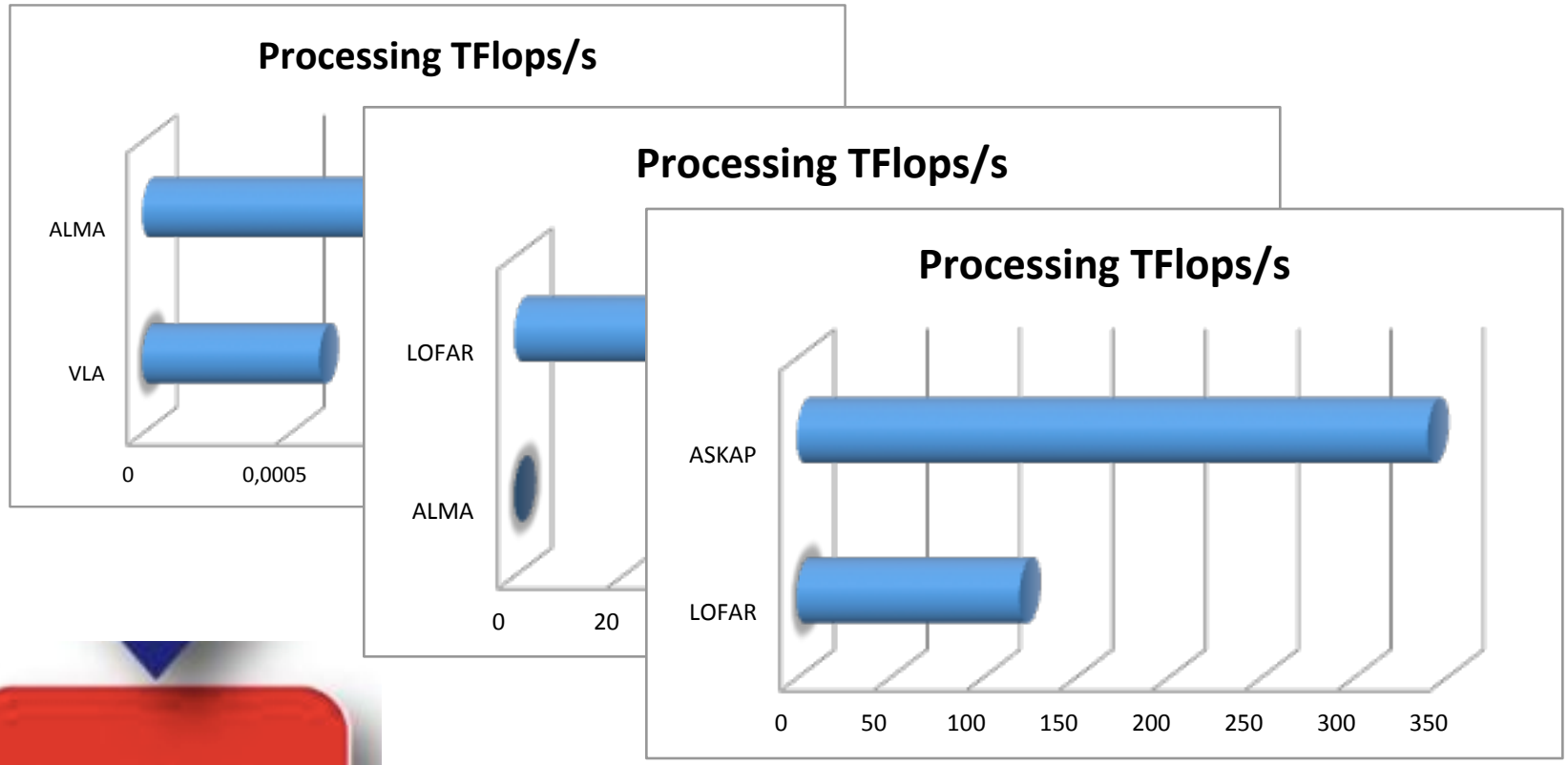
Challenges in SKA and pre-SKA era



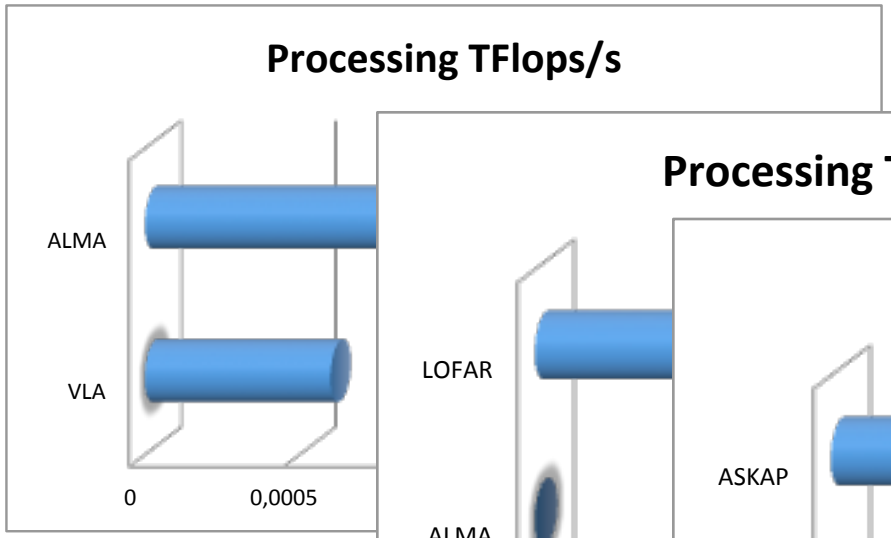
Challenges in SKA and pre-SKA era

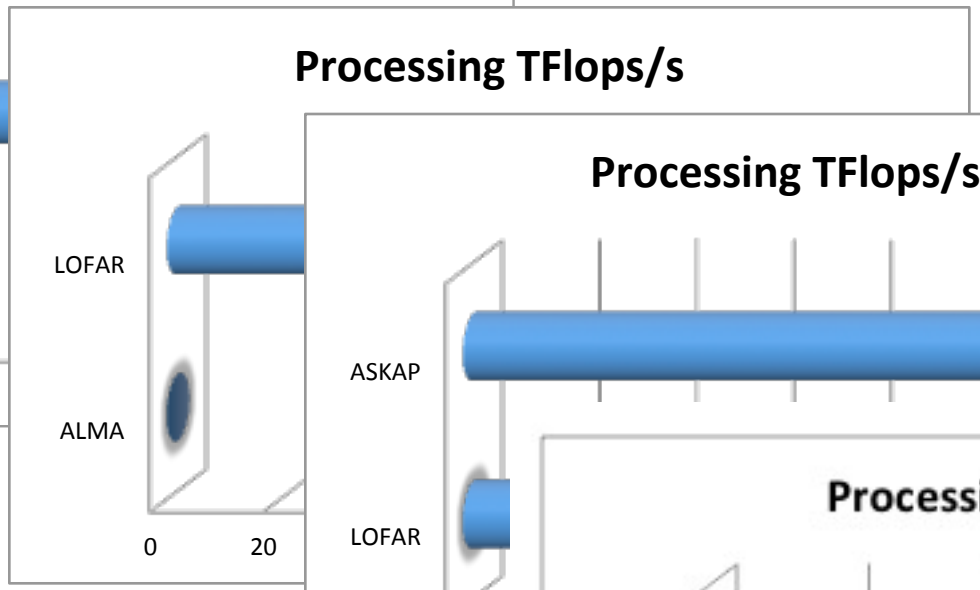


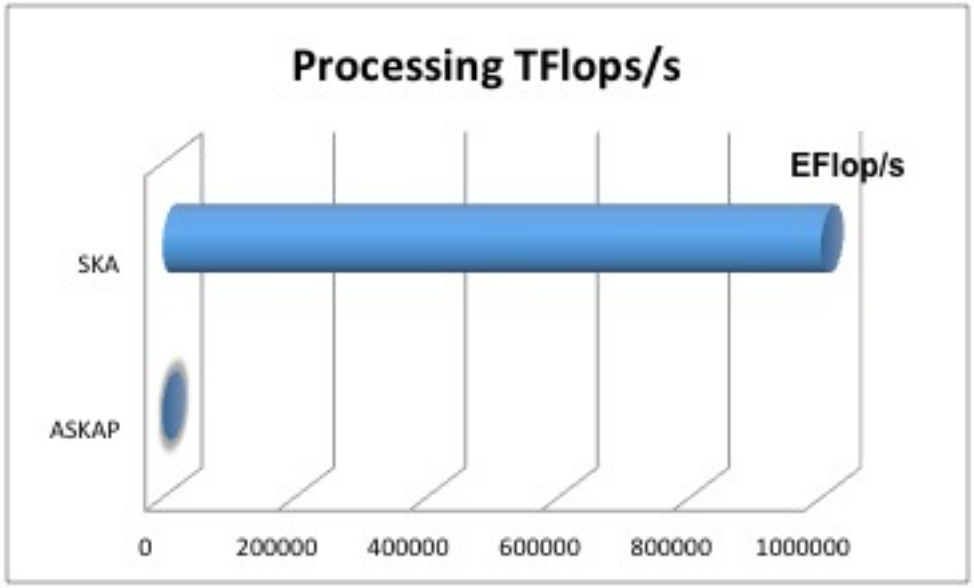
Challenges in SKA and pre-SKA era



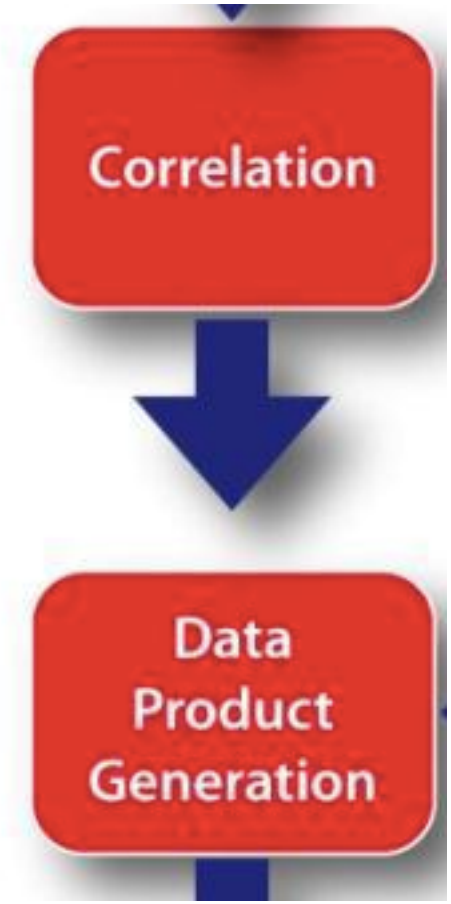
Challenges in SKA and pre-SKA era

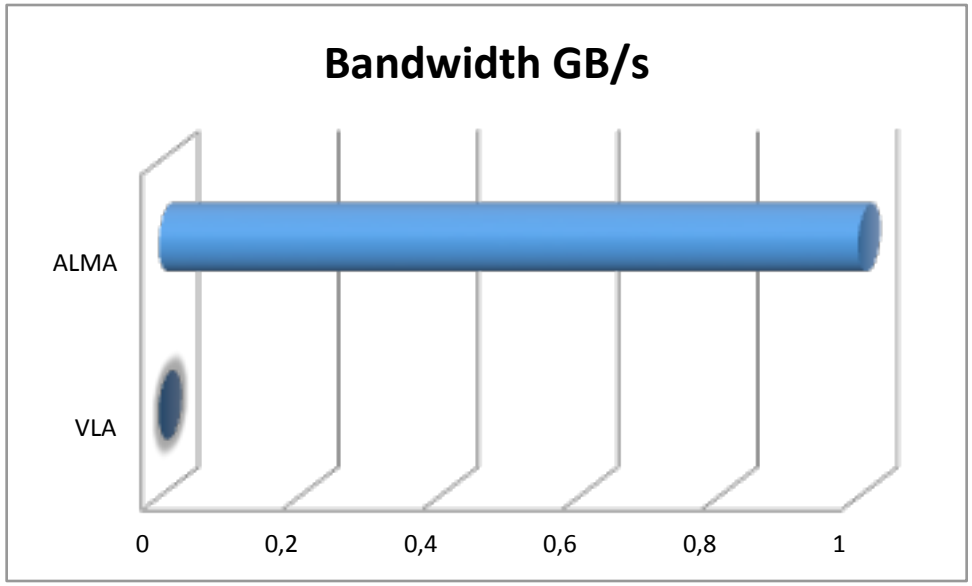




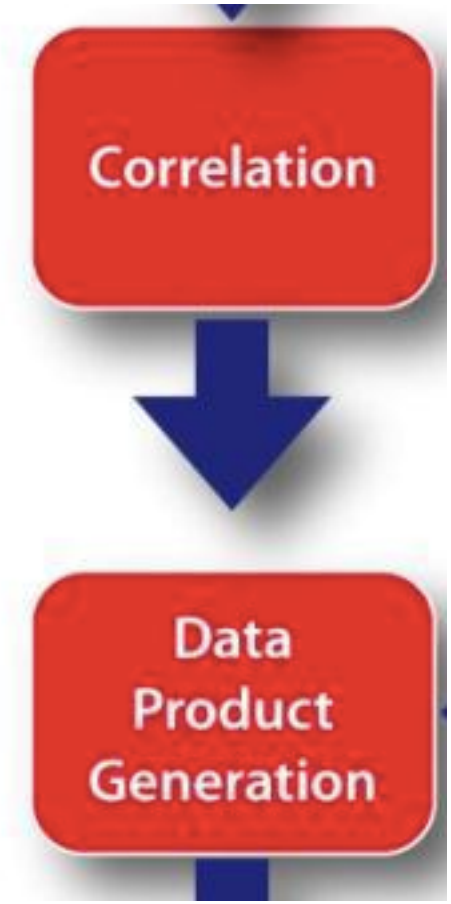


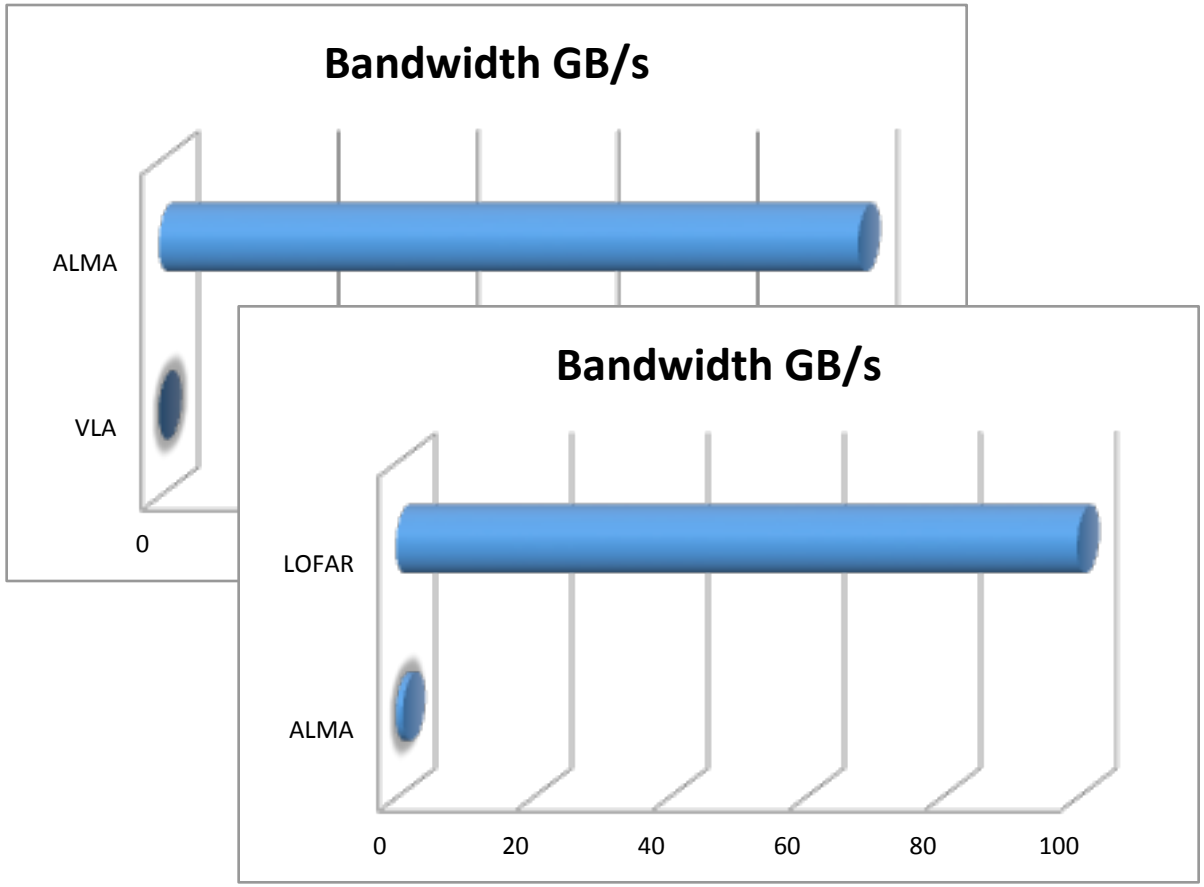
Challenges in SKA and pre-SKA era



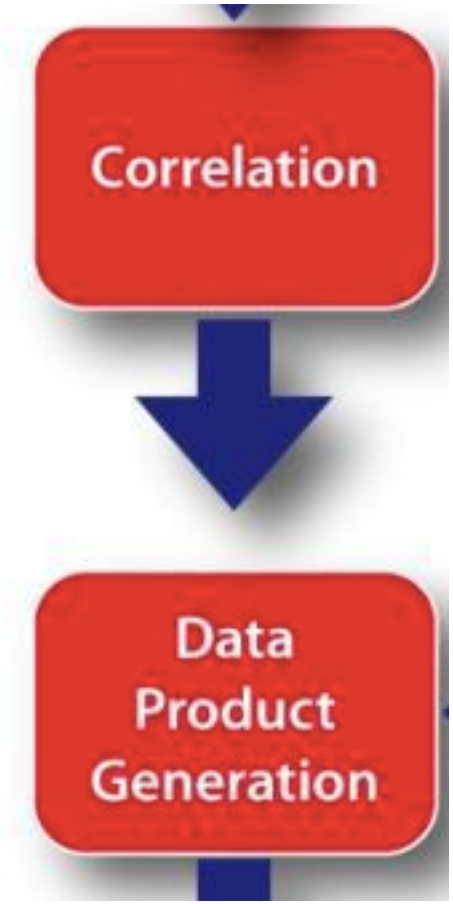


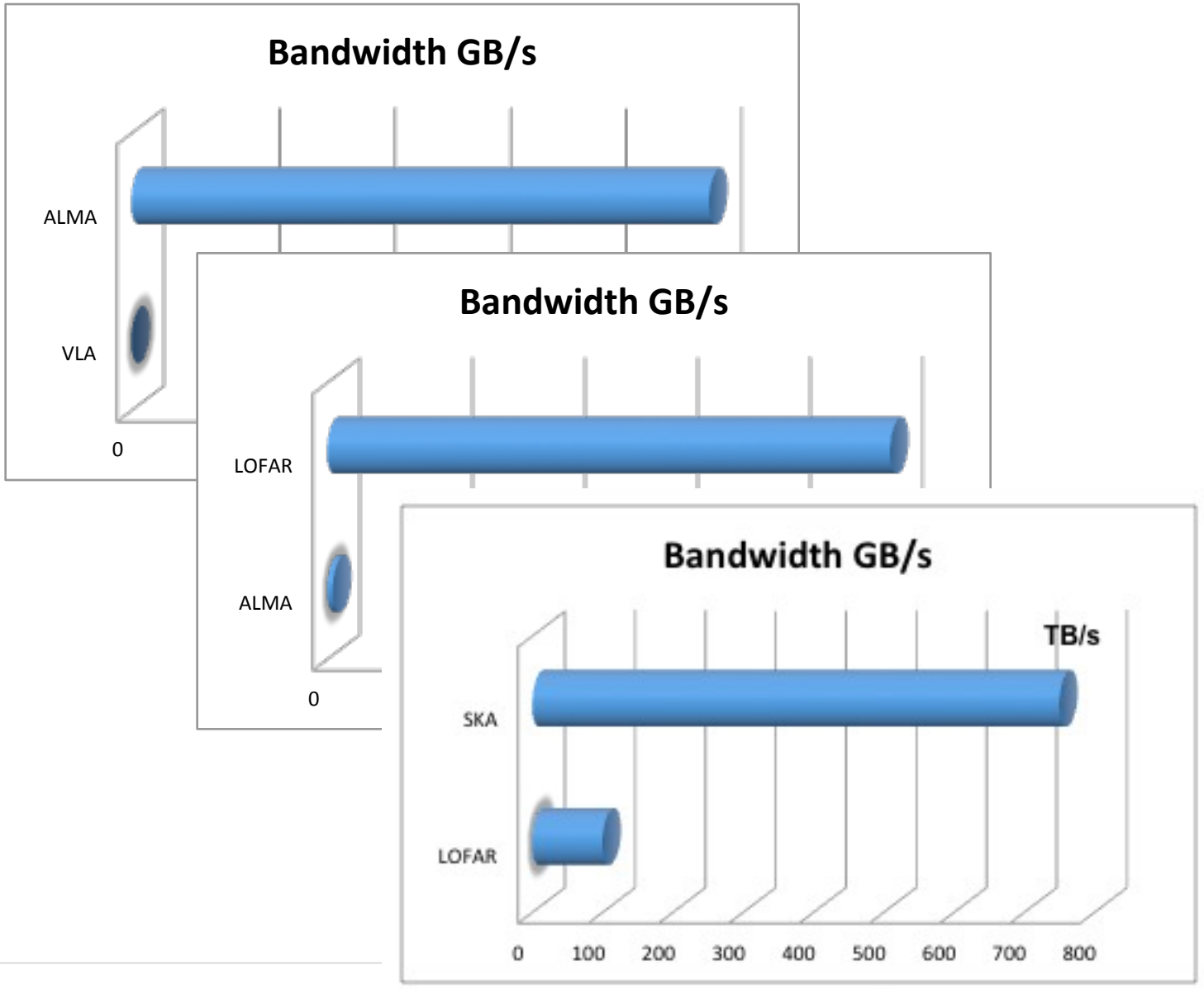
Challenges in SKA and pre-SKA era



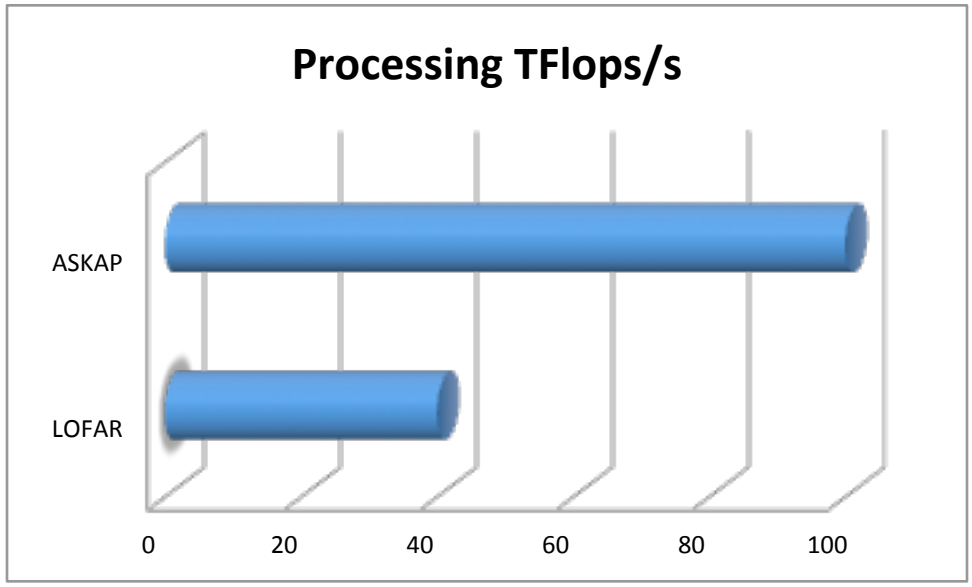


Challenges in SKA and pre-SKA era

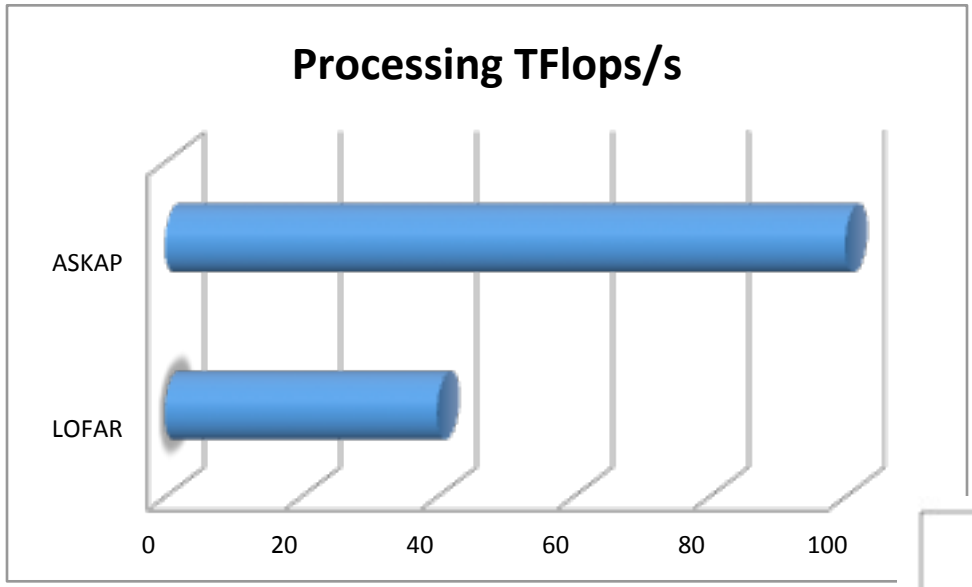


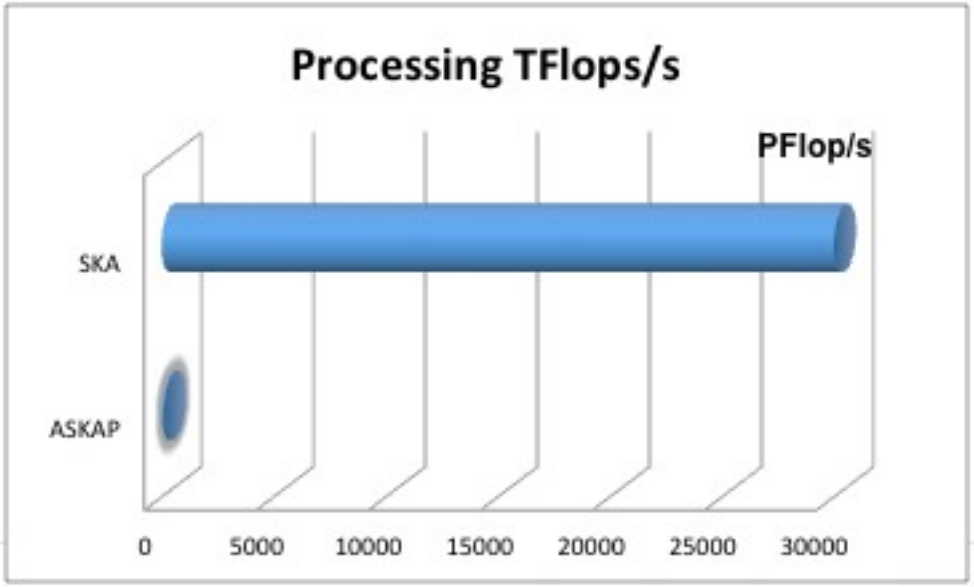


Challenges in SKA and pre-SKA era

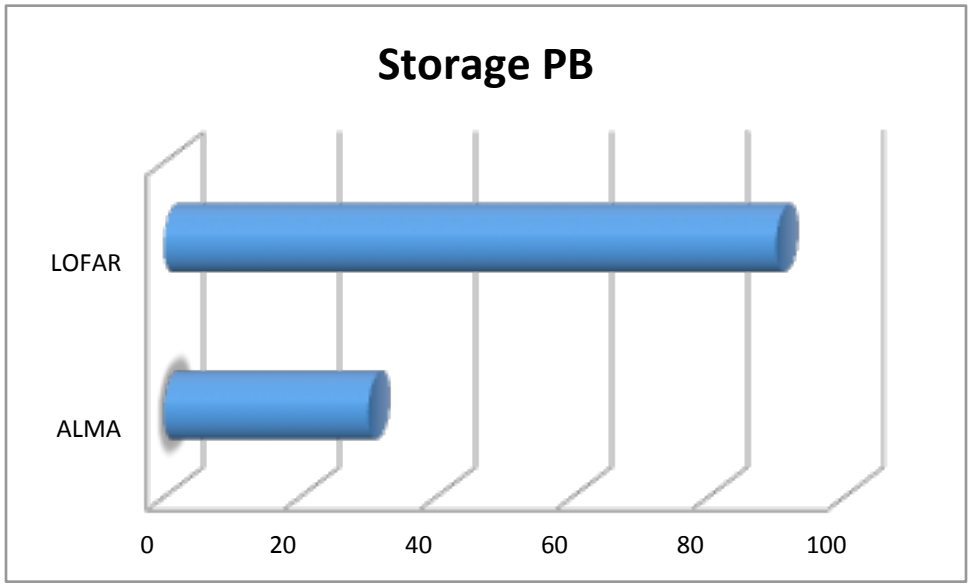


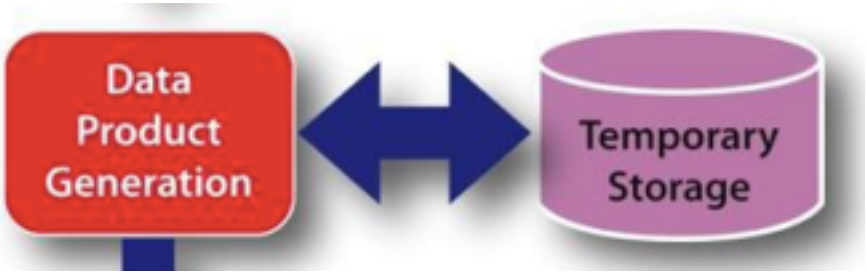
Challenges in SKA and pre-SKA era



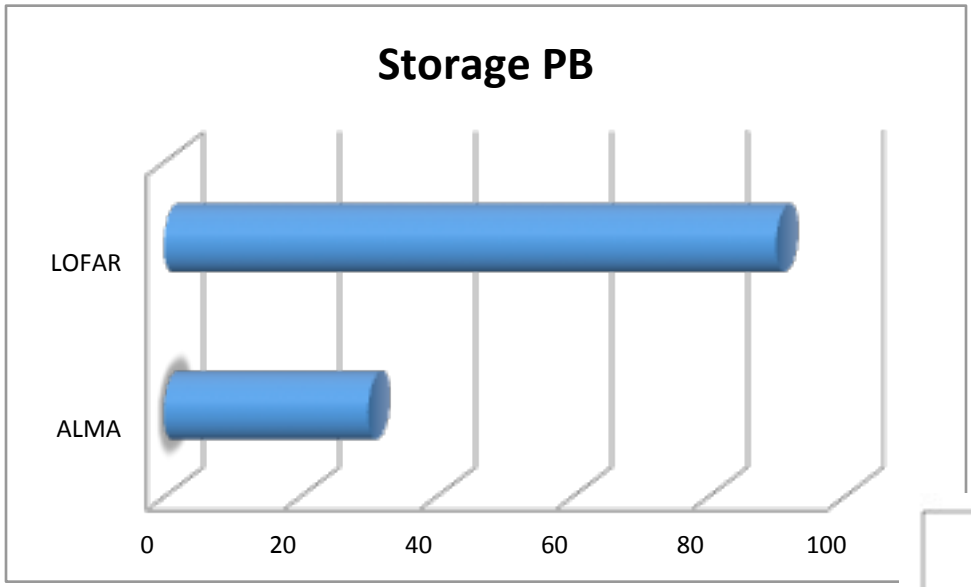


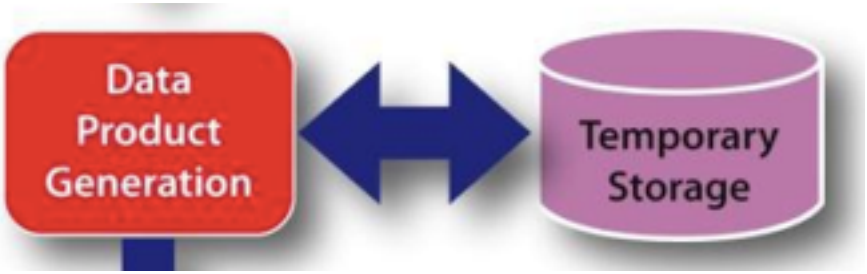
Challenges in SKA and pre-SKA era

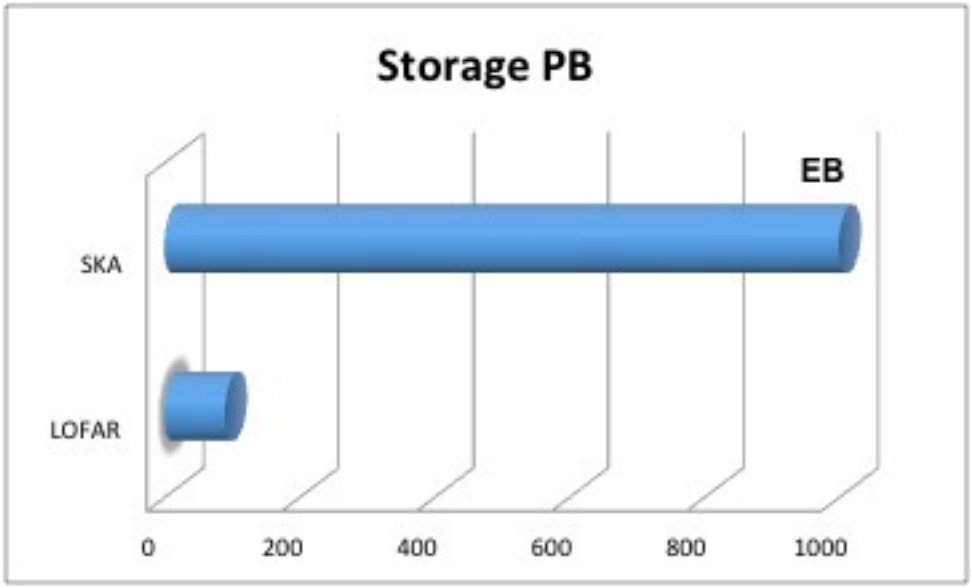




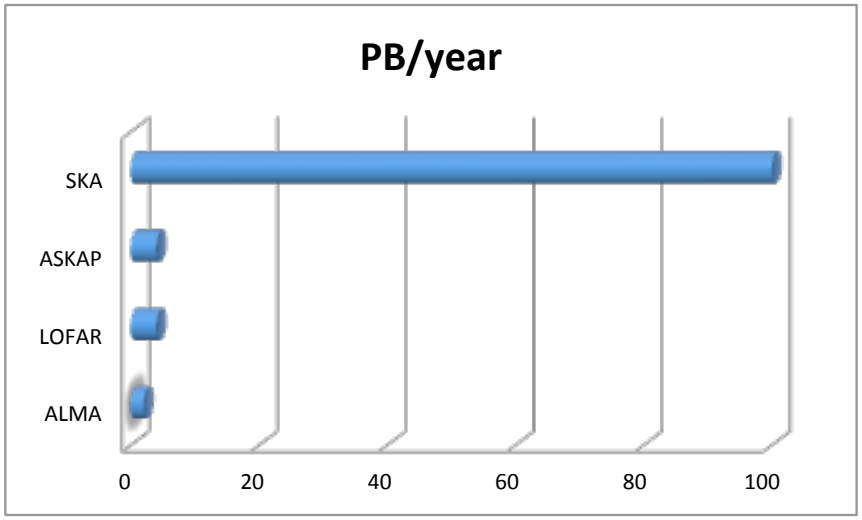
Challenges in SKA and pre-SKA era

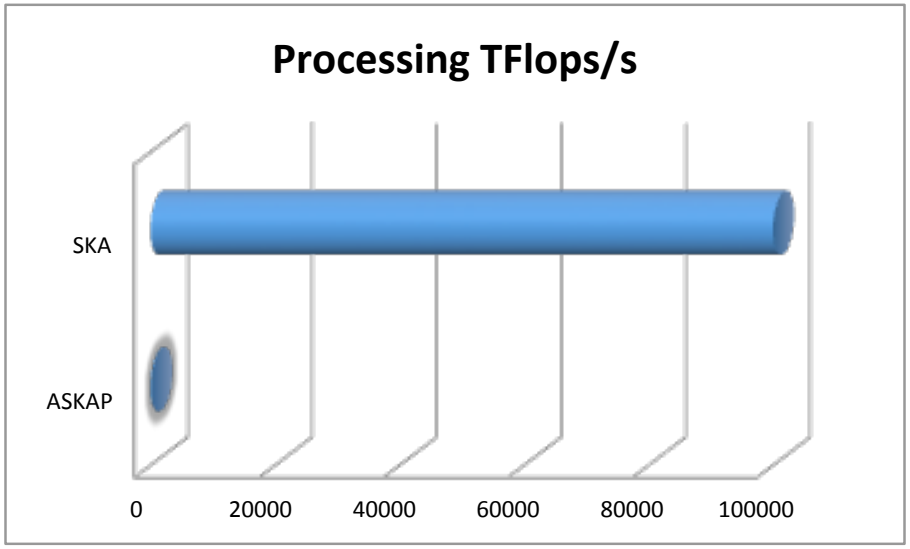


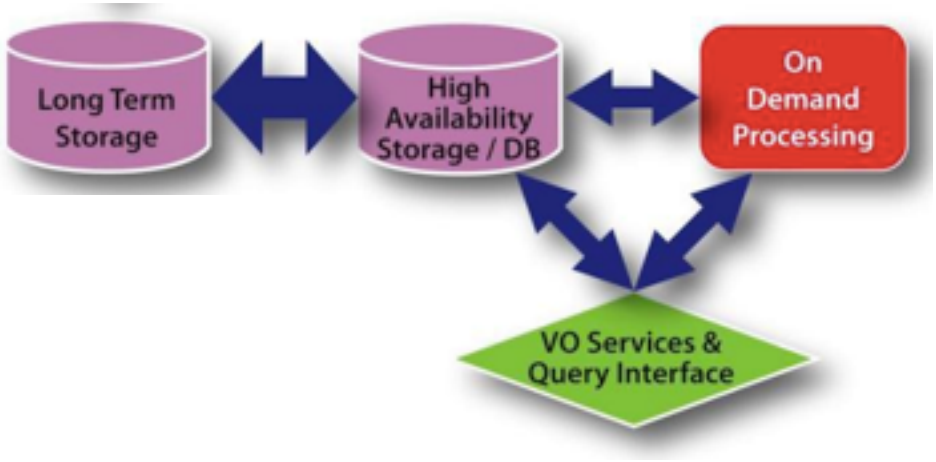




Challenges in SKA and pre-SKA era







Challenges in SKA and pre-SKA era

Processing

SKA processing needs are equivalent to **1 billion top range PCs**

Bandwidth

SKA aperture arrays will produce **250 times the current Global Internet traffic**

- SKA Pathfinder Cubes ~ 4.4 TB which implies 7.3 min read time at 10GB/sec
- Typical survey consists of ~**1000 cubes = 5 days read time**
- **Need:**100-1000 GB/sec for on-demand processing single cubes and cube groups

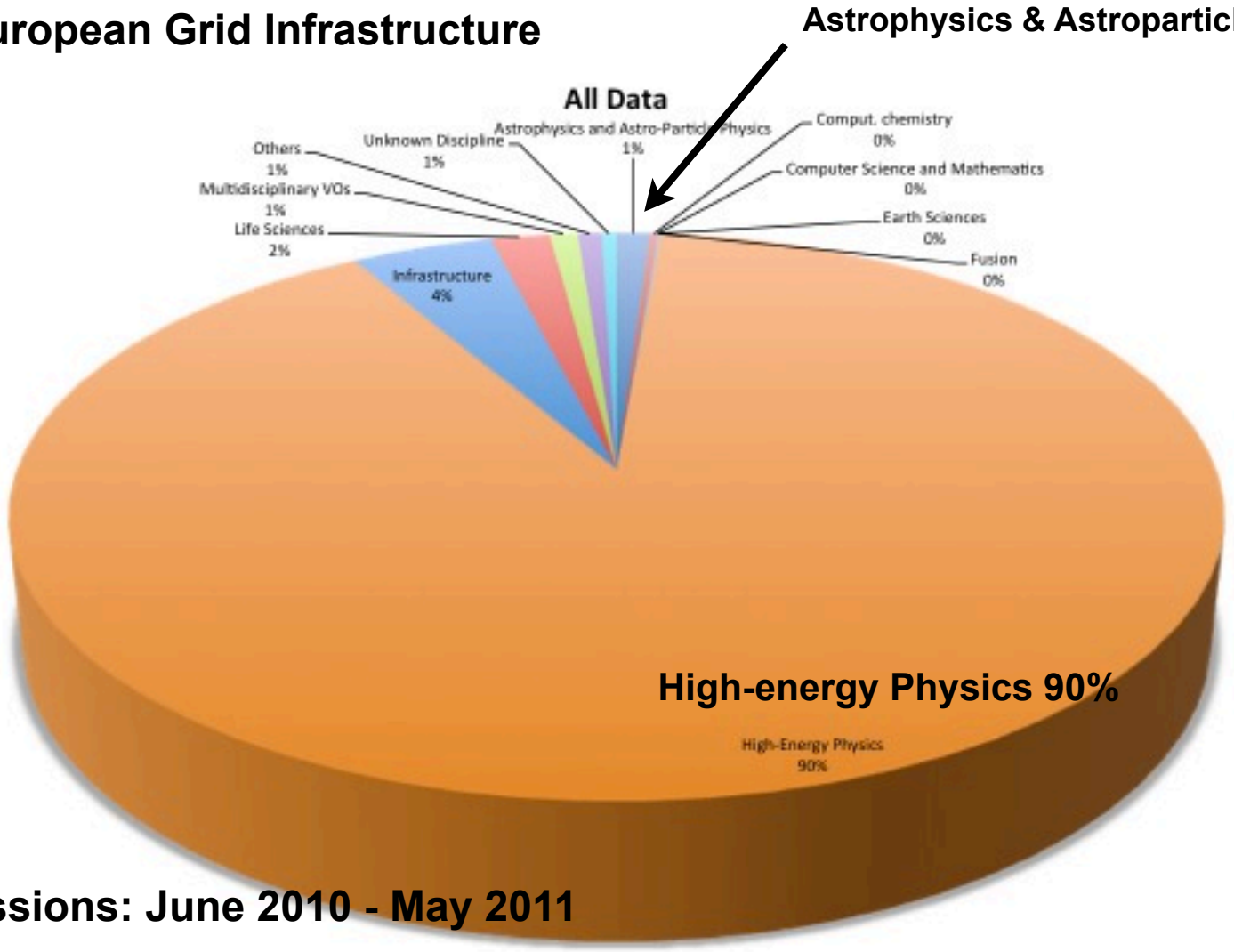
Power

- SKA HPC power consumption 1EF/s ~100 MWatt

Storage

- SKA will produce in **one day the annual data product of all mankind**

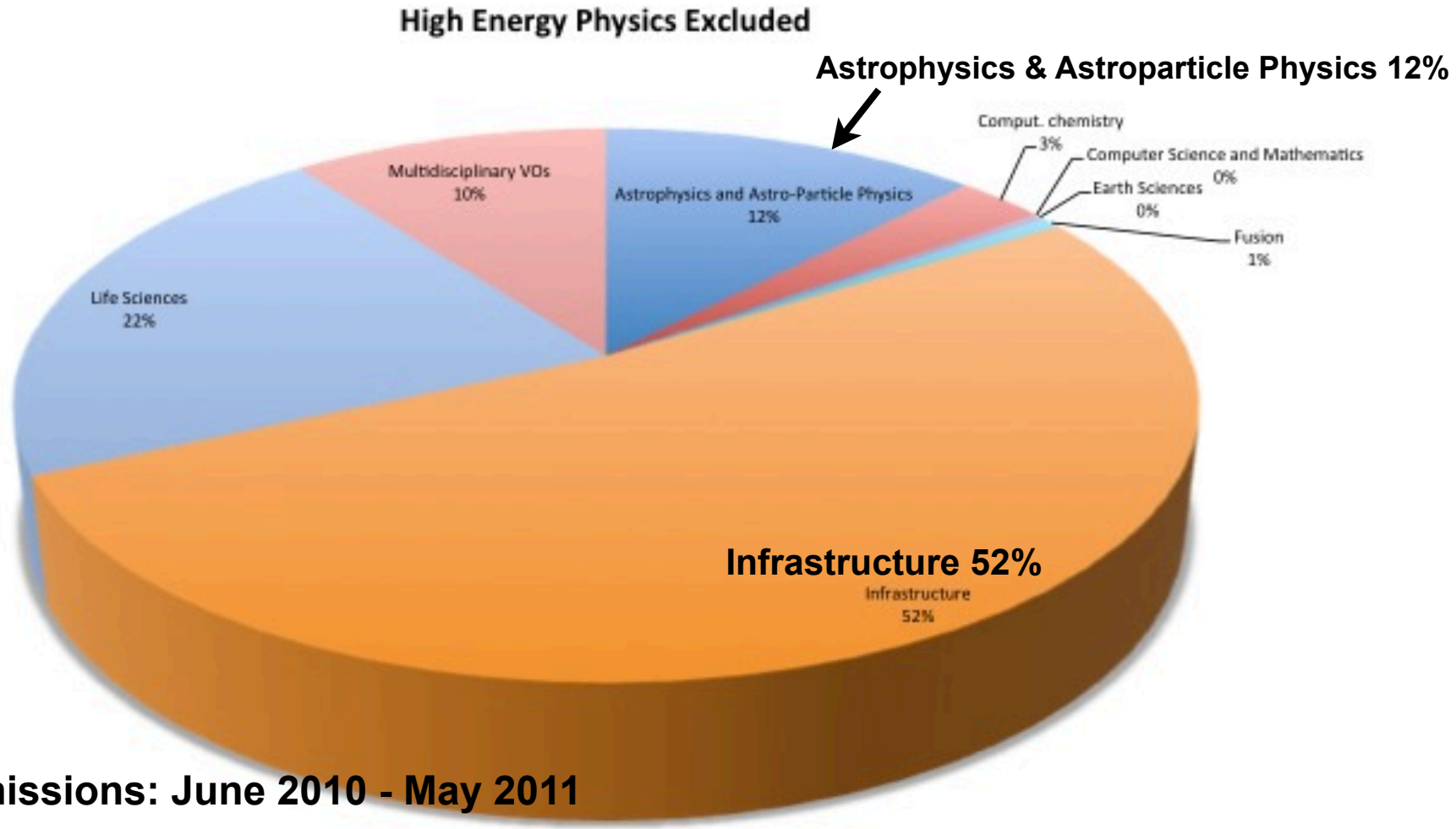
Use of European Grid Infrastructure



Job submissions: June 2010 - May 2011

Obtained using EGI statistics tool http://accounting.egi.eu/

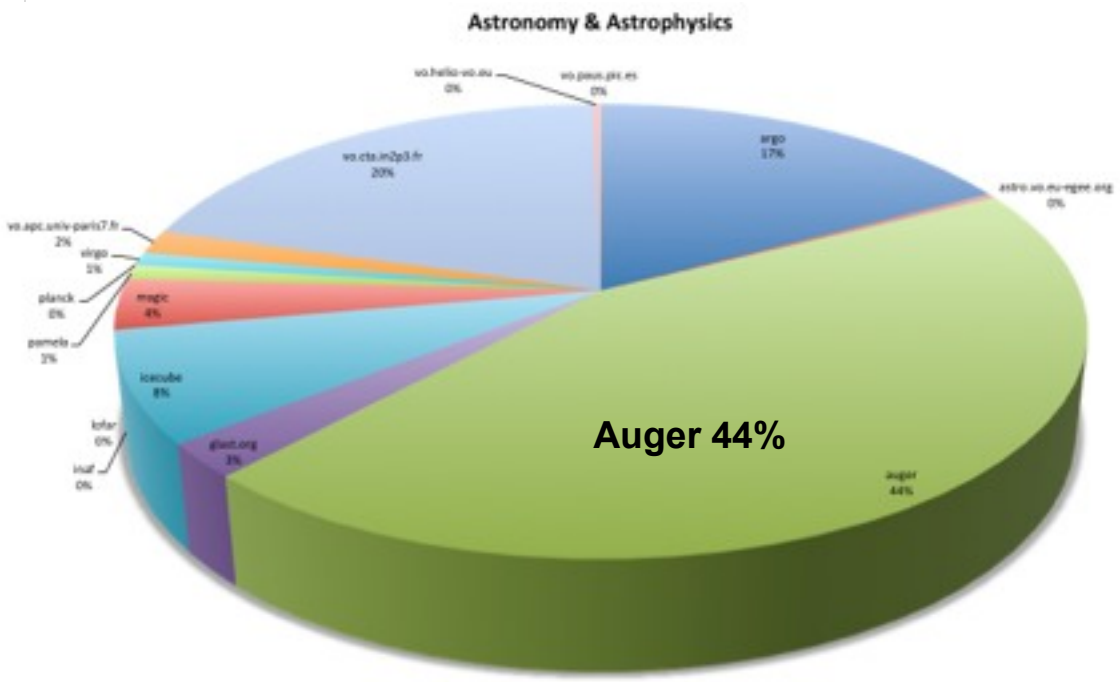
Use of European Grid Infrastructure



Job submissions: June 2010 - May 2011

Obtained using EGI statistics tool http://accounting.egi.eu/

Current approach



Existing European e-infrastructure not used by the SKA oriented community

Unexisting software infrastructure to enable the use at the level required to support radioastronomers

Different available resources: variety of policies for usage, interfaces for access, and a collection of programming models

Challenging to even the most sophisticated of users

THE REAL CHALLENGE STARTS HERE

Extraction of scientifically relevant information from huge volumes of data

- » Visualization of enormous catalogs into multidimensional parameter spaces
- » Efficient packaging of scientific methodology
- » Collaborative science

Transfer of knowledge to society

- » Friendly visualization tools
- » e-Science@school
- » Citizen science

Not only SKA, but EELT will face the same problem

Challenges in SKA and pre-SKA era

THE REAL CHALLENGE STARTS HERE

Extraction of scientifically relevant information from huge volumes of data

- » Visualization of enormous catalogs into multidimensional parameter spaces
- » Efficient packaging of scientific methodology
- » Collaborative science

Transfer of knowledge to society

- » Friendly visualization tools
- » e-Science@school
- » Citizen science

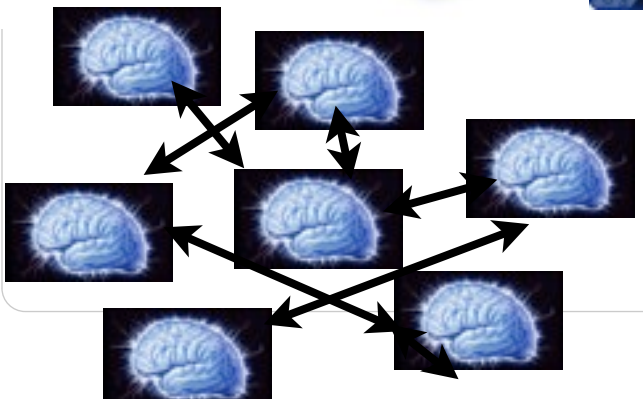
THE REAL CHALLENGE STARTS HERE

Extraction of scientifically relevant information from huge volumes of data

- Visualization of enormous catalogs into multidimensional parameter spaces
- Efficient packaging of scientific methodology
- Collaborative science

Transfer of knowledge to society

- Friendly visualization tools
- e-Science@school
- Citizen science



Extraction of scientifically relevant information from huge volumes of data

- Visualization of enormous catalogs into multidimensional parameter spaces
- Efficient packaging of scientific methodology
- Collaborative science

Transfer of knowledge to society

- Friendly visualization tools
- e-Science@school
- Citizen science

Towards e-Science solutions

- Strategy: build on pathfinders
- e-FALCONS & eSIRA collaborations

e-FALCONS: e-Infrastructure for FAst & Large Capacity Online Networked Systems

Data infrastructure for Large Data Volume Science Applications

e-SIRA: e-Science Infrastructure for Radio Astronomy

User oriented services for the SKA era

Towards e-Science solutions

University of Oxford, Coordinator

Science and Technology Facilities Council

University of Cambridge

The Netherlands Institute of Astronomy

National HPC and e-Science support centre

JIVE

Max Planck Gesellscaft zur Foerderung der Wissenschaften E.V.

Ludwig-Maximillians University, Munich

Forschungszentrum Juelich GMBH

Instituto de Astrofísica de Andalucía (CSIC)

Barcelona Supercomputing Centre

Fundación Centro Supercomputación Castilla y León

RedIRIS

National Institute for Astrophysics

Aalto University

European Grid Infrastructure

Delivery of Advanced Network Technology to Europe Limited

To lay the foundation for a scalable, expandable, modular, geographically distributed and network-connected, high-volume, high-speed data storage infrastructure that will address the needs of the community of LDVSAs such as LOFAR and SKA.

» **Sinergies:** infrastructure providers collaborate with radio astronomical organisations

EGI, PRACE, Géant, BiG Grid, Target, **GRID- IAA (CSIC)** + LOFAR

» **To develop:** **federating components (FCSL)** and provisioning of tools for efficient usage of the infrastructure and suitability for transport **(RedIRIS)** and archiving

» **Methodology:**

To identify main **bottlenecks** for high data volume apps, streamed or distributed Benchmarking architectures using existing science projects

- Processing of raw LOFAR correlator output
- Pulsar Surveys
- GLOSTAR EVLA Survey
- Prototype for kinematical modelling of extragalactic data cubes (IAA)

Goal: to enable radio astronomers to do **new science**

» **Methodology:**

» Knowledge exchange across the science and infrastructure communities

(Spain: BSC + IAA)

» Engage with the scientific community through

- A network of users provided by the SKA consortium + ASKAP + CyberSKA

- A set of user-driven tools that will enable astrophysicist to readily access data, processing and analysis. **Scientific workflows.**

- Training tools (adapted workflows)

» Access to both tools and data by students and citizens. **Educational workflows.**

Scientific Workflows are about sharing and collaboration

e-Science tool where data access and transformations explicitly interconnected through web services, **fully packaging the methodology of an experiment**

Scientific Workflows are about sharing and collaboration

e-Science tool where data access and transformations explicitly interconnected through web services, **fully packaging the methodology of an experiment**

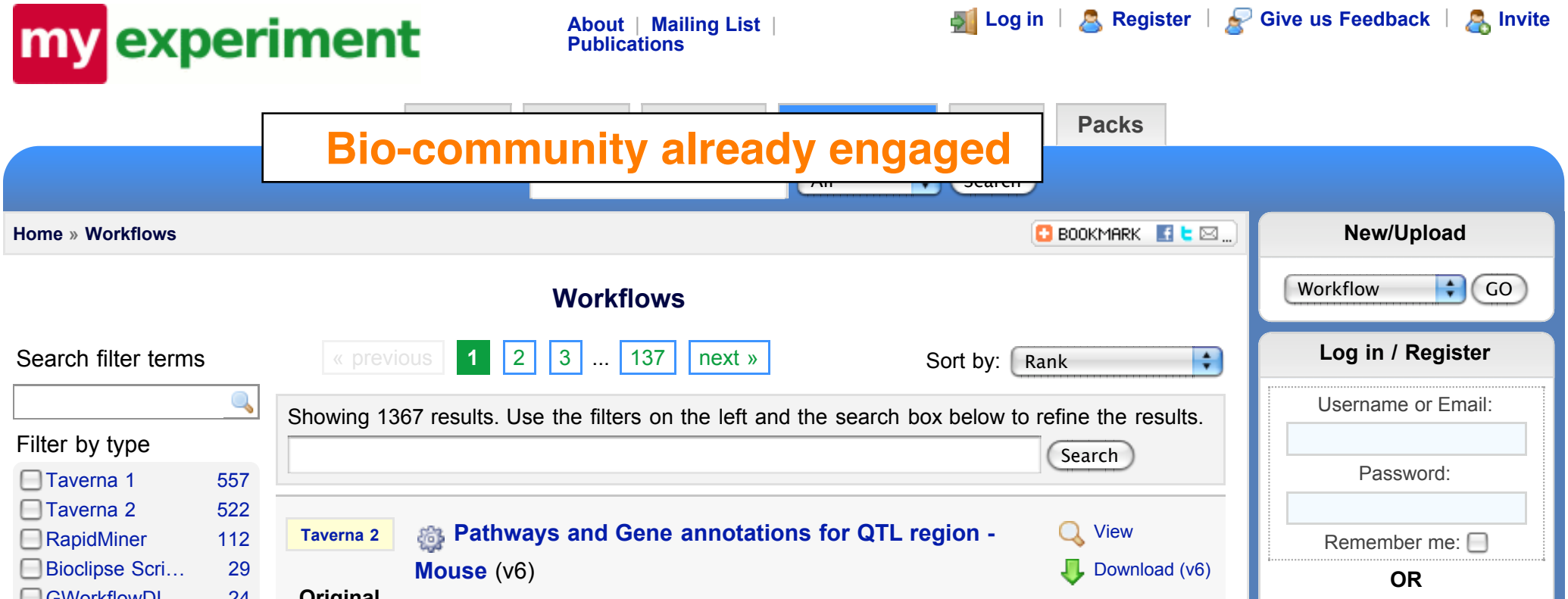
Modularity + capability to encapsulate methodologies, allow scientists

to create, reuse, and share them, e.g. through myExperiment (http://www.myexperiment.org/), a forum encouraging collaborative work.

Scientific Workflows are about sharing and collaboration

e-Science tool where data access and transformations explicitly interconnected through web services, **fully packaging the methodology of an experiment**

Modularity + capability to encapsulate methodologies, allow scientists **to create, reuse, and share them**, e.g. through myExperiment (http://www.myexperiment.org/), a forum encouraging collaborative work.



Workflows

Search filter terms

« previous 1 2 3 ... 137 next »

Sort by: Rank

Showing 1367 results. Use the filters on the left and the search box below to refine the results.

Search

Filter by type

- Taverna 1 557
- Taverna 2 522
- RapidMiner 112
- Bioclipse Scri… 29
- GWorkflowDL 24
- LONI Pipeline 20
- Kepler 17
- BioExtract Ser… 13
- Trident (Packa… 10
- Chemistry Plan 7

Filter by tag

- example 178
- mygrid 103
- localworker 99
- bioinformatics 89
- benchmarks 77
- cheminformatics 65
- protein 60
- ebi 56
- BLAST 49

Taverna 2

Pathways and Gene annotations for QTL region - Mouse (v6)

View

Download (v6)

Created: 19/11/09 @ 18:18:52 | **Last updated:** 05/04/11 @ 10:59:25

Credits: Paul Fisher

License: Creative Commons Attribution-Share Alike 3.0 Unported License

Original Uploader

Paul Fisher

This workflow searches for genes which reside in a QTL (Quantitative Trait Loci) region in the mouse, Mus musculus. The workflow requires an input of: a chromosome name or number; a QTL start base pair position; QTL end base pair position. Data is then extracted from BioMart to annotate each of the genes found in this region. The Entrez and UniProt identifiers are then sent to KEGG to obtain KEGG gene identifiers. The KEGG gene identifiers are then used to searcg for pathways in the KEGG path...

Rating: 4.5 / 5 (2 ratings) | **Versions:** 6 | **Reviews:** 0 | **Comments:** 4 | **Citations:** 1

Viewed: 1916 times | **Downloaded:** 400 times

example 178
mygrid 103
localworker 99
bioinformatics 89
benchmarks 77
cheminformatics 65
protein 60
ebi 56
BLAST 49
pathway 47

Filter by user

Alan Williams 210
Paul Fisher 89
Antoon Goderis 82
Peter Li 64
Hamish McWil… 52
Francois Belleau 43
Franck Tanoh 27
Andreas Hohe… 26
Anja Le Blanc 25
Egon Willigha… 24

Filter by licence

by-sa 895
by 273
by-nd 188
GPL 7
CC0 3
LGPL 1

Filter by group

myGrid 214
SabrOndexPr… 32
helio 26
MediGRID 22
TextGrid 21

position. Data is then extracted from BioMart to annotate each of the genes found in this region. The Entrez and UniProt identifiers are then sent to KEGG to obtain KEGG gene identifiers. The KEGG gene identifiers are then used to searcg for pathways in the KEGG path...

Rating: 4.5 / 5 (2 ratings) | **Versions:** 6 | **Reviews:** 0 | **Comments:** 4 | **Citations:** 1

Viewed: 1916 times | **Downloaded:** 400 times

Tags (13):

data-driven | disease | genotype | kegg | mouse | nbiconworkflows | pathway | pathway-driven | pathways | phenotype | qtl | shim | subworkflow

---

Taverna 1

Original Uploader

Marco Roos

BioAID_DiseaseDiscovery_RatHumanMouseUniprotFilter (v4)

View

Download (v4)

Created: 15/12/08 @ 20:46:09 | **Last updated:** 26/01/11 @ 14:43:31

Credits: Marco Roos AID

License: Creative Commons Attribution-Share Alike 3.0 Unported License

This workflow finds disease relevant to the query string via the following steps: 1. A user query: a list of terms or boolean query - look at the Apache Lucene project for all details. E.g.: (EZH2 OR "Enhancer of Zeste" +(mutation chromatin) -clinical); consider adding 'ProteinSynonymsToQuery' in front of the input if your query is a protein. 2. Retrieve documents: finds 'maximumNumberOfHits' relevant documents (abstract+title) based on query (the AIDA service inside is based on Apache's Luce...

Rating: 4.0 / 5 (2 ratings) | **Versions:** 4 | **Reviews:** 0 | **Comments:** 2 | **Citations:** 0

Viewed: 3390 times | **Downloaded:** 483 times

Tags (9):

AIDA | BioAID | biorange_nl | disease | pkna | protein | text_mining | text_mining_network | VL-e

Forgot Password?

Popular Tags

25 tags

[All Tags]

benchmarks | bio2rdf | bioinformatics | BLAST | cheminformatics | data integration | ebi | example | gene | graph | impact | kegg | Kegg Pathways | localworker | mygrid | ondex | pathway | pathways | phenotype | protein | pubmed | sequence | taverna | text mining | workflow

Workflow Entry: Pathways and Gene annotations for QTL region - Mouse

Created at: 19/11/09 @ 18:18:52 **Last updated:** 05/04/11 @ 10:59:25

| License | Credits (1) | Attributions (0) | Tags (13) | Featured in Packs (2) | Ratings (2) | Attributed By (6) | Favourited By (5) | | Citations (1) | Version History | Reviews (0) | Comments (4) |

Version 6 (latest) (of 6)

View version: 6 (latest)

Version created on: 05/04/11 @ 10:59:23 **by:** Paul Fisher | Revision comments

Title: Pathways and Gene annotations for QTL region - Mouse

Type: Taverna 2

Preview

(Click on the image to get the full size)

Workflow Type

Taverna 2

Original Uploader

Paul Fisher

License

All versions of this Workflow

end position = 29500000

Taverna

Monday 25 October 2010 @ 16:23:46 (BST)

I like this workflow

Kawther

Thursday 14 April 2011 @ 18:00:32 (BST)

I could not run this workflow in Taverna, the error message Faile to oppen workflow name: doesn't appear to be a workflow!

Other workflows work fine

Where do u think the problem?

Linked Data

What is this?

Non-Information Resource URI: http://www.myexperiment.org/workflows/16

Alternative Formats

» **State of the art:**

Most widespread methodology:

- General-purpose software + specific tools developed within a single research group

---> reinvention and effort duplication

Does not scale with complexity level + size of upcoming data, nor with computing infrastructures involved.

- Very few cases of standardization of the methodology. Those existing, developed just for specific parts of the full discovery process (e.g pipelines).

Data:

- Digital sky
- Data formats and access protocols standardized in astronomy via the VO

PDF publication: Bibliographical archives

» ADS provides interlinking to Astronomical Objects DB

» Vizier provides interlinking to Bibliographic Archives

» **What is missing?**

The astronomical workflows:

Helio project on-going:

virtual observatory for solar physics: data access + sharing + description of the knowledge in the field (via ontologies), and their processes (via workflows).

Working environment: Preservation

development of standards for workflow preservation, which will enable

workflow classification, indexing, and inspection of used and generated data

Workflows for ever Project (started December 2010)

Advanced Workflow Preservation Technologies for Enhanced Science

Funded under FP7 ICT-2009-6

PI: iSOCO

Participants: UPMadrid, U. Manchester, Poznan PSNC, U. Oxford, Leiden U. Medical Center, IAA-CSIC

Wf4ever in one slide: Preservation of Scientific Workflows

- novel definition of a Research Object, which **packages** workflow descriptions, provenance of their executions, and links to all the related resources

- models for **repeatability and reproducibility**

- models for workflow abstraction, to facilitate workflow **classification and indexing,** comparison

- strategies for **sharing and reusing** workflows or fragments

- mechanisms for personalized workflow **recommendation**

- methods and tools to proactively **evaluate** workflow information quality

Use cases: Astronomy and Genomics

kinematical modelling of extragalactic data cubes

VO Scientific Workflow

3D Kinematical Modeling

Source Name

Sesame NarreResolver

Coordinates

Raw Data

CASA Calibration

HIALMA 3D VO Archive Calibrated Data

VO Catalog

CO ALMA 3D VO Archive Calibrated Data

3D VO Archive Modeled Data

HI Datacube

Center, Inclinations, etc.

CO Datacube

Gipsy 3D Modeling

VO Publishing

3D HI Residuals

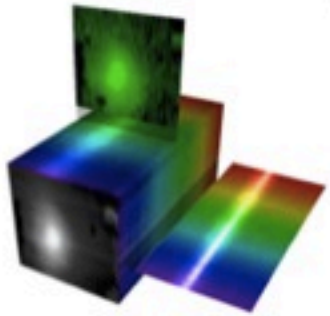
Subtraction

Subtraction

3D CO Residuals

Modeled HI Datacube

Modeled CO Datacube



Rotation Curve Extraction

CO Rotation Curve

CO Rotation Curve

CO Dark Matter Model

CO Dark Matter Model

Current developments of AMIGA group

AMIGA4GAS: **AMIGA for GTC, ALMA, and SKA pathfinders**

Proposed to AYA National funding (2012-2014), coordinated with FCSCL

Science goal:

driving mechanisms of secular evolution via high resolution 3D studies of isolated galaxies

Technological e-Science developments:

- Migration to the GRID and Cloud Computing of tools for modelling datacubes
  - Development of IVOA standards for VO services on 3D data
    - Deployment of workflows on distributed infrastructure of heterogeneous resources

Collaboration with:

- CATON private company
- Barcelona Supercomputing Center
- IberGrid initiative: Instituto de Telecomunicaçoes in Portugal
- IAA-CSIC Grid technical group
- TarGet project, which covers the data processing needs for the LOFAR array

Integrated into the umbrella of PrepSKA

The time is NOW since:

- LOFAR and EVLA working now
  - ASKAP: expected to be completed by 2013 (coll. with “WALLABY” project)
    - MeerKAT: expected to start operations in 2014 (particip. in “MHONGOOSE”)
      - Apertif@WSRT: 2013 commissioning and 2014-2017 scientific exploitation

      (member of Science Team)

Contracts to be placed soon

2012-2015 Pre-construcion Phase