ASTERICS - H2020 - 653477

First ASTERICS-OBELICS workshop and training event

ASTERICS GA DELIVERABLE: D3.2 and D3.6

Document identifier:	ASTERICS-D3.6.doc
Date:	**23 June 2017**
Work Package:	**WP3 OBELICS**
Lead Partner:	**LAPP**
Document Status:	**Report**
Dissemination level:	Public
Document Link:	www.asterics2020.eu/documents/ ASTERICS-D3.6.pdf

Abstract

This report provides a global overview of the presentations and discussions that took place at the 1st ASTERICS – OBELICS workshop on Science Data Cloud & Computing models in Astronomy and Astroparticle Physics, in Rome from 12-14 December 2016. This workshop was a result of merging two deliverables namely WP3 training event (deliverable D3.2) and the first WP3 general workshop (deliverable D3.6). The workshop addressed the user engagement, dissemination as well as networking objectives of the WP3-OBELICS, thanks to participants from H2020 projects, industries, consortia and academia.

I.   COPYRIGHT NOTICE

II.   DELIVERY SLIP

	Name	Partner/WP	Date
From	Giovanni Lamanna	LAPP	
Author(s)	Jayesh Wagh	LAPP	07-03-2017
Reviewed by	Rob van der Meer	ASTRON	10-05-2015
Approved by	AMST		22-06-2017

III.   DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1	07-03-2017	First Draft	Jayesh Wagh, LAPP
2	10-05-2017	second draft	Rob van der Meer, ASTRON
3	18-05-2017	final draft	Jayesh Wagh, LAPP
4	02-06-2017	almost final version	Rob van der Meer, ASTRON
5	21-06-2017	revised final version	Jayesh Wagh, LAPP
6	23-06-2017	final version	Rob van der Meer, ASTRON

IV.   APPLICATON AREA

This document is a formal deliverable for the GA of the project, applicable to all members of the ASTERICS project, beneficiaries and third parties, as well as its collaborating projects.

V.   TERMINOLOGY

ASTERICS	Astronomy ESFRI & Research Infrastructure Cluster
CTA	Cherenkov Telescope Array
ESFRI	European Strategy Forum on Research Infrastructures
KM3NeT	Cubic Kilometre Neutrino Telescope
LOFAR	The Low Frequency Array
LSST	The Large Synoptic Survey Telescope
OBELICS	Observatory E-environments LInked by common ChallengeS
SKA	The Square Kilometre Array
VO	Virtual Observatory

A complete project glossary is provided at: http://www.asterics2020.eu/glossary/.

VI.   PROJECT SUMMARY

ASTERICS (Astronomy ESFRI & Research Infrastructure Cluster) aims to address the cross-cutting synergies and common challenges shared by the various Astronomy ESFRI facilities (SKA, CTA, KM3Net & E-ELT). It brings together for the first time, the astronomy, astrophysics and particle astrophysics communities, in addition to other related research infrastructures. The major objectives of ASTERICS are to support and accelerate the implementation of the ESFRI telescopes, to enhance their performance beyond the current state-of-the-art, and to see them interoperate as an integrated, multi-wavelength and multi-messenger facility. An important focal point is the management, processing and scientific exploitation of the huge datasets the ESFRI facilities will generate. ASTERICS will seek solutions to these problems outside of the traditional channels by directly engaging and collaborating with industry and specialised SMEs. The various ESFRI pathfinders and precursors will present the perfect proving ground for new methodologies and prototype systems. In addition, ASTERICS will enable astronomers from across the member states to have broad access to the reduced data products of the ESFRI telescopes via a seamless interface to the Virtual Observatory framework. This will massively increase the scientific impact of the telescopes, and greatly

encourage use (and re-use) of the data in new and novel ways, typically not foreseen in the original proposals. By demonstrating cross-facility synchronicity, and by harmonising various policy aspects, ASTERICS will realise a distributed and interoperable approach that ushers in a new multi-messenger era for astronomy. Through an active dissemination programme, including direct engagement with all relevant stakeholders, and via the development of citizen scientist mass participation experiments, ASTERICS has the ambition to be a flagship for the scientific, industrial and societal impact ESFRI projects can deliver.

VII. EXECUTIVE SUMMARY

The 1st ASTERICS – OBELICS workshop on Science Data Cloud & Computing models in Astronomy and Astroparticle Physics addressed high priority subjects of ESFRI projects such as Authorization and Authentication, Data Storage, Transfer & Preservation, Large Databases, Workflow management and Interoperability. The format of the event included invited talks, panel discussions, live demonstrations, and tutorials on services and software that are aimed to be explored and evaluated for likely being used in the ASTERICS ESFRI Projects. Thanks to the participation of over 87 registered attendees from ASTERICS partner institutes, partner H2020 projects, EU consortia as well as major industries, all the workshop sessions were addressed from different perspectives. A call for expression of interest for industrial cooperation was launched to benefit from industrial presence at the workshop. Therefore, this workshop helped WP3 OBELICS achieve its objectives of dissemination, user engagement and training as well as establishing contacts with industries to address innovation related goals.

Table of Contents

1. Introduction

In the project proposal, we estimated in Task 3.1 MAUD we could organize the first main thematic WP3 training event in M12 as deliverable D3.2 and the first WP3 general workshop in M18 as D3.6. At the same time, it was expected that OBELICS partners would dedicate most of the first year in hiring and building up the teams. In order to identify important subjects for Astronomy ESFRI and other world class projects for this workshop, multiple face to face and telecom meetings with OBELICS members, experts from other H2020 projects and industries were organized to understand their potential contribution and participation to the workshop programme. Due to this exhaustive organizational approach, the combined training event (D3.2) and General Workshop (D3.6) was organized on 12 – 14 December 2016 (M20) in Rome. This workshop entitled 1st ASTERICS – OBELICS workshop addressing Science Data Cloud & Computing models in Astronomy and Astroparticle Physics was organized at Casa I CAPPUCCINI, Rome, Italy. The workshop was organized by the French CNRS LAPP (Laboratoire d'Annecy-le-Vieux de Physique de Particules) laboratory, the leading OBELICS institute and hosted locally by the Italian INAF Rome institute, one major partner of the project.

2. Structure of the meeting

The reason for the workshop was to bring together many parties and persons interested in the ASTERICS goals, in particular the OBELICS scope. We wanted people that could teach others, people that could bring new insights to the group, people that are eager to learn new things to bring their research (infrastructure) to a higher level and people that are interested to learn new people to extend their network and therefore their knowledge.

The workshop had a multifold purpose:
- Dissemination of research activities under OBELICS.
- Address technical challenges of ASTERICS ESFRI projects through the series of talks, tutorials with live demos and panel discussions involving other H2020 projects, e-infrastructures, EU consortia as well as industries.
- Understand the complementarity with major industries to address the innovation related objectives of OBELICS.
- Building mutual engagements between ASTERICS consortium and e-infrastructures.

The meeting was built on three types of interaction:
- invited talks, presenting a wide and complete overview of the subject
- tutorials, presenting a very specific subject, with a step by step walkthrough and demonstration

- panel discussion; At the end of each session we had a panel discussion to see if the audience got enough information and to discuss if we are moving in the right direction with this subject. The panel consisted of the speakers of the session.

During this three-day workshop, 7 tutorials with live demos, 20 invited talks and 5 panel discussions addressed five broad themes:
- Authorization and Authentication
- Data Storage, Transfer & Preservation
- Large Databases
- Workflow management
- Interoperability

The meeting started on the first day around noon to allow for travel in the morning. After the introductory presentations, one theme was covered in the afternoon. The other two days had one theme in the morning and one in the afternoon. At the end of the third day there was a short concluding overview.

The detailed programme of the workshop can be found on the meeting website, https://indico.astron.nl/conferenceDisplay.py?confId=47 and in appendix 1 of this document.

All of the sessions were integrally recorded and reviewed to make sure all interesting remarks and discussions would be preserved for future reference.

A detailed description of all presentations can be found in appendix 2.
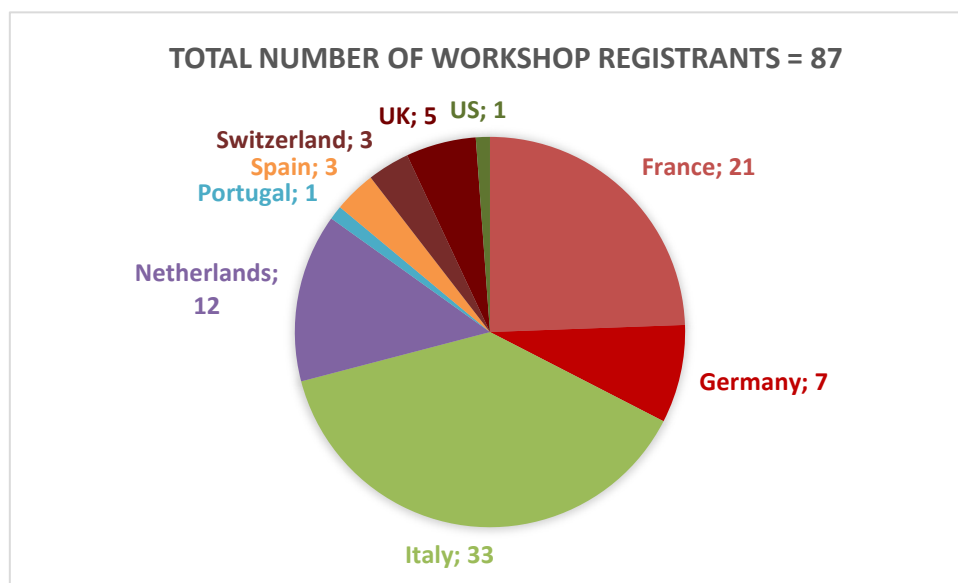A summary of highlights and results of the meeting will be given in section 4.

Call for expressions of interest for industrial cooperation

On this occasion, a call for expressions of interest for industrial cooperation was launched, to encourage the partnering process between ASTERICS partners and the industries on innovation-related objectives of OBELICS work package.

3. Participation to the meeting

The workshop brought together 87 scientist and technologist from ASTERICS partner institutions, H2020 projects such as IndigoDataCloud, HNSciClound, AARC2, industries such as ATOS, NVIDIA, ORACLE Europe, IBM, E4 and consortia such as EU-T0, APPEC, RDA, and EGI. In addition, experts from CERN as well as LIGO Laboratory, Caltech, USA were amongst the participants. The complete list of participants, speakers and the presentations are available on workshop website https://indico.astron.nl/conferenceDisplay.py?confId=47

TOTAL NUMBER OF WORKSHOP REGISTRANTS = 87

UK; 5   US; 1
Switzerland; 3
Spain; 3
Portugal; 1
France; 21
Netherlands; 12
Germany; 7
Italy; 33

Workshop participants by country

Although we invited people from all partners and independent industry, the distribution over countries in Europe was a bit biased towards the host and organising countries. We had 5 participants from non-partner countries.

4. Results of the meeting

Overall, the workshop was extremely well received by the scientific community with active participation in discussions, talks and tutorial sessions. The event also produced significant impact on social media outreach of ASTERICS with 6450 tweet impressions.

Detailed information on all the presentations can be found in appendix 2

The panel discussion under Authentication and Authorization session discussed that EGI, EUDAT and INDIGO-DataCloud are plannig to prepare a common bid for the H2020 INFRADEV-12 call to improve the e-infrastructure provided, streamlining duplications and improving sustainability. The proposal will include a number of use cases and pilots in order to involve the various scientific domains, and ASTERICS members are most welcomed to participate in this process. In addition, some of the guidelines discussed during this session will also be introduced in the skeleton of best practices document, which could be included in upcoming OBELICS deliverables.

Some of the tutorials and demos organized under Large Databases session as a part of training activities during the workshop can be potentially implemented in ESFRI experiments ex. Qserv can have potential applications under LOFAR and SKA. QServ applications will also be included in upcoming OBELICS deliverables. The tutorial and demonstrations gave participants a detailed overview of QServ architecture, MonetDB database systems and Apache Spark computing framework.

The event allowed ASTERICS members to network and share ideas with the industrial participants from NVIDIA, IBM, ATOS, ORACLE and E4. This networking was further supported by Call for expression of interest for industrial cooperation on innovation related objectives of OBELICS workpackage. The workshop had discussions on synergies in the future with new started H2020 AENEAS, and EOSCpilot projects building up together with OBELICS a cooperative ecosystem of e-infrastructures for Big Data scientific projects. Discussions on opportunities for cooperation with H2020-INDIGODataCloud and EGI for the H2020 programme on "OneData" service was one of the outcome of the workshop. The workshop participants included directors from the European data centres such as CNAF, PIC and CC-IN2P3. Their participation in the discussions helped the ESFRI project members from CTA, EUCLID and Virgo to further strengthen their collaborative ties with the data centres.

Interoperability session addressed structuring the collaboration with United States in the domain of Gravitational Waves observation under interoperability, high performance computing, high throughput computing as well as parallel programming. It discussed to invite tutors from US to deliver tutorials during the second main thematic event under OBELICS, as well as opening up the attendance of the school to concerned US scientists for LSST and Ligo. As a follow, up of the discussions and inputs from the workshop, the second main thematic

event will be organized for students, postdocs, and senior researchers from the domain of astrophysics and astroparticle physics to provide theoretical and hands-on training to acquire efficient and fast computer programming techniques, as well as skills for improving their scientific data analysis software.

5. Evaluation of the meeting

As expected, the merger of two events had a positive impact on the project activities in terms of knowledge sharing, training and networking with industries, partner H2020 projects as well as EU consortia.

The programme included dedicated sessions for the training objectives and other sessions for knowledge sharing objectives fitting in the general workshop. We attracted participants from diverse expertise and expectations, ones that had more interest in the training, but now learned about the other subjects as well and ones that were more interested in the collaboration aspect, who learned some things in detail they would otherwise not have seen. There was also very good interaction between the participants.

This organization approach including tutorials and demo sessions was particularly beneficial where the number of subjects that were evolved enough to discuss them in larger groups or provide trainings. We intend to keep the similar format for the upcoming general workshop in the month of the October 2017.

As an output of this workshop we have managed to reach audience outside the collaboration and this dissemination level will be further improved with upcoming thematic training events as well as general workshops. There are two more trainings and two more general workshops planned. These will be organized separately from each other. A post workshop follow up with the industries did result in collaboration with NVIDIA and E4 submitting two OBELICS expressions of interest proposals for industrial cooperation. The MoUs for this industrial collaboration are already in drafting stages. The post workshop dissemination of summaries of the workshop discussions through website and social media would be improved further in upcoming events.

Appendix 1.   Programme

Day 1: 12 December 2016

11h30-12h15: Registration.
12h15-12h25: Welcome Address by Lucio Angelo Antonelli, INAF Osservatorio Astronomico di Roma.
12h25-12h50: OBELICS & Objectives of the workshop by Giovanni Lamanna, LAPP/IN2P3/CNRS.

12h50-13h50: Lunch Break

13h50-14h00: H2020- Astronomy ESFRI and Research Infrastructure Cluster (ASTERICS) by Marco De Vos, ASTRON.

Authentication & Authorization Session

Chair of the session: Fabio Pasian, INAF.
Time: 14h00-18h20

14h00-14h20: CTA Data Management Challenges.
Presenter: Nadine Neyroud, (CTA-LAPP, Technical Director)
14h25-14h55: A&A in CTA: from User Requirements towards a Research Infrastructure.
Presenter: Eva SCIACCA (Researcher, INAF), Alessandro COSTA (Senior Technologist and Computer Scientist at INAF)
15h00-15h20: H2020 AARC2: Community Driven Developments in the Identity Space.
Presenter: Licia FLORIO, AARC Project Coordinator- GEANT.
15h25-15h55: H2020-INDIGO Datacloud achievements: How Indigo DataCloud broker identities and do authentication and authorization.
Presenter: Andrea CECCANTI (INFN-CNAF).

16h00-16h20: Coffee Break

16h20-16h40: LOFAR Presentation
Presenter: Hanno HOLTIES, ASTRON.
16h45-17h15: EGI A&A Demo.
Presenter: Mario DAVID LIP Portugal.
17h20-18h00: Panel Discussion - How to harmonise A&A mechanisms across the world-class projects
Chair: Fabio Pasian, Head, Astrophysical Technologies Group INAF
Panelists: Nadine Neyroud, Eva Sciacca, Alessandro Costa, Licia Florio, Andrea Ceccanti, Hanno Holties, Mario David, Cristina Knapic, Giuliano Taffoni.

Day 2: 13 December 2016

Data Storage, Transfer & Data Preservation
Chair of the Session: Paul Alexander, UCAM
Time: 9h00-13h00

9h00-9h20: ESFRI Project Presentation - SKA.
Presenter: Paul ALEXANDER (Head of Astrophysics, Cavendish Laboratory, University of Cambridge)
9h25-9h45: Oracle Long-Term Storage Solutions.
Presenter: Philippe DEVERCHERE; (Oracle EMEA Storage CTO)
9h50-10h10: Private and public Cloud experiences at CERN.
Presenter: Domenico GIORDANO (CERN)

10h15-10h35: Coffee Break

10h35-10h55: Data Management at ESO.
Presenter: Michael F. STERZIK (Head of Data Management Operations, ESO)
11h00-11h20: The Research Data Alliance (RDA) on practical policies.
Presenter: Raphael RITZ (Max Planck Computing and Data Facility)
11h25-11h45: A globally distributed data management solution.
Presenter: Oliver KEEBLE (data management group CERN IT department)
11h50-13h00: Panel Discussion.
Chair**:** Michael WISE (Head of the Astronomy Group at ASTRON)
Panelists: Paul Alexander, Philippe Deverchere, Domenico Giordano, Michael F. Sterzik, Raphael Ritz, Oliver Keeble

13h00-14h00: Lunch Break

Large Databases
Chair of the Session: Volker GUELZOW, DESY
Time: 14h00-18h00

14h00-14h20: LSST Presentation.
Presenter: Dominique BOUTIGNY (LSST-France Principal Investigator, LAPP)
14h25-14h55: Tutorial: MonetDB in the context of the new (high-cadence) facilities.
Presenter: Bart SCHEERS (CWI Amsterdam)

15h00-15h30: QServ Presentation & Demo.
Presenter: Fabrice JAMMES (LPC IN2P3 Clermont Ferrand)

15h35-15h55: Tea-Coffee Break

15h55-16h25: SPARK Demo & Presentation.
Presenter: Andre SCHAFF (Strasbourg astronomical Data Center (CDS))
16h30-17h30: Panel Discussion.
Chair: Volker Guelzow, Head of IT, DESY.
Panelist: Dominique BOUTIGNY, Bart SCHEERS, Fabrice JAMMES, Andre SCHAFF

20h00-22h00: Social Dinner

Day 3: 14 December 2016

Workflow management & system preservation
Chair of the Session: Manuel Delphino, Director of the Port d'Informació Científica (PIC)
Time: 9h00-13h00

9h00-9h20: EUCLID Presentation
Presenter: Marc SAUVAGE (Astrophysicist, CEA, Saclay, France)
9h25-9h45: H2020 Indigo DataCloud presentation
Presenter: Davide SALOMONI (Technology Director INFN, INDIGO-DataCloud coordinator)
9h50-10h20: Tutorial: DIRAC Systems presentation
Presenter: Luisa ARRABITO (Computer Engineer at LUPM IN2P3/CNRS)

10h25-10h50: Tea-Coffee Break

10h50-11h10: ESFRI Presentation - KM3NeT
Presenter: Kay GRAF (Friedrich-Alexander - Universität Erlangen-Nürnberg)
11h15-11h35: Industry Presentation - ATOS
Presenter: Ana Juan FERRER (Head of NG Cloud Lab, Atos – Research and Innovation).
11h40-12h00: H2020- HNSciCloud Presentation
Presenter: Robert JONES (Coordinator of the HNSciCloud Horizon 2020 Pre-Commercial
Procurement project (http://www.hnscicloud.eu/)
12h00-13h00: Panel Discussion
Chair: Manuel DELFINO (Director of the Port d'Informació Científica (PIC).)
Panelists: Marc SAUVAGE, Davide SALOMONI, Luisa ARRABITO, Kay GRAF, Ana Juan FERRER,
Robert JONES.

13h00-14h00: Lunch Break

Interoperability Session: How to efficiently handle heterogeneous architectures (on distributed data centers)
Chair of the Session: Giovanni Lamanna, LAPP
Time: 14h00-18h00

14h00-14h20: VIRGO Presentation
Presenter: Michele PUNTURO (INFN)
14h25-14h45: LIGO-VIRGO Collaboration Presentation
Presenter: Peter COUVARES (LIGO Laboratory, US)
14h50-15h10: E4 experience and expertise in HPC
Presenter: Daniele GREGORI (E4)
15h15-15h35: NVIDIA Presentation
Presenter: Piero ALTOE, NVIDIA
15h40-16h00: Low-Power computing with ASTRI & CTA use cases
Presenter**:** Denis BASTIERI (Padova University & INAF)

16h05-16h25 Tea-Coffee Break

16h25-16h45: IBM Presentation
Presenter: Giorgio RICHELLI (IBM)
16h50-17h30: Panel Discussion
Chair: Giovanni Lamanna (LAPP)
Panelist: Dr. Daniele GREGORI, Dr. Giorgio RICHELLI, Dr. Denis BASTIERI, Dr. Piero ALTOE, Dr. Peter COUVARES, Dr. Michele PUNTURO.
17h30-17h50: Conclusion
Speaker: Giovanni Lamanna, LAPP.

Appendix 2.   Presentations, Tutorials and Discussions

a. Welcome Address

Angelo Antonelli welcomed all the workshop participants. He also gave an overview of research activities at INAF under the CTA and OBELICS context. Marco de Vos discussed the ASTERICS project and its objectives. Giovanni Lamanna gave a brief overview of ASTERICS project with more focus on research activities under OBELICS workpackage. He briefly introduced various ESFRI projects and world-class infrastructures involved in ASTERICS. He also described the objectives of the workshop. The workshop was aimed to explore existing solutions and discuss potential engagements/partnerships with Industries, Consortia, and H2020 e-projects.

b. Authentication and Authorization Session

The afternoon session began with presentation on CTA Data Management Challenges by Nadine Neyroud, Technical Director CTA-LAPP. She noted that CTA follows IVOA standards, but mainly for data discovery and VO-compatibility of the archive, which is expected to grow at a rate of 27 PB/y (6 on disk, 21 on tape). Given the data rights policy, the CTAO for A&A plans to use instead a centralised and coherent system connecting users through an Observatory Science Gateway supporting single sign-on (SSO) or a CTA Observatory application. Tools under evaluation for A&A are Grouper and Unity. Several challenges for the CTA central A&A system were listed.

Alessandro Costa and Eva Sciacca demonstrated one of the CTA prototypes. An eduGAIN authentication is matched with a Grouper-based authorisation to obtain a role-based authorisation system (when a user or group is added to a role, the user or group automatically inherits the various security permissions). CTA has been chosen as a use case for the AARC2 project funded by H2020.

Licia Florio remarked the role of AARC and its approach to solving the projects' A&A needs. Building on existing tools and framework, the goal is avoiding new research collaborations to develop independent AAIs. A proxy model is the chosen approach, at the moment the only way to provide SSO. An AARC Blueprint is available: RI can implement it, using a set of AARC building blocks on top of eduGAIN, which is the foundation for federated access. CILogon was mentioned as a token translations service. AARC is as counsellor, rather than a provider; it can provide to projects: a) policies and best practices for proxy operators, b) recommendations/support to RIs on assurance, data protections and security incidents, c) guideline documents on group membership, non-web access, authorisation, d) training (bilateral ad-hoc meetings with communities can be arranged). The CTA pilot in AARC2 is a good example of this approach.

Andrea Ceccanti gave a tutorial (2 demos) on the A&A mechanisms provided by the INDIGO-DataCloud project: the system is providing support for federated authentication & social

logins, and orthogonally for an authorisation attribute-based, dynamic and consistent across heterogeneous infrastructures. Key concepts are delegation (services act on behalf of a user) and token translation (integration of services relying on heterogeneous authentication mechanisms), The Authentication Identity Access Management (IAM) allows identities based on SAML, X.509 and OIDC to be translated in OpenID-connect (standard adopted by industry, mobile friendly) to be used internally to INDIGO.

Hanno Holties presented the current status of LOFAR. The Long-Term Archive keeps growing at the rate of 7 PB/y. For A&A, X.509 certificates (both personal and robot certificates) are used. Services allow anonymous data queries, while data retrieval require authentication and role-based authorisation. Auditing is also implemented (auditability may be just as important a driver as A&A for data access). Experience has shown that personal certificates are not designed for large communities, while robot certificates are not practical for complex authorisation models.

Mario David demonstrated the EGI AAI Check-in service that, implementing the AARC blueprint, provides a federated IAM solution (eduGAIN, certificates, social including Google and ORCID) which passes a unique user identity to the services provided by EGI or others. A key point is guaranteeing the trust chain between Users and Identity Providers (IdPs) and Service Providers (SPs) and vice-versa. Different Levels of Assurance (LoA) can be used depending on the risk level accepted by the various services (the minimum LoA required is determined by the user community and the service provider requirements): EGI supports four LoAs. The EGI Checkin service generates the EGI User ID which is: personal, persistent, non-reassignable, non-targeted, globally unique, opaque. A Credential Translation service is supported.

Panel Discussion on "How to harmonise A&A mechanisms across the world-class projects" began with two short talks by Cristina Knapic,SKA and Raphael Ritz, RDA. Cristina Knapic noted that in SKA the A&A(&A) requirements refer to an enterprise solution (i.e. valid for all activities of the whole Consortium. SKA will use a 3-step approach: a) implement a SERVICE provider for SKA to authenticate identities already present in EduGAIN, b) implement an identity provider for SKA in order to manage identities inside the SKA, c) support other technologies for AIM (authentication interface management). Interoperability within the Virtual Observatory is a requirement: GMS, currently being discussed as an IVOA standard, could be a solution for SKA authorisation.

Raphael Ritz summarised B2ACCESS, the A&A framework in EUDAT which allows SSO via the users' primary identity providers (including Google, Microsoft and Facebook) and Integrated with EduGain, allowing interoperability among different protocols (SAML, OAuth2 and X.509). Developed on top of Unity, it allows service managers to specify authorisation decisions.

The discussion was focused on the key interest in the ASTERICS project, i.e. to identify a way to harmonise A&A mechanisms across the world-class astronomical and astroparticle projects. As a basis for the discussion, Fabio Pasian reported about a meeting held during the ADASS

XXVI conference (Trieste, 18 October), in which representatives from Europe and Canada (involved in CTA, SKA, ESAC, CADC, IVOA, ASTERICS + some data centres) reached agreement on the following approach: a) authentication should be clearly separated from authorisation for every project, b) federated authentication across projects (EduGain-like) seems to be acceptable for everyone, c) intra-project federation of authorisation across partners is a topic to be investigated (e.g. VOSpace CADC-INAF experiment in EGI-Engage, to be extended to other partners using GMS), d) share information on A&A activities.

The international breadth of projects like SKA and CTA, and of the Virtual Observatory initiative, was reminded: solutions proposed from the European partners in these projects need to be accepted by a wider community. The IVOA standards are a good example of working international collaborations, as they are a unifying/interoperability layer on top of the various technical solutions chosen. SSO and Credential Delegation have been IVOA standards for quite a while, and GMS is expected to become a standard within 2017.

A discussion took place on the identification of the weak points in the trust mechanism. A possible approach is the adoption of the Levels of Assurance (LoA) mechanism: not always a high LoA is needed, and where the trust and security need to be maximum (e.g. for access to sensitive info), the LoA would be maximum as well (C, in the EGI scale). On the average, eduGAIN (A) would be good for most cases in our communities, while the LoA provided by social media (0) would be sufficient for public info.

The duplication of A&A solutions among e-infrastructure providers (EGI, EUDAT, INDIGO-DataCloud) was noted. It was also noted that project in the interest of ASTERICS could take advantage of the guidance that can be provided by AARC (now by AARC2). A discussion followed on what can happen when H2020 projects end, since they last less than the big projects (e.g. ESFRIs, space missions). E-infrastructures need to provide sustainability for the core products built within the H2020 projects. It was finally mentioned that EGI, EUDAT and INDIGO-DataCloud plan to prepare a common bid for the H2020 INFRADEV-12 call, and this is expected to improve the e-infrastructure provided, streamlining duplications and improving sustainability. The proposal will include a number of use cases and pilots in order to involve the various scientific domains, and the astrophysics and astroparticle communities are warmly invited to participate in this process. Moreover, the database should keep track of and synchronize all changes applied to raw data during the analysis chain (DL0 to DL5) and conserve information about the data provenance. Currently, no official estimation on the size or query rates of such a database are available. Here again, insights from ASTERICS and other ESFRI projects might be extremely valuable.

c. Data Storage, Transfer & Data Preservation Session

Second day of the workshop started with a talk on SKA data challenges by Paul Alexander. In the first phase (2020) 5000 PB/day will be transferred from SKA antennas to DSP and by 2030 this figure will reach to 100 000 PB/day. Therefore, Data management is one of the big challenges for SKA. The SKA data has to be buffered during the entire time of the observation

which is about 12 hours. The high-performance buffer of SKA is about 85 PBytes in size and is being realized through Solid state drives. SKA regional centres will be following the CERN model for managing data products up to 1PB/day. In SKA phase 2 (Complete SKA) the data rates and processing are expected to increase by a factor of 100 i.e 3-30 EBytes/ year of fully processed data for SKA2.

Philippe Deverchere presented ORACLE long-term storage solutions. A general overview of near future data explosions from the scientific domain, social media and aviation industry were presented. Oracle addresses those challenges by providing differentiated technologies and solutions that go from enterprise-grade tape to massively scalable storage cloud services, enabling long term preservation and distribution/sharing of scientific data through a variety of hybrid architectures. General industry trends for data storage with Cloud, disks and Tape were presented. Cloud can provide the end-user the cost effectiveness of tape without requiring the expertise. Philippe noted that Oracle Storage Cloud service provides the lowest cost per gigabyte in the industry.

Domenico Giordano described the history, the architecture, the tools and the technical decisions behind the CERN OpenStack Cloud Infrastructure. OpenStack is one of the largest open source communities with contribution from over 2300 developers. The contributors also involved industries such as IBM, HP, and Red Hat. 90% of CERN's computational resources are now delivered on top of OpenStack. OpenStack infrastructure at CERN has been operational for three years to deliver LHC compute capacity.

Michael F. Sterzik explained the strategies and challenges in scientific data management for the European Southern Observatory (ESO). For ESO the main challenge is develop its own tools and operate them. Data archiving infrastructure cost is smaller compared to the content management.Data. ESO welcomes more interaction from H2020 project.

Raphael Ritz presented the templates of common policies as well as implementation examples of the policies in different systems illustrating best practices that were put together by working group of Research Data Alliance (RDA). RDA has 46 organizational and 6 affiliate members. It has 4672 individual members to date from diverse background. The next RDA plenary meeting will take place in Barcelona. The plenary meetings allow participants to exchange knowledge, information and potential solutions.

Oliver Keeble presented Worldwide LHC Computing Grid (WLCG) of CERN, focusing on data management and discussing the potential for reuse of its resources, components and concepts. WLCG is the production offline computing platform for the 4 LHC experiments. It is a multiple PB/month infrastructure. The pledged resources for WLCG infrastructure consists of 400000 CPU cores, 400Pbytes of Tape, 300PB disk. In 2025, the new accelerator will produce 10 times more data and will require 60 times more CPU cores.

The panel discussions began by identifying the common challenges in data management. Paul Alexander mentioned that the technology is not being delivered to the end users (astronomers) is the big common challenge. According to Oliver Keeble identifying common

challenges is the easier than finding common applicable solutions. Michael Sterzik said it is not necessary to find common solutions as it causes unnecessary complexities such as in A&A. Domenico Giordano mentioned that a new generation of developers is required who will address the requirements of end users i.e. scientists.

Another important topic that was addressed during the panel discussion was metadata chains, as any wrong inputs at certain stage of metadata processing causes errors in further analysis stages. Manuel Delphino from audience emphasized on the importance of breaking the chains of metadata. Addressing to the question on data preservation, Philippe Deverchere mentioned that hierarchy of storage multiplication of different storage types is increasing complexity. Raphael Ritz mentioned RDA policies on data optimization. Oliver Keeble brought up the issue of data selection for storage i.e. which data should be stored on the tapes and which data can be recreated on demand. Data popularity & software defined networking were amongst other subjects addressed during the discussion. One of the ambitious goal for ESFRI projects in ASTERICS would be to operate in synchronization in response to a specific event. Paul Alexander identified data transfer and OpenStack as follow up topics for OBELICS that are priorities for SKA.

d. Large Databases Session

The general importance of software development was underlined in all presentations. Only a rock-solid management of the large amounts of data can ensure the scientific harvest. Databases form a critical piece in this chain. Dominique Boutigny described the way on how the QServ Database System is in use by LSST. This open source system is planned to be distributed over 1000 nodes and should hold 15 PB of data. Currently large scale test benches are deployed at NCSA and CC-IN2P3. Fabrice Jammes went into a bit more detail on the databases. He described the collaboration of the database and the data access team between slac, CC-IN2P3, NCSA and others, which gives a significant amount of person power. He described a two way approach for the use of databases, first for the "real time alert DB", wher no long analysis is done and secondly the immutable database for analysis. This group is aiming to enable the majority of analytics via databases and is looking for very short response times like a scan throughout the entire sky in less than 1 hour, which means high performance. This shall be achieved by using QServ, which is open source. It sits on MonetDB, MySQL Proxy and other existing components like XRootD. With a distributor/combiner model it follows the MapReduce method. The target for production is to run the solution about 500 node clusters an 2 two international data centres.

Bart Scheers presented a different approach to databases, away from a row wise view to a column wise view implemented by MonetDB. This work has started 1094 at CWI by Kersten&Boncz and relies on the following key features: no expensive code-parsing sice the interpreter is removed, strict columnar architecture, adapted to the computer architecture, hard coded operations on simple arrays. He presented examples of possible relational algebraic operations and based upon data scenarios on BackGem/MeerLicht.

Andre Schaaff presented the possibilities of the Apache Spark system as a tool to be used in the purpose of cross matched services, a cross correlation of sources from large catalogues. Promising performance tests are carried out and significant improvements are expected.

In the following panel discussion, a key question was about what industry can do for us, whether we could use commercial databases and whether we could keep up with the developments in industry. The panelists pointed out the importance of the possibility of directly influencing the development of the databases. They do not expect to have appropriate solutions available from industry and last but not least not to a reasonable price. This flexibility was considered as a key element for own developments even inside a community.

e. Workflow management session

Third day of the workshop began with talk from Mark Sauvage on Data acquisition, processing and distribution for Euclid Mission. The Euclid Mission is the second medium-class mission of the European Space Agency's Cosmic Vision program. Scheduled for a launch in late 2020, it will survey the sky between the Galactic and Ecliptic planes in order to map the time-evolution of the Dark Matter distribution and to elucidate the nature of Dark Energy. The presentation discussed how the Euclid Consortium, and more precisely the Science Ground Segment is building the system that will let the mission face the complexity of data handling, processing, and preservation, to ensure it can be efficiently exploited by scientist. To date Euclid Science Ground Segment has essentially passed its reality based assessment (i.e. move from paper concepts to ones that start to be implemented) and most elements of the pipeline will have a running prototype of significant maturity in 2017.

Davide Salomoni presented H2020-INDIGO DataCloud project. The mission of INDIGO is to develop an open source data and computing platform, aimed at scientific communities, able to be deployed over hybrid e-infrastructures. INDIGO-DataCloud is led by the Italian National Institute for Nuclear Physics (INFN) with 26 beneficiaries in 11 countries, including e-Infrastructure providers, resource centers, scientific communities and industrial partners. The first public INDIGO release was issued at the beginning of August 2016. Indigo services are already available in several testbeds. Indigo is now looking for early adopters/people willing to test and run Indigo components with their applications and requirements.

Luisa Arrabito presented the main functionalities of DIRAC systems for the workload and workflow management with a demonstration of how DIRAC is used for the Monte Carlo production and analysis of CTA (Cherenkov Telescope Array). DIRAC (Distributed Infrastructure with Remote Agent Control) is general framework for the management of tasks over distributed heterogeneous computing infrastructures. It has been originally developed to support the production activities of the LHCb (Large Hadron Collider Beauty) experiment and today is extensively used by several particle physics and biology communities. The main DIRAC components are the Workload and Data management Systems, together with a workflow engine, referred as 'Transformation System'.

Kay Graf gave an overview on the e-needs of the Cubic Kilometre Neutrino Telescope (KM3NeT) project and e-Infrastructure commons that in use and planned with a special focus on workflow management and system preservation. KM3NeT is a future European deep-sea research infrastructure hosting a new generation neutrino telescope (currently in Phase-1) with a volume of several cubic kilometres that - located at the bottom of the Mediterranean Sea - will open a new window on the Universe. There is a need to preserve data including central repository and processing chain for 10 years after publication. During the presentation various examples of data preservation, system preservation and workflow management systems were discussed. Within KM3NeT, new e-needs challenges arise due to a new technical design and more data channels compared to the successful precursor experiments ANTARES, NEMO and NESTOR.

Ana Juan Ferrer presented ATOS, the model of its e-science collaborations with H2020-Helix Nebula and H2020- INDIGO DataCloud and ATOS vision on research challenges. ATOS is a European IT company with branches 72 countries. In 2014 its revenue was €11 billion. Atos is a leader in digital services delivering Systems Integration services, Consulting, Managed Services & BPO, Cloud operations, Big Data & Cyber-security solutions, as well as transactional services. Disruptive Technologies such Swarm computing were also discussed.

Robert Jones gave an overview of H2020- Helix Nebula Science Cloud project. HNSciCloud demonstrates the PCP instrument can be used to incite public and commercial providers to develop innovative, open-source based and standards-compliant services that satisfy the needs of Europe's research communities. Helix Nebula Hybrid Cloud model and Helix Nebula Science Cloud Joint Pre-Commercial Procurement (PCP) were presented. The procurement budget is 5.3M€. Under this PCP a hybrid cloud platform for the European research community. Helix Nebula Science Cloud is part of a foreseen series of EC co- funded procurement projects which will contribute to the European Open Science Cloud.

The panel discussion addressed complex points such as data abstraction. Luisa Arrabito mentioned that data abstraction layer is necessary not only in CTA context but also many other projects on data federation. Mark Sauvage said that at EUCLID they have distributed systems. He also added that it's important for projects to identify what they are processing and for what purpose to address the question of data abstraction layer. Another important topic that was addressed during the panel discussion was super computers. Bob Jones said that supercomputers can be potentially considered for on demand cloud services and it is being investigated. Davide Salomoni mentioned that there are facilities that allow exploitation of supercomputing infrastructure and it has also been mandated by the European commission. Volker Guelzow from audience said supercomputing facilities are comparatively expensive and inconvenient for use. Discussing use of commercial solutions in public funded scientific projects, Ana Juan Ferrer said that from providers' point of view it is a very interesting concept such as Helix Nebula Science Cloud.

f. Interoperability

The Interoperability session began with an overview presentation of LIGO & Virgo Analysis and computing by Michele Punturo. The first direct detection of the gravitational wave signal emitted by the coalescence of two stellar mass black holes occurred the 14 September 2015 in the two LIGO detectors and has been announced on 11 February 2016 by the LIGO scientific collaboration and the Virgo collaboration. The data analysis pipelines implemented in the network of gravitational wave detectors requires large and distributed computational resources to analyse a relatively small amount of data; different analysis algorithms with different requirements in terms of computing power and latency have been implemented. The Virgo data storage is handled by CNAF and CC-IN2P3, jobs are usually executed by EGI facilities or locally.

Peter Couvares presented some of LIGO-Virgo's data analysis computing challenges, including scientific prioritization, resource allocation, optimization, development practices, distributed workflow execution, data movement, job scheduling, and accounting. The talk briefly presented a pyCBC GPU version which is under development in LIGO. LIGO have many remaining computing challenges and much to learn from other collaborations and projects. The talk outlined some of the successes and challenges and solicit ideas for new solutions.

Daniele Gregori presented E4 experience and expertise in HPC with various examples. E4 Computer Engineering is a leader in HPC and GRID Computing for Universities and Research Centres since more than 12 years. The presentation highlighted some of the main installations and working methods to ensure high reliability and the hardware of the system solutions which is used to configure the cluster. E4 is open to collaborate with academic institutions under EU projects to develop new technologies

During the last decade, NVIDIA designed a complete hardware and software ecosystem to support the major computing challenges of the scientific community. Piero Altoe discussed advances made in NVIDIA products such as: i) new Pascal architecture (up to 5.3 TFlops), ii) OpenACC, CUDA and CUDA fortran programming models, iii) PGI complier free community edition, iv) domain specific libraries for deep learning (cuDNN, …), v) NVIDIA's GPU Educators programs, and vi) deep learning institute (DLI). The talk also addressed requirements of scientific projects in the field of Astronomy and Astroparticle physics.

Denis Bastieri presented Low-power computing with ASTRI & CTA use cases. ASTRI is a mini array of Cherenkov telescopes, under construction in a remote site far away from human activities, in order to achieve optimal observation conditions for gamma-ray astronomy. ASTRI, CTA, and Gamma-Ray Astronomy at large, are an optimal test ground for Low-Power Computing and High-Throughput Computing. The presentation covered the architecture and the accomplished performance for ASTRI followed by a "what's next" in Low-Power Computing, where it was showed how traditional accelerators, like mainframe GPUs, could be supported or substituted by ARM processors and FPGA.

Giorgio Richelli described the IBM hardware and software components in high performance computing. IBM has delivered multiple, productive, reliable systems including SP/2, Blue Gene/L, Blue Gene/P and Blue Gene/Q – systems that have perennially ranked in the top echelon of the TOP500, GREEN500, and most recently, GRAPH500 benchmarks. Currently, IBM strategy is centered on OpenPower, coupled with accelerators (GPU, FPGA) connected by innovative interconnects (OpenCAPI, NVLINK). Summarizing IBM strategy Gregori mentioned that IBM has differentiated solutions with accelerators and networking with CAPI and NVLink and IBM Research is paving the way to exascale through innovation and collaboration. During the Panel discussion, chair of the session Giovanni Lamanna evoked some of the sociological issues in astrophysics and astroparticle physics along with the scientific and technological issues. Addressing the issue of coding in X86 architecture, Giorgio Richelli said there has been significant evolution in architecture. Piero Altoe emphasized on clear communication and cooperation between scientists and programmers. In response to skill development subject, Denis Bastieri said professional programmer reduce developing time compared to scientists trained to programme. Fabrice Jammes from audience said hiring professional programmer will also reduce maintenance time. Giovanni Lamanna raised the idea of improving programming skills of young researchers in an academic or laboratory environment by professional training through consulting. So that these young researchers can work together with astrophysicists to improve pipelines. He also discussed the SKA approach of recruiting interdisciplinary professors. On the similar lines, Peter Couvares said, it is a trade-off to have scientific expertise in domain knowledge and professional training in software design. Fabio Pasian from audience said, in the papers released by European Open Science Cloud there is a specific reference to need to have create data scientists. He added that the same approach has been introduced in Bioinformatics and similar approach should be used in astrophysics.

g. Conclusion

As a continuation of discussions from Interoperability session, the participants were informed about First ASTERICS-OBELICS International School that will take place at LAPP-Annecy in June 2017. The school will be specifically intended for PhD students, postdocs and researchers from the domain of astrophysics and astroparticle physics to improve their programming skills. On this occasion, the call for expression of interests for industrial cooperation under ASTERICS-OBELICS was also launched. The industrial cooperation through this call is expected to address the innovation related objectives of OBELICS workpackage. The participating industries were encouraged to explore the possibility of collaboration with OBELICS through this call. In his closing remarks, Giovanni Lamanna thanked all the speakers and participants from ESFRI projects, H2020 projects, EU consortia and industries for outstanding participation in discussions throughout the workshop and thanked the local host INAF for their support in very successful organization of the workshop.