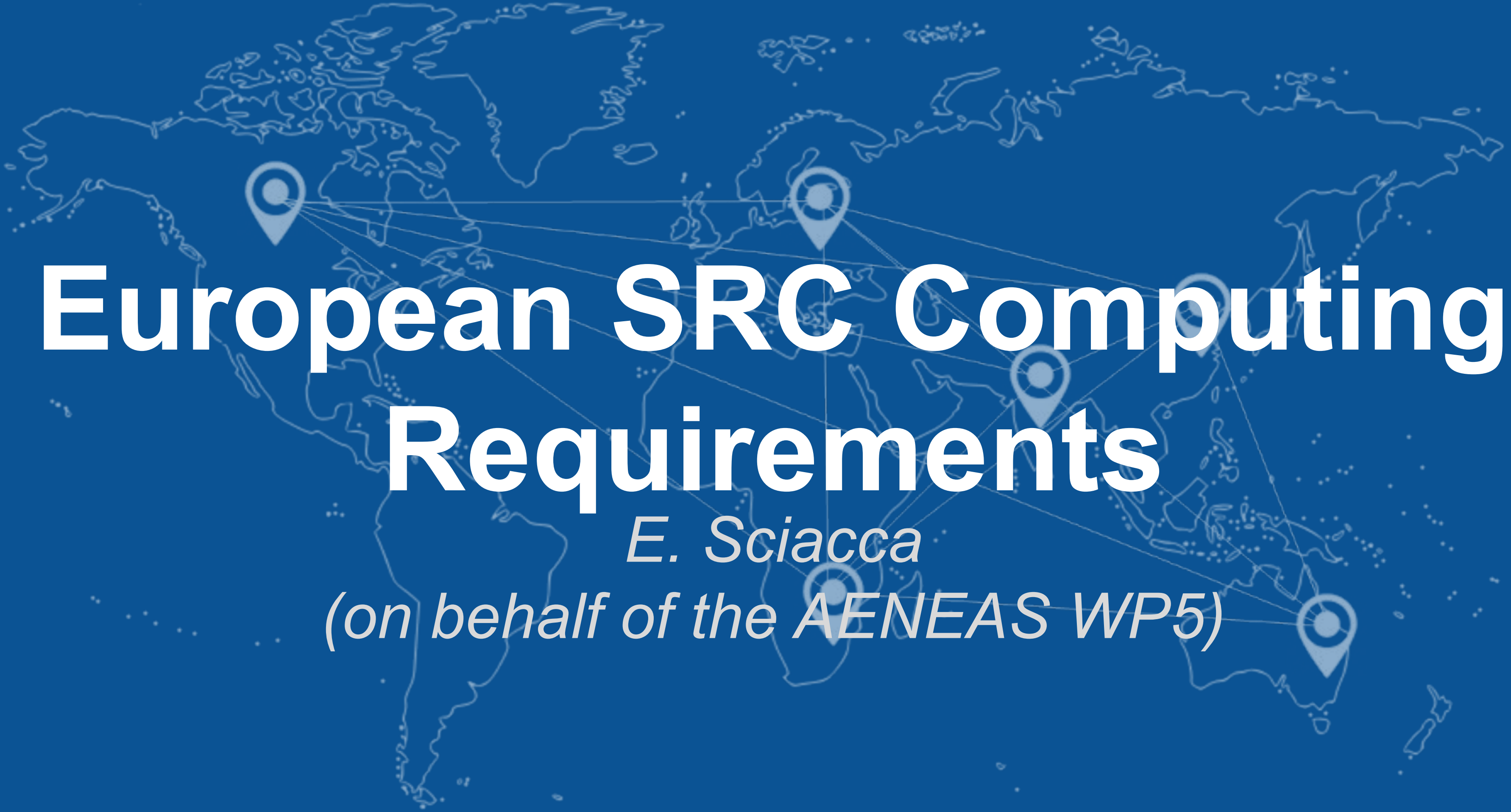


European SRC Computing Requirements

E. Sciacca

(on behalf of the AENEAS WP5)

H2020 AENEAS Outputs

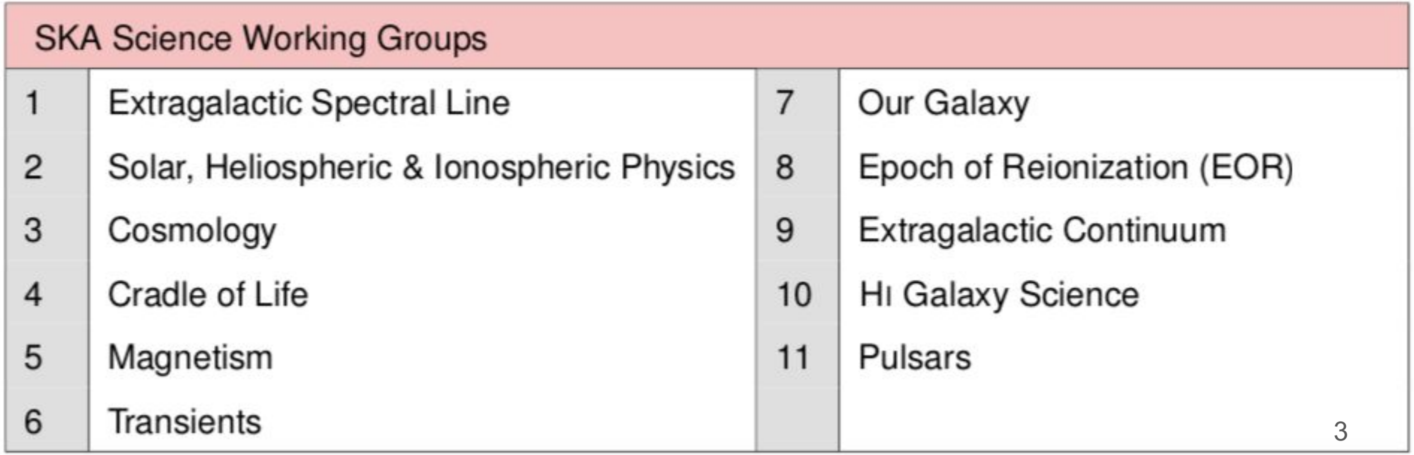
The ultimate **objective** of the AENEAS project is to develop a plan for the **implementation** of a **European Science Data Centre** for the **Square Kilometre Array**.

Work Package 3 focus on the **computing requirements** and in particular identify and assess the **components**, both in **hardware and software**, necessary to deploy the functionality required by the SKA science community.

Based on the full SKA science case, WP3 takes a perspective of **total science delivery** and consider requirements, **computing and storage** scales, and **assess relevant technologies**.

SKA Science Cases

- 11 Science Working Groups (SWGs)
- 13 High Priority Science Objectives (HPSOs)
- 15 SKA SDP Data Products

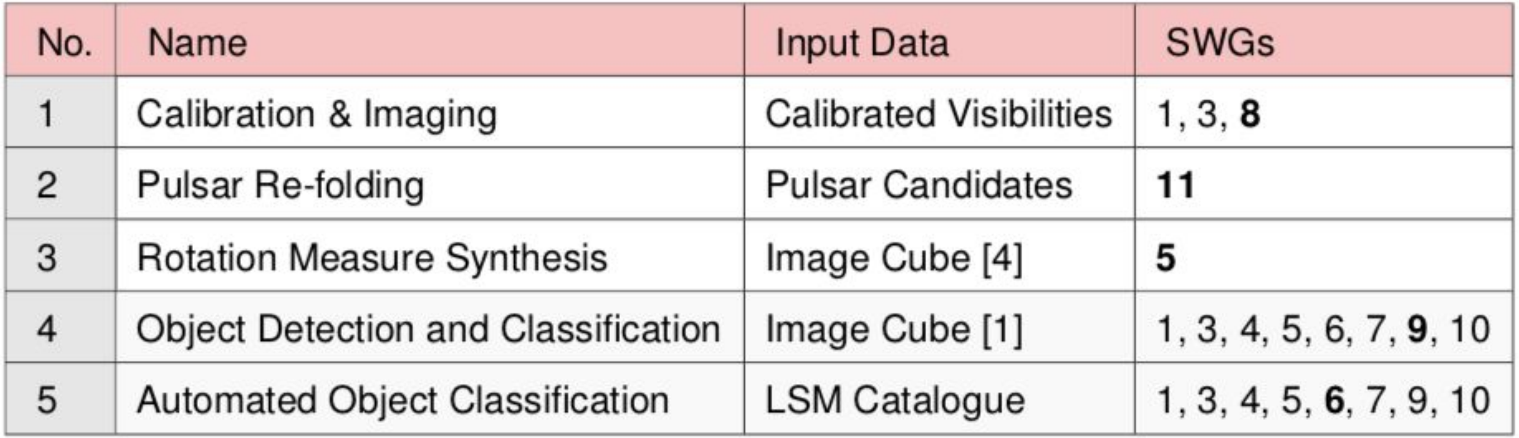


SKA Science Working Groups			
1	Extragalactic Spectral Line	7	Our Galaxy
2	Solar, Heliospheric & Ionospheric Physics	8	Epoch of Reionization (EOR)
3	Cosmology	9	Extragalactic Continuum
4	Cradle of Life	10	HI Galaxy Science
5	Magnetism	11	Pulsars
6	Transients		

ESRC Use Cases

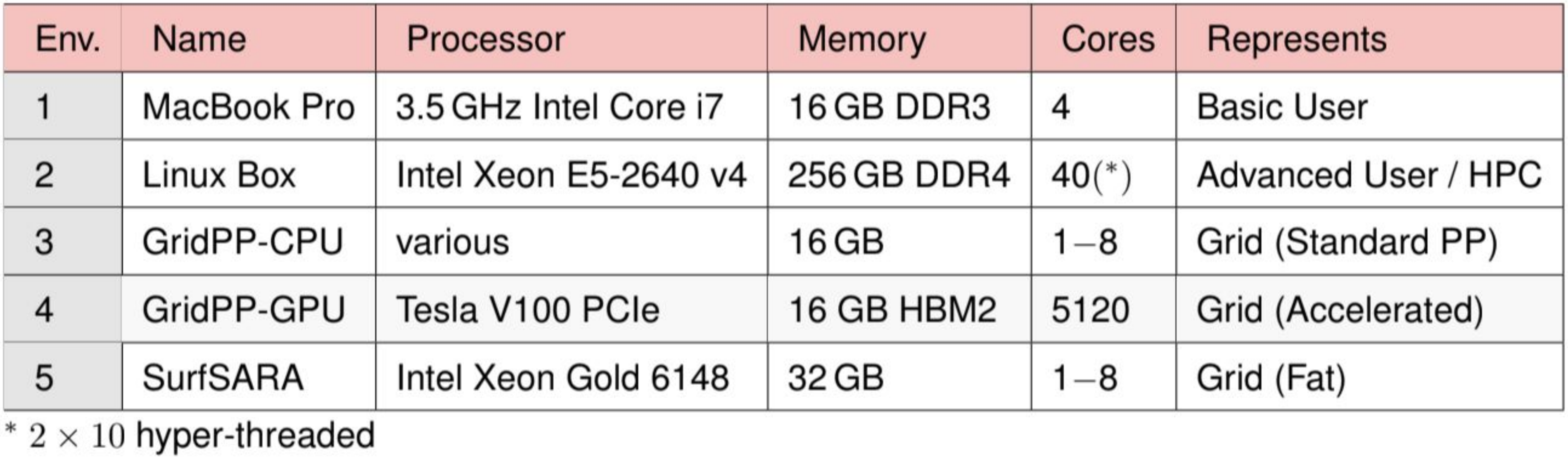
AENEAS Use Cases were selected:

- to be representative of a wide range of processing models
- to cover a range of SKA science working groups
- to use a variety of SDP data products as input data
- to be high usage cases, i.e. they can be run as standard processing on the majority datasets



No.	Name	Input Data	SWGs
1	Calibration & Imaging	Calibrated Visibilities	1, 3, **8**
2	Pulsar Re-folding	Pulsar Candidates	**11**
3	Rotation Measure Synthesis	Image Cube [4]	**5**
4	Object Detection and Classification	Image Cube [1]	1, 3, 4, 5, 6, 7, **9**, 10
5	Automated Object Classification	LSM Catalogue	1, 3, 4, 5, **6**, 7, 9, 10

ESRC Use Cases Computing Environment



Env.	Name	Processor	Memory	Cores	Represents
1	MacBook Pro	3.5 GHz Intel Core i7	16 GB DDR3	4	Basic User
2	Linux Box	Intel Xeon E5-2640 v4	256 GB DDR4	40(*)	Advanced User / HPC
3	GridPP-CPU	various	16 GB	1 – 8	Grid (Standard PP)
4	GridPP-GPU	Tesla V100 PCIe	16 GB HBM2	5120	Grid (Accelerated)
5	SurfSARA	Intel Xeon Gold 6148	32 GB	1 – 8	Grid (Fat)

* 2×10 hyper-threaded

ESRC Use Cases @ GCP

Google Cloud Platform

Tested AENEAS use cases for the ESRC, taken from LOFAR pathfinder, selected as one of the INAF Proof of Concepts.

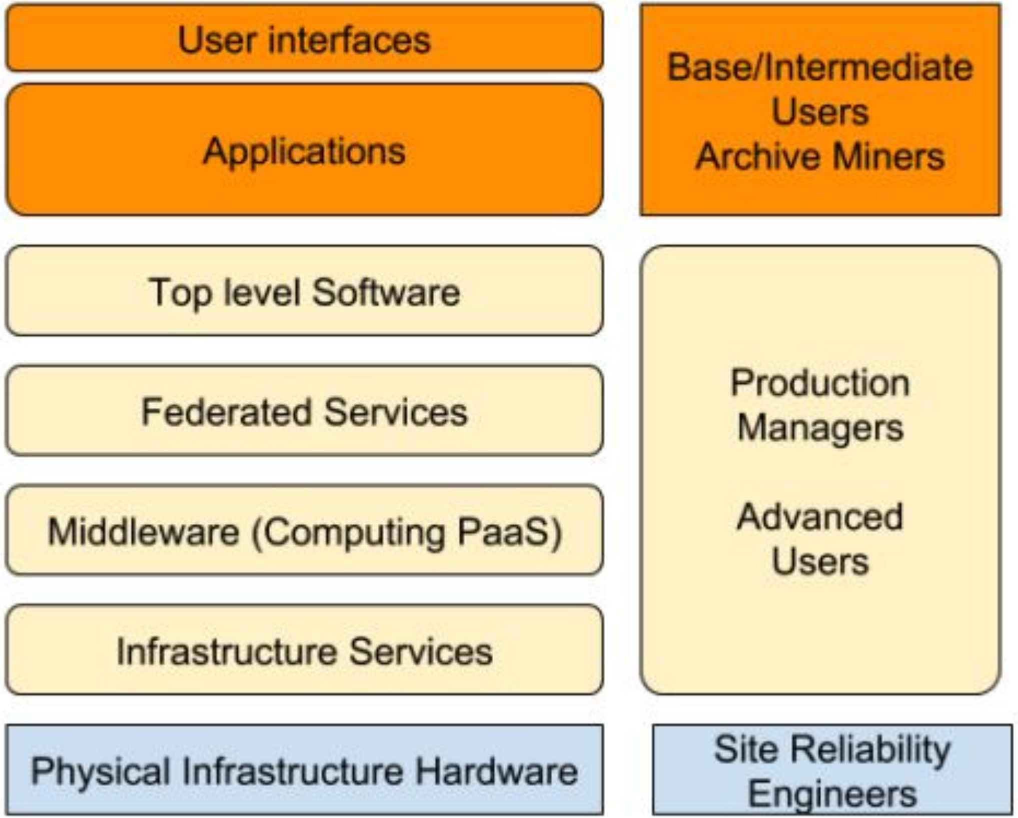
Computing architecture & run strategy

- LOFAR pipelines running on Singularity containers
- Max data size: 300 GB
- Single instances (40 vCPU) used in Google Compute Engine
- Storage: local disk, GC storage bucket (mounted with FUSE)

Results

- Easy porting thanks to Singularity & GC dashboard & documentation
- Good scalability, usability and reasonable costs of the platform wrt the tested use cases
- Significant impact on the computing times (x 2) observed with data accessed in the GC storage

Evaluation of Computing Sw Stack



User interfaces

Applications

Base/Intermediate Users
Archive Miners

Top level Software

Federated Services

Middleware (Computing PaaS)

Infrastructure Services

Production Managers

Advanced Users

Physical Infrastructure Hardware

Site Reliability Engineers

Requirements for data processing/storage/transfer

Data Processing -> ESRC will need to provide a ***minimum of 26 PFlops*** of sustained ***processing*** to achieve the key science goals of the SKA.

Data Storage -> ESRC will need to provide ***storage*** growing at a minimum of ***750 PB per year*** for the first 10 years of key science operations, and ***200 PB per year*** for the following 5 years. We suggest that the first 10 years will be representative of the longer-term overall data rates. The total storage required at an SKA SRC in order to meet the requirements of the SKA HPSOs is therefore 8.5 EB over the course of 15 years.

Data Transfer -> ESRC will need to be capable of ***ingesting data*** from the SDP at a ***rate of 60Gbit/s*** to support SKA key science. This is considered to be a minimum requirement as the ESRC will also need to support the ingest of PI science data products commensally with key science.

Requirements for the computing SW stack

Top-level software stack -> ESRC should adopt a computing stack supporting running ***parallel applications*** at the ESRC in a native way, e.g. without any application modification or porting required to be done.

The computing stack should support running advanced user pipeline applications on resources provisioned either through the computing middleware or through the virtual cluster approach.

The adopted computing stack should support running ***interactive user applications*** (e.g. Jupyter notebooks) pulled from registered external ***repositories*** and/or available in ESRC archive.

Requirements for the computing SW stack

Federated services -> ESRC ***federated A&A*** maps to existing data center A&A services, e.g. ***through proxy services***, rather than forcing existing data centers to change A&A technologies in use. Authentication to use ESRC computing resources should be as simple and less bureaucratic as possible for SKA users.

A ***workload management solution*** should be provided based on mature approach and technologies. At the present status and on the basis of the prototyping activities performed both DIRAC and Rucio provide these functionalities.

Middleware complexity and relative interaction model is masked to base users through a ***web portal/gateway***. In addition to the gateway, advanced users and production managers can also use the standard middleware API and CLI.

Requirements for the computing SW stack

Computing services -> Both ***HPC and Big Data schedulers*** should be supported. Employed technology solutions should be actively developed and supported by large communities.

Schedulers should be ***highly scalable*** supporting more that 100K+ jobs.

To support the needs of advanced users (flexibility and ease of porting new applications), a virtual cluster-on-demand functionality is provided.

ESRC applications should be delivered to the users through ***containers*** to ensure ***easy portability*** pulled from registered external repositories and/or available in ESRC archive.

References

AENEAS web page https://www.aeneas2020.eu

AENEAS D3.1 **Analysis of compute load, data transfer and data storage anticipated as required for SKA Key science** [link]

AENEAS D3.2 **Report on suggested solutions to address each of the key software areas associated with running a distributed ESDC** [link]