The 3rd Italian Conference on Big Data and Data Science

Workshop «Scientific HPC in the pre-Exascale era»

18 September 2024 – CNR - Pisa

The challenge of the data in the SKA Regional Centres network

Giuseppe Tudisco
(on behalf of Andrea Possenti)

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca

SKAO:
two sites…

SKA1-MID, Karoo, South Africa:

133 SKA1 + 64 MeerKAT dishes. Max baseline ~150km

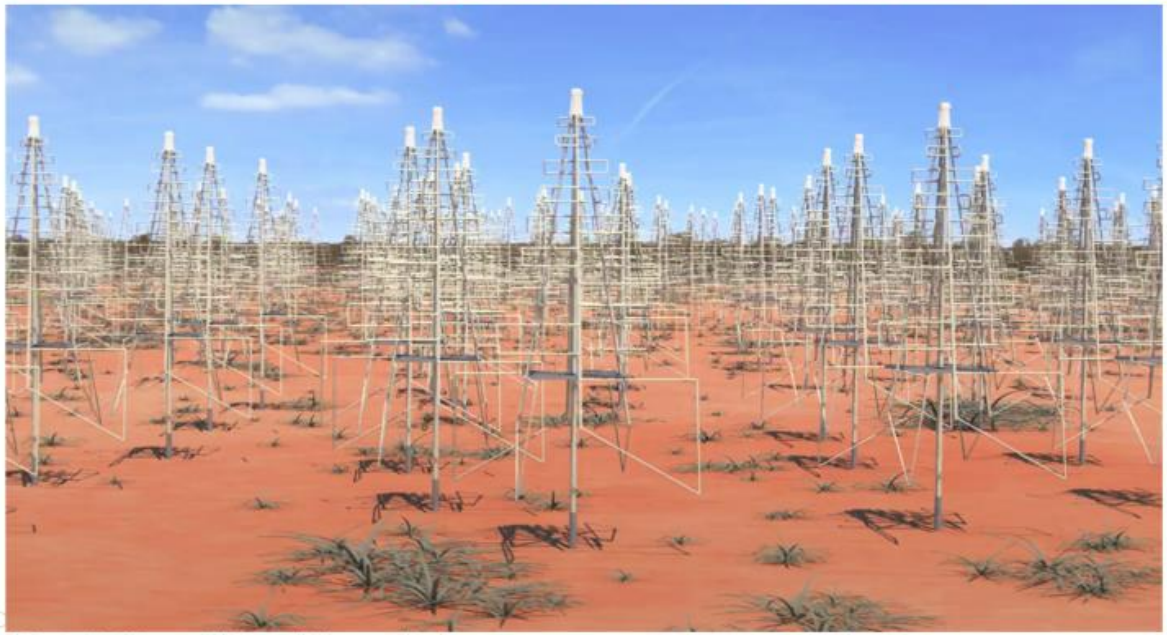
Bands: **2** (0.95–1.76 GHz), **5** (4.6–14(24) GHz), **1** (0.35–1.1 GHz)

[© R. Braun]

SKA1-LOW, Murchison, Australia:

130,000 dipoles (512 stations x 256 antennas); 50–350 MHz

~80km baselines; large areal concentration in core



[© R. Braun]

… two kinds of **ant**ennae

SKAO: some numbers of an unprecedented observatory

Element	SKA1	SKA2
Dishes, feeds, receivers	~200	~2500
Aperture arrays	~130,000	~1,000,000
Signal transport	~1 Pb/s	~10 Pb/s
Signal processing	~exa-MACs	~exa-MACs
High performance computing	~100s peta-flops	~exa-flops
Data storage	Exa-byte capacity	Exa-byte
Power requirements	~10MW	~50MW

Around 2040 ?

[© R. Braun]

Not only antennae…

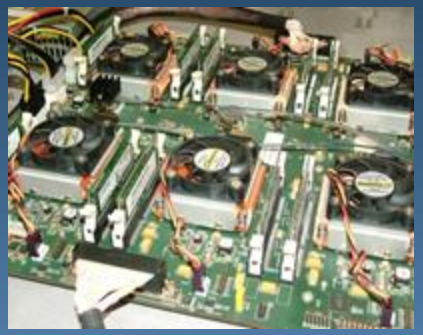
Computing and data archiving are the key ingredients for extracting the best of the science from antennae, network and receivers

[© SKA organization 2021]

Thus, computing and data archiving are the real «limiting factors» for the capabilities of the new Observatory: SKAO will become more and more efficient with the improvement of those

The SKAO data flow: from the antennae to the Regional Centers

CSP: Central Signal Processor



e.g. FPGAs in the ASKAP correlator

5 Tb/s
data buffer of 2 weeks

SDP: Science Data Processor

e.g. SDP prototype, Cambridge

5 + 9 Tb/s
data
buffer of
2 minutes

600 PB/yr
data persistence

SRC: SKA Regional Centre network

SRC
SRC
SRC
SRC
SRC

SKA portal

Distributed facilities

USERS

Adapted from Philippa Hartley (SKAO)

it | SRC
SKAO Regional Centre Italy

The aim of the Ska Regional Centres (SRCs) and the birth of the SRCnet

July 2016: the SKA Board deliberated:

"**The SKA Observatory will coordinate a network of SKA Regional Centres that will provide the data access, data analysis, data archive and user support interfaces with the user community**"

From 2018 to 2023 **a lot of in-kind efforts were done by SKA members (including Italy) to collect the requirements and to draft alternate possible design solutions for the SRCNetwork**

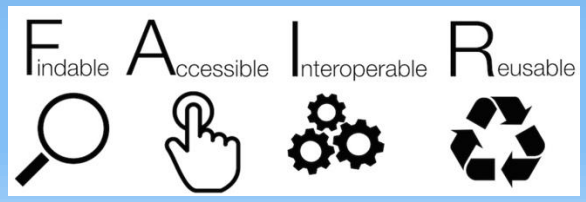
In February 2024: **the SKA IGO Council reviewed and finally endorsed the implementation of the SRCNetwork**

it | SRC
SKAO Regional Centre Italy

The responsibilities of the SKA Observatory and of the Ska Regional Centres (SRCs)

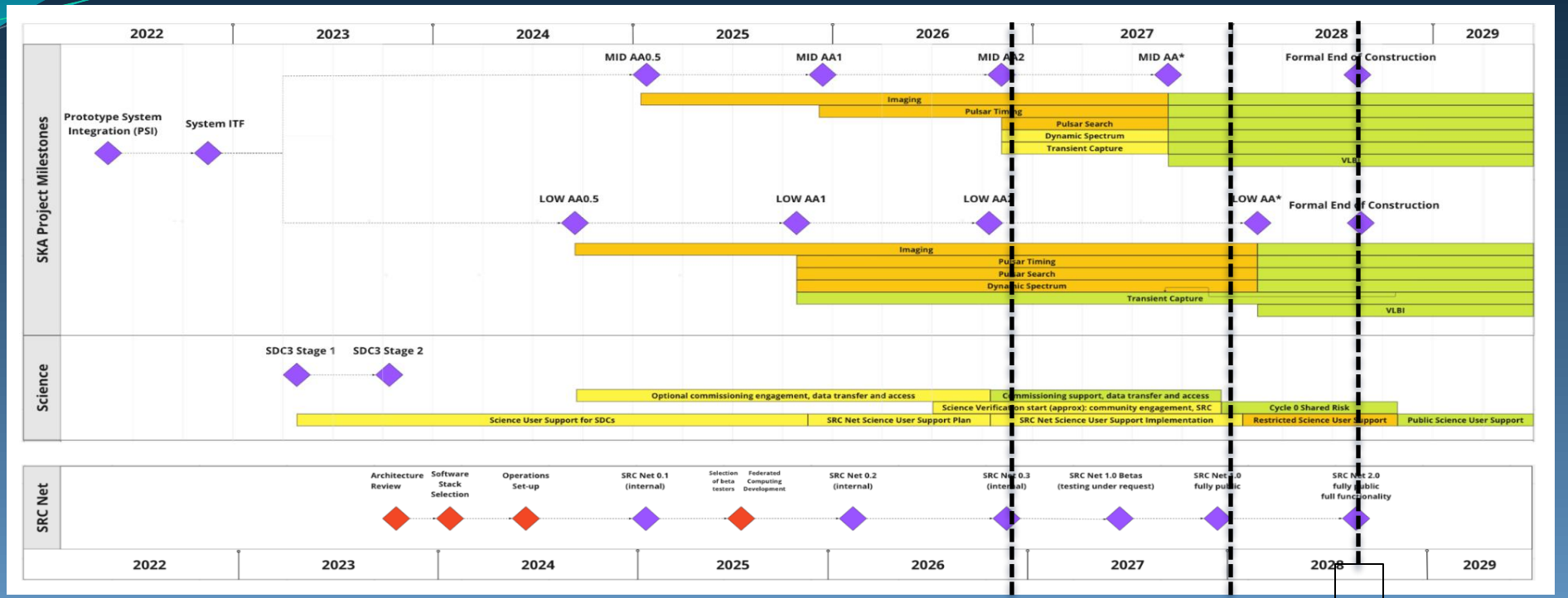
The SKA Observatory and the SRCnet will be jointly responsible for:

- maximizing the quality of SKA data delivered to users;
- the production of Advanced Data Products;
- storing, archiving and curation of the primary SKA output data and of the Advanced Data Products;
- ensuring that the approved science program can be accommodated within available resources;
- ensuring the availability of a platform of distributed services across computational and data infrastructures to support the user community to deliver SKA science, under the FAIR principles.



it | SRC
SKAO Regional Centre Italy

The Roadmap of the SRCnet



Preliminary Plan

SRCnet available to support AA2

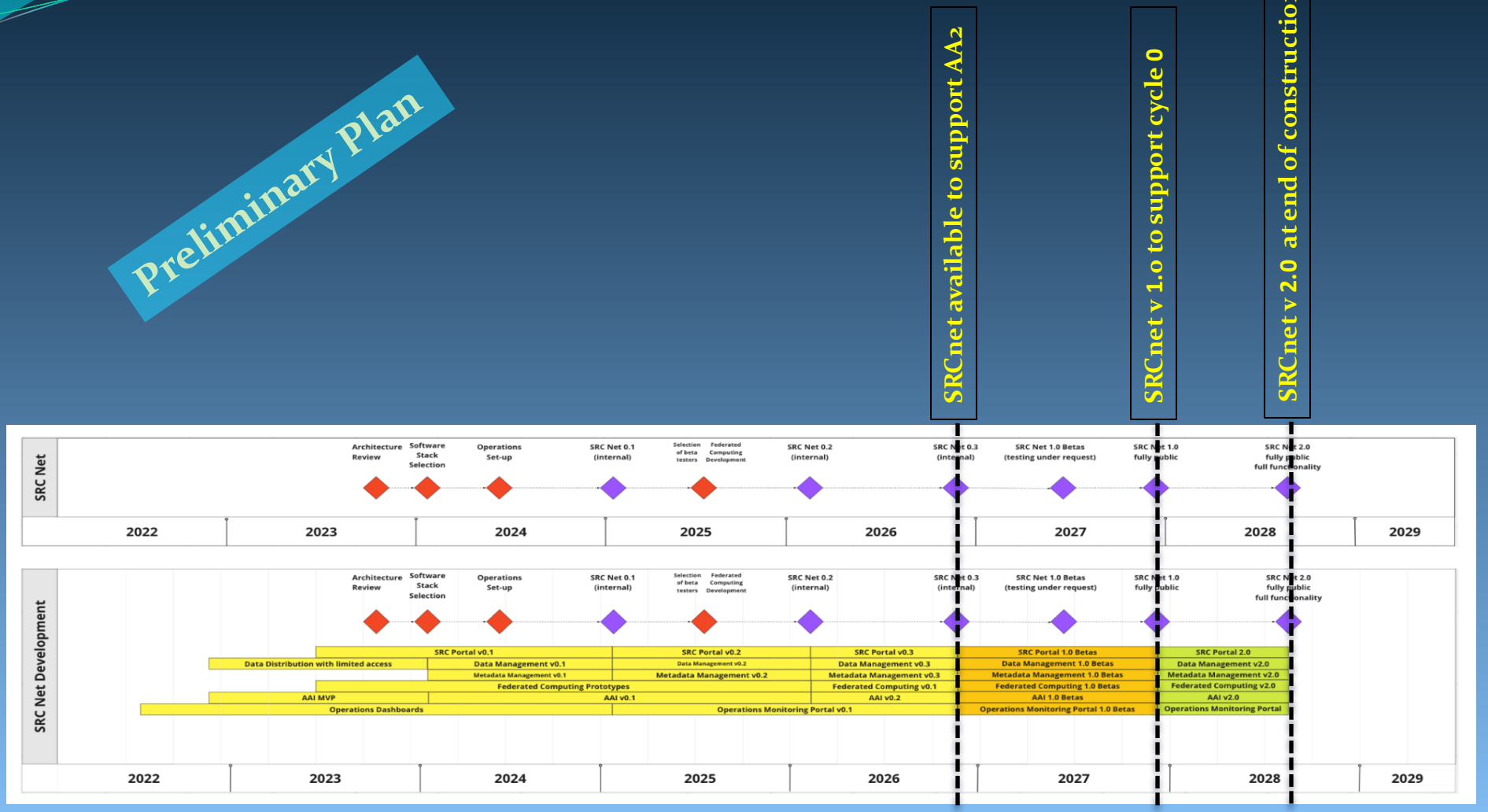
SRCnet v1.0 to support Cycle 0

SRCnet v2.0 at end of construction

it | SRC
SKAO Regional Centre Italy

The Roadmap of the SRCnet

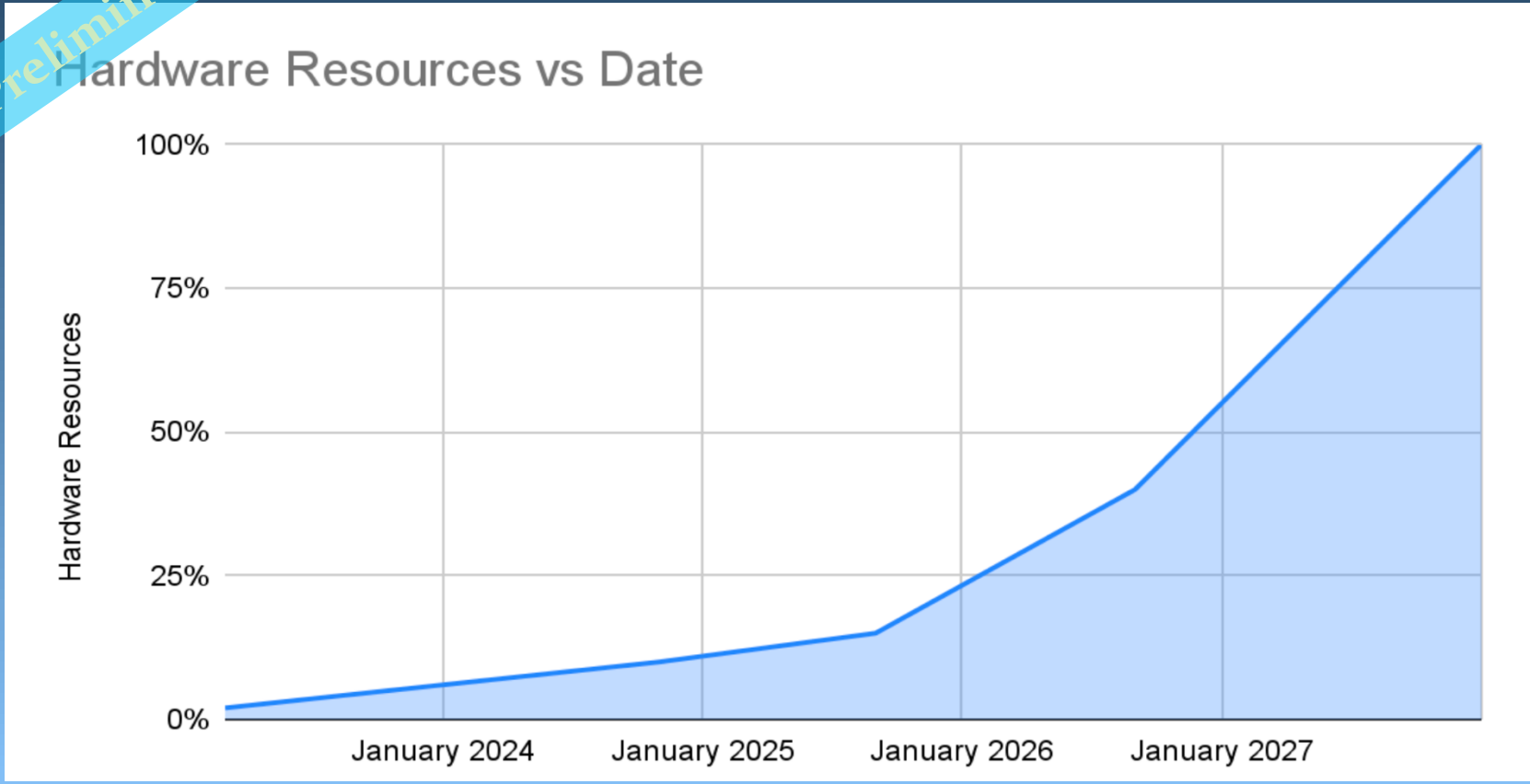
Preliminary Plan



it SRC
SKAO Regional Centre Italy

Hardware resources growth

Preliminary Plan



it SRC
SKAO Regional Centre Italy

Some early estimates

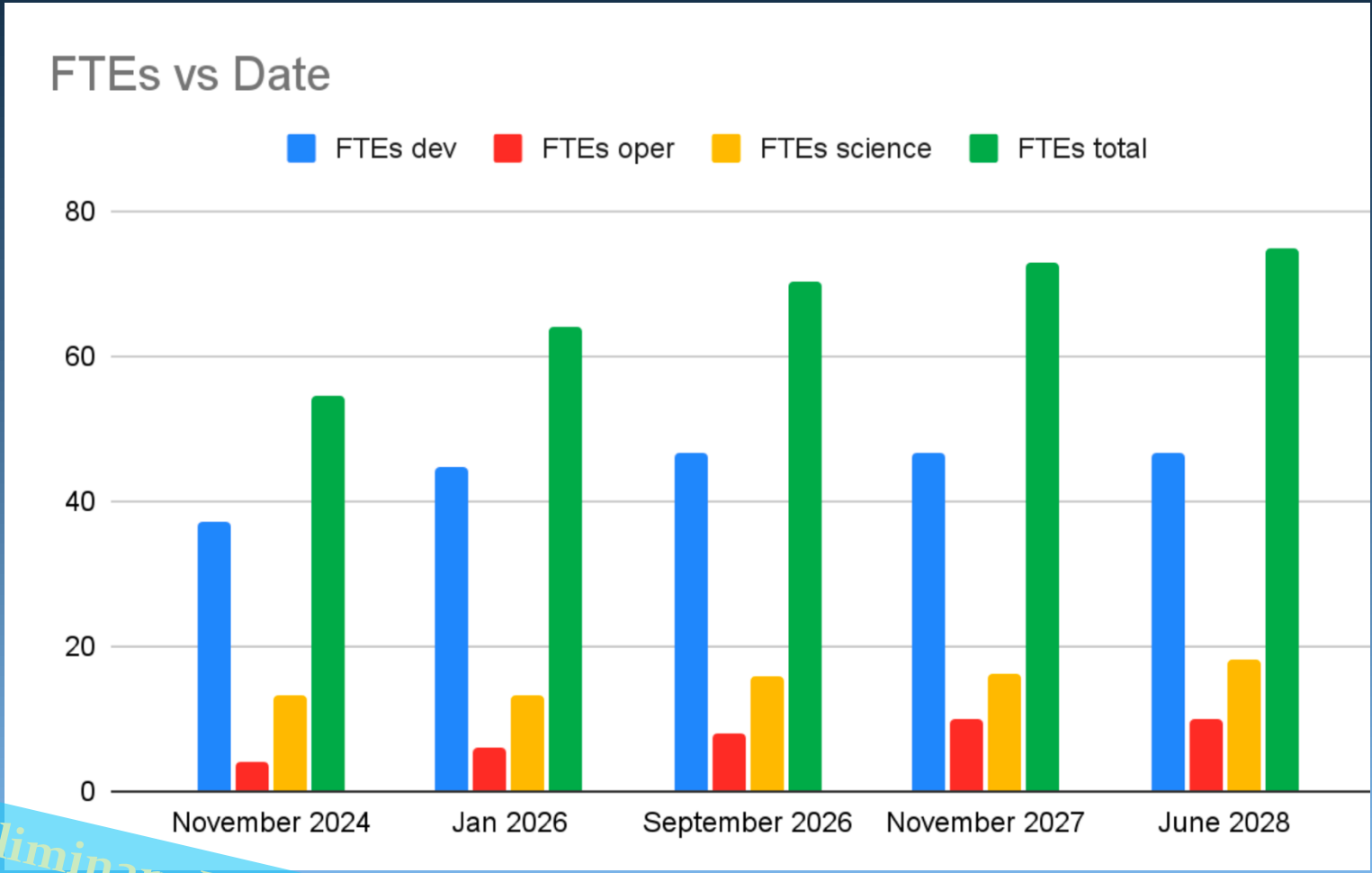
Preliminary Plan

		SRC Net v0.1	SRC Net v0.2	SRC Net v0.3	SRC Net v1.0b	SRC Net v1.0
		Jan 2025	January 2026	Sep 2026	Nov 2027	Jun 2028
Deployment (%)		2.00	10.00	15.00	50.00	100.00
Country	Share (%)	Computing (PFLOPS)	Computing (PFLOPS)	Computing (PFLOPS)	Computing (PFLOPS)	Computing (PFLOPS)
Italy	6	**0.04**	**0.21**	**0.32**	**1.05**	**2.10**
Total	100	0.70	3.50	5.25	17.50	35.00

		SRC Net v0.1	SRC Net v0.2	SRC Net v0.3	SRC Net v1.0b	SRC Net v1.0
		Jan 2025	January 2026	Sep 2026	Nov 2027	Jun 2028
Deployment (%)		2.00	10.00	15.00	50.00	100.00
Country	Share (%)	Storage (PB)	Storage (PB)	Storage (PB)	Storage (PB)	Storage (PB)
Italy	6	**1.27**	**6.36**	**9.54**	**31.80**	**63.60**
Total	100	21.20	106.00	159.00	530.00	1060.00

it SRC
SKAO Regional Centre Italy

Human resources growth



Preliminary Plan

it SRC
SKAO Regional Centre Italy

Italian expected outcome of the SRC network foundation

✓ 1. The identification of a kernel of "modi operandi" in the interactions among the various actors to secure an efficient, persistante, and always developable science-needs driven system

✓ 2. The establishment of a SRC network with a significant node located in Italy

✓ 3. The recognition of the local investments in both hardware and human expertise, and its conversion into incentives as soon as possible

it | SRC
SKAO Regional Centre Italy

The Italian contribution to the global effort

science

≈ 100 Italian astro-scientists are members of the SKA Science Working Groups!

Developed requests and imagine solutions to the USE CASES for the SRC network

+

Staying at the frontline in ADAPTING to the new way for doing data reduction and computation in the SKA era

+

Exploit experience in precursors & pathfinders to provide suggestions and solutions

SKA Science Regional Centres - SCSRC community input

Survey Flow

Standard: Questionnaire Preamble (2 Questions)
Standard: Section 0 - Some general questions (6 Questions)
Standard: Section 1. Data products and scientific requirements (17 Questions)
Standard: Section 1. Data products and scientific requirements_Loop (66 Questions)
Standard: Section 2. Archive mining and VO Interface (19 Questions)
Standard: Section 3. Post-processing – Analysis – Visualisation (53 Questions)
Standard: Section 4. User support (11 Questions)

The Questionnaire for the SWGs: 174 questions!

it | SRC
SKAO Regional Centre Italy

The Italian contribution to the global effort

expertise

Country	Total anticipated fractional FTE for PI21 (SRCNet global activities)
Australia	0,00
Canada	2,00
Switzerland	1,00
China	4,60
France	0,80
Germany	0,00
India	0,00
ITALY	**2,15**
Japan	1,20
Korea	0,80
Neaderland	3,05
Portugal	0,00
South Africa	0,00
Spain	4,10
Sweden	0,80
United Kingdom	4,35
SKAO	5,30
Total	30,15

it | SRC
SKAO Regional Centre Italy

The Italian foreseen contributions

it | SRC
SKAO Regional Centre Italy

personnel: **FTE for International Effort (2022+) e.g. (13 March 2024-01 September 2024) ≈ 2.2 FTE**

Anticipated FTE for the National Effort (2025+) ≈ 2.5 FTE

hardware: already funded assets and soon available to SKAnet v. 0.1

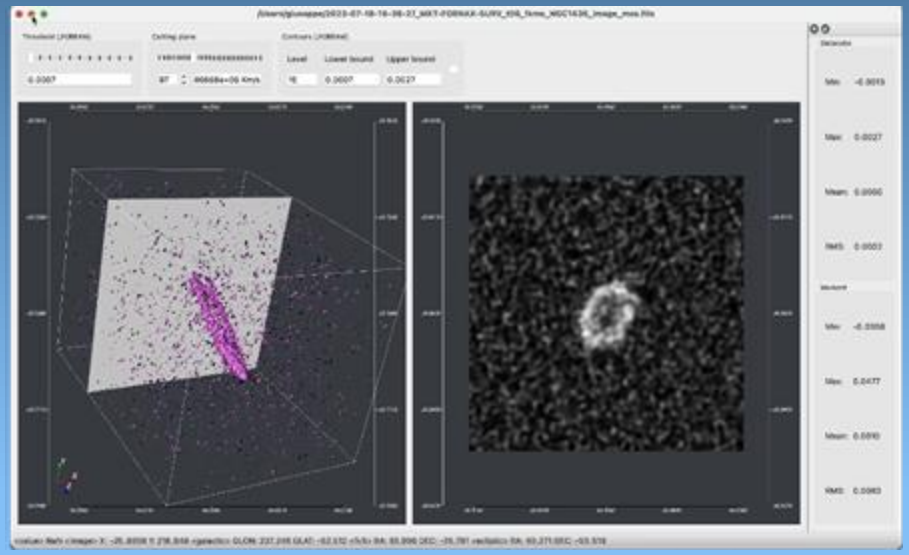
Item	Early 2025	End 2025	
CPU power	0.1 PF/s (Tier 3 – dedicated, CPU only)	1.5 PF/s (Tier 3 – dedicated, CPU+GPU)	**15 PF/s** (Tier 1 – **shared**, CPU+GPU)
Storage	0.3 PB on-line disk (S3), 1.2 PB Tape	2 PB on-line disk (S3), 5 PB Tape	**10 PB Flash** (LUSTRE, **shared**)
Network	10 Gb/s (LAN + WAN)	100 Gb/s (LAN) 100 Gb/s (WAN)	400 Gb/s (LAN) 100 Gb/s (WAN)

The work of the Orange Team

(INAF led and devoted to Visualization tools)

Started with working on Prototype 4: Visualization in PI15 (June 2022)

- Contributing to the **definition** of visualization **use cases** for SRCNet
- **Visualization Tools review** (dependencies, interfaces, workshop)
- **Collection of data products** and data formats from **precursors and pathfinders**
- **Adapting Visualization Tools to address use cases** and work with SRC architecture and its data lake
- Development, testing and deployment of **SODA** (Server-side Operations for Data Access) into SRCNet, integrated with **Rucio Data Lake** and **Discovery services**
- Review of Solutions and Technologies for the **Computing Services API**
- Testing and **deployment** of visualization tools and data access services **into SRC nodes**



Courtesy Tudisco & Vitello

it SRC
SKAO Regional Centre Italy

The work of the Orange Team
(INAF led and devoted to Visualization tools)

- **Mini SRCNet Demonstrator**
  - Benchmark and Optimization of SODA performances
- **Data Lake Integration**
  - Adapt SODA to work with storage manager service (SRM)
  - Collection and classification of heterogeneous datasets from precursors and pathfinders for testing services and tools
- **Computing API**
  - Adapt SODA so that it can process requests through the Computing API (Application Programming Interface)
  - Adapt VisIVO to invoke cutout and visualize data through the Computing API (Application Programming Interface)

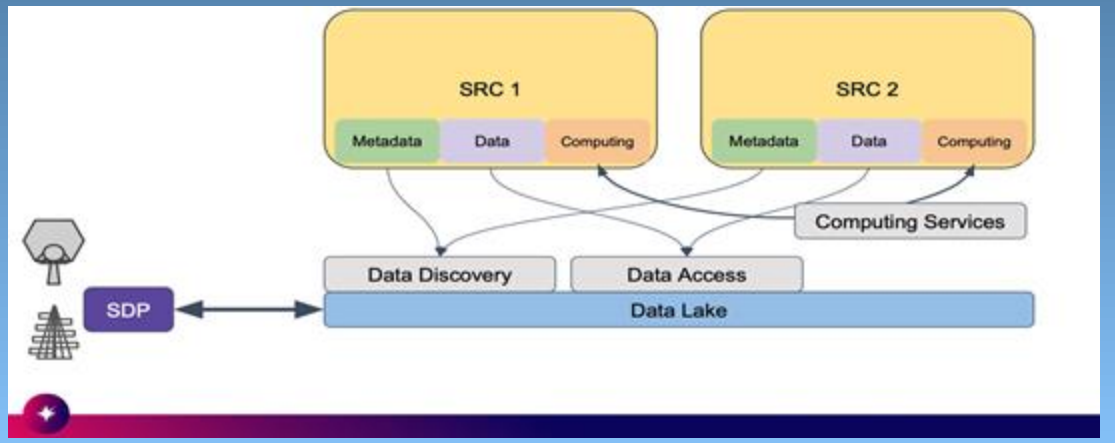


image credits:Jesús Salgado - SKA Regional Centre Architect

Courtesy Tudisco & Vitello

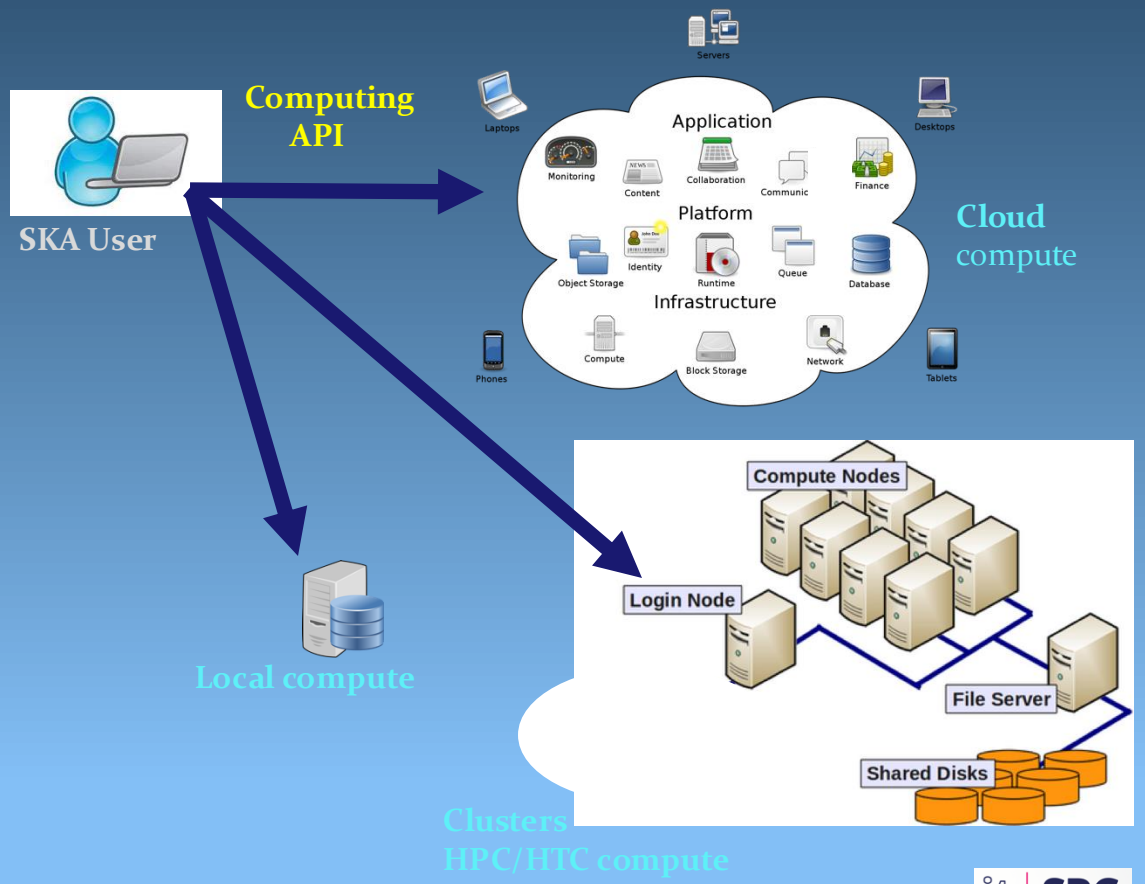
it SRC
SKAO Regional Centre Italy

The work of the Olive Team

Contributions on Computing API (Application Programming Interface)

Discussion

- Use cases
- Requirements
- Authentication and Authorization
- Solutions
  - ExecutionPlanner ?
  - Dirac ?
  - Others ?



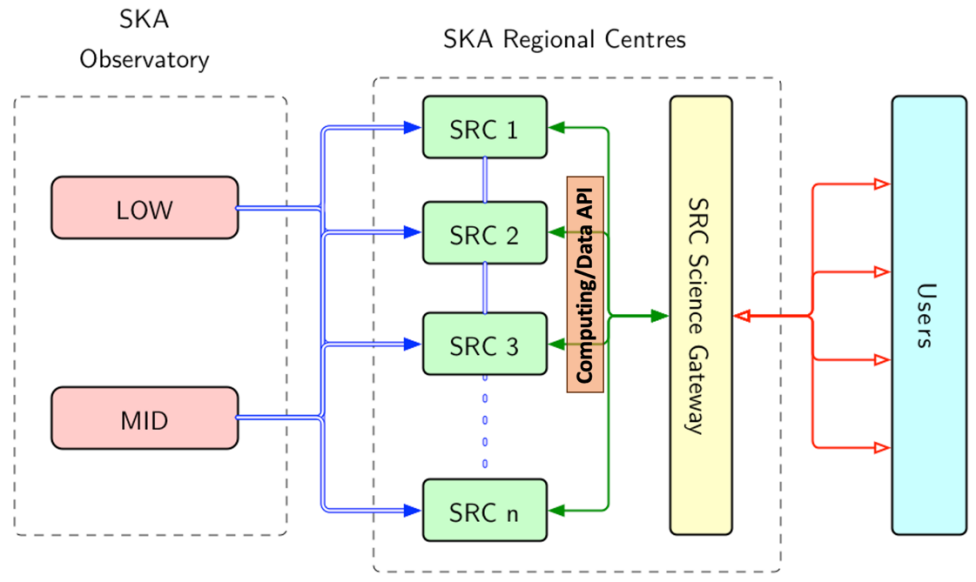
Courtesy Bertocco & Maggio

it SRC
SKAO Regional Centre Italy

Additional work about Computer Service API (Application Programming Interface)

Aim: design & implement an Application Programming Interface (API) to

1. Discover computing services, answering the question "what computing services are available and suitable to run my task?"
2. Access computing services, answering the question "how can I run my task on the selected computing service (and when)?"



THE FOCUS IS NOT ON THE TECHNOLOGY, BUT TO PROVIDE AN ABSTRACTION ABOVE TECHNOLOGY

Courtesy Gheller

it SRC
SKAO Regional Centre Italy

The Italian contribution to the global effort

hardware: initial assets

Acquisition of ≈1.5 **PetaFlop/s** (with a combination CPU and GPU) **and** ≈11 **PBy** (combined between fast disks for computing and tapes for long-term preservation) **Tier-3** computing system, to be installed inside of one of the CINECA areas at **the Bologna Technopole**

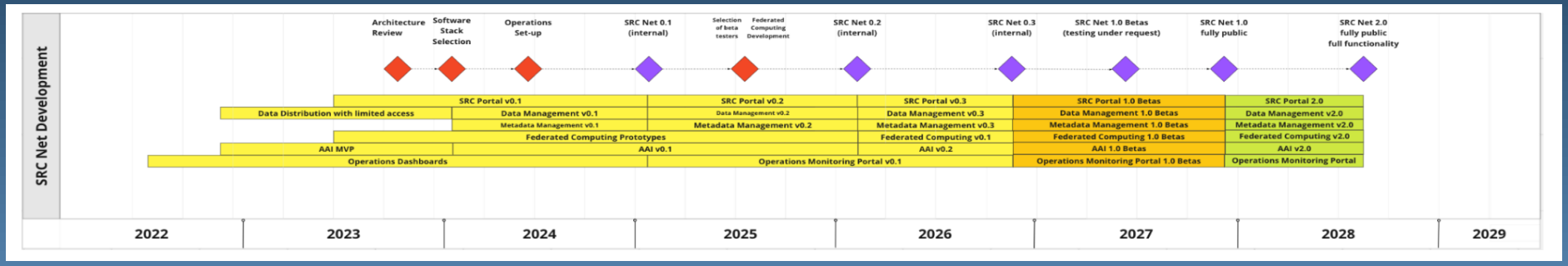
The Technopole already hosts the European weather centre ECMWF, the Leonardo super-computer and will host the United Nations University on Climate Change

Use of **a Tier-2 sizing system** integrated into a Tier-1 sizing system and becoming the **kernel of the Italian node of the SKA Regional Center.** Investment of the CN-PNRR for the needs of INAF and CNR owned by CINECA, with guaranteed (non-exclusive) use for INAF. Expected for INAF: **about 4 PetaFlop/s** (Data Centric Nodes and Booster Nodes) and **≈2 PBy** high speed storage

it | SRC
SKAO Regional Centre Italy

Longer term roadmap of the INAF node

Immediate Objective: create a first prototype of Italian SKA Regional Center integrated in SRCnet v 0.1



2nd Step in the Global effort: by mid 2028, a **Tier-2/Tier-1.5 dedicated** system with capability of **≈ 2 Pflops** and **≈ 60 PBy** of storage (20 PBy on-line and 40 PBy near on-line), connected at **100 GB/s** with the other nodes (**v 1.0** of the SRCnet node)

3rd Step in the Global effort: by 2030, a **Tier-1 dedicated** size infrastructure with capability of **≈ 3+ Pflops** and **≈ 80 PBy/yr** of storage (30 PBy on-line and 50 PBy/yr near on-line), connected at **100 GB/s** with the other nodes (**v 2.0** of the SRCnet node)

Technopolo - Bologna

Thank you !

Thank you !