LOFAR.IT

THE COMPUTING INFRASTRUCTURE

Ugo Becciani – Giuliano Taffoni

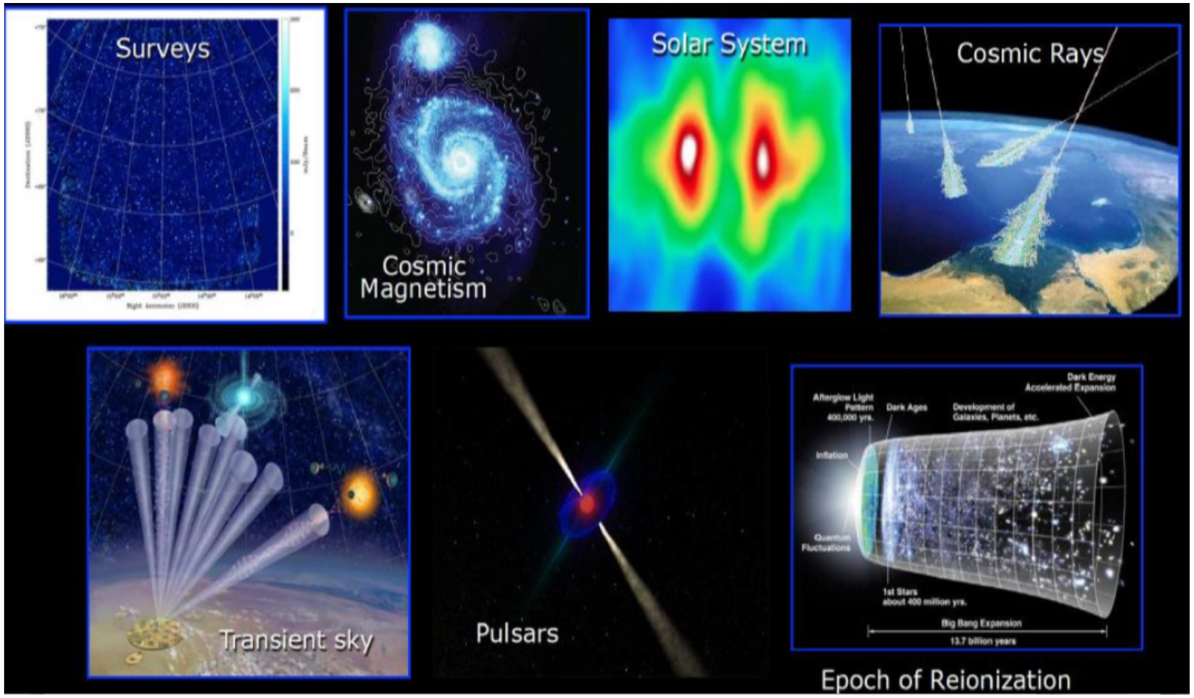
INAF - OACT , INAF-OATS

INAF SCIENCE ARCHIVES & THE BIG DATA CHALLENGE
17-19 June 2019
Rome
INAF-HEADQUARTERS

The International LOFAR Telescope

ILT consists of an **interferometric array of dipole antenna** stations, an it is distributed throughout **9 EU Countries**: NL, Germany, France, Italy, Poland,UK, Sweden, Ireland, Latvia.

Mainly based on the activities of 7 LOFAR Key Science Projects (KP)



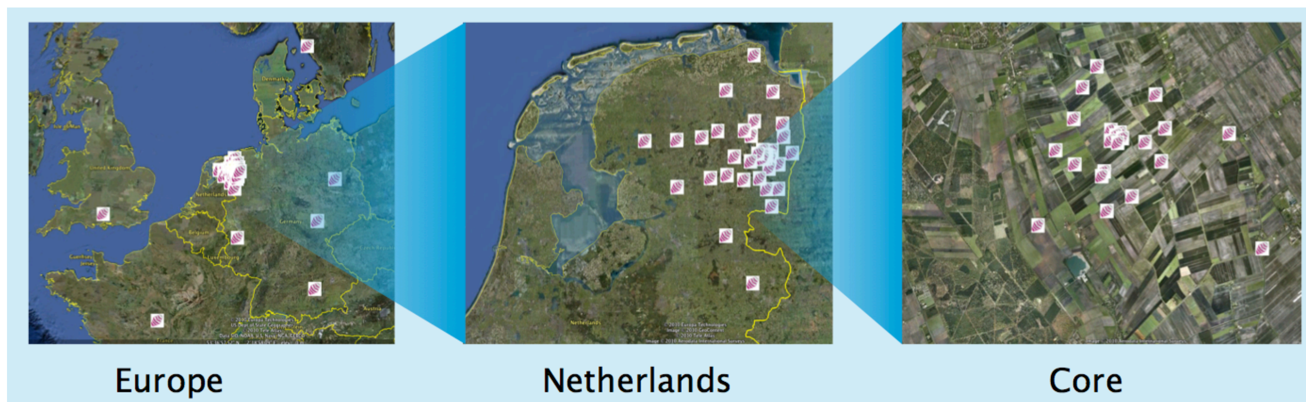
LOFAR - Radio Telescope

- **53 Stations** (**24 core (Exloo)**, 16 remote, 13 International)

- A LOFAR core station consists of **96 Low Band Antennas (LBAs),** operating from **10 to 90MHz** and **48 High Band Antenna (HBA)** tiles that cover the frequency range from **110 to 250 MHz**

- Remote stations in the Netherlands have the same number of HBA tiles, and LBAs

- International stations provide a single cluster of 96 HBA tiles and 96 LBAs (6 station in Germany, 3 in Poland, 1 in France, Ireland and UK, IT)



LOFAR.IT: better late than...

Why were we interested in joining ILT?

LOFAR is the only instrument today to **produce imaging with a resolution of a few arcsec** (…. potentially < arcsec)

New scientific insight in low frequency radio Astronomy.

Potentially a **large community** is interested in LOFAR data.

Key Projects are the scientific groups that drive the evolution and technical knowhow in LOFAR data acquisition and analysis

SKA precursors/pathfinders drive frontier research

Offer to Italian researcher the possibility to improve knowledge on LOFAR data acquisition and analysis, to build a community of researchers…. ready for SKA.

LOFAR is the biggest SKA pathfinder (SKA-Low)

Main Scope of the Italian Consortium

- Build a **LOFAR 2.0 station** in Medicina (2021-2022)
- Build a **LOFAR data analysis infrastructure**
- Implement a **technical and scientific collaboration with ASTRON**
- Develop a **community that is able to work with LOFAR data** (for science and technology)
- Participation of Italian community to **Key Projects** (surveys in particular)

Participation to the KPs, LOFAR guarantee time

LOFAR.IT Board and WGs

Consortium for the participation to International LOFAR Telescope

Board: **Gianfranco Brunetti(Coordinator** INAF-IRA**) Ugo Becciani**(INAF-OA Catania) Segretario, **Federica Govoni**(INAF-OA Cagliari, UTG II), **Francesco Massaro**(UniTo), **Jader Monari**(INAF-IRA), **Roberto Scaramella**(INAF-OA Roma)

- **Science Advisory Committee**:**Andrea Ferrara** (Chair), Matteo Murgia, Mauro Messerotti, Grazia Umana, Gianni Bernardi, Ettore Carretti, Isabella Prandoni, Laura Pentericci, Marta Burgay, Rossella Cassano, Andrea Chiavassa(UniTO)
- **Technological joint WG ASTRON-INAF:**Established on March 2018 . Leaded by Astron. Primary objective: joint development of RCU for LOFAR 2.0 and eventually LBA2.0.
- **Data WG**: **Giuliano Taffoni**(INAF-OA Trieste) - Chair, Alessandro Costa (INAF-OA Catania), Francesco Bedosti (INAF-IRA), Cristina Knapic (INAF-OA Trieste), Manuela Magliocchetti (INAF-IAPS Roma), Annalisa Bonafede (UniBo, Associata INAF IRA)

Italy Road Map status

Jun 2017 -- Feb 2018: Negotiation with ILT board

Oct 2017 -- Working group to define a RoadMap to join ILT

March 2018 -- The roadmap is presented and approved by INAF

Apr 2018 -- ILT Board approves LOFAR-IT as new member

Aug 2018 -- Call for proposal (observing time) & Call for KP

Oct 2018 – LOFAR.IT computing infrastructure v1

Nov 2018 -- Call for proposal (observing time) & Call for KP + ideas

Jun 2019 -- First Italian LOFAR school

Impact on Italian Community

KP involvement (rapidly evolving with time):

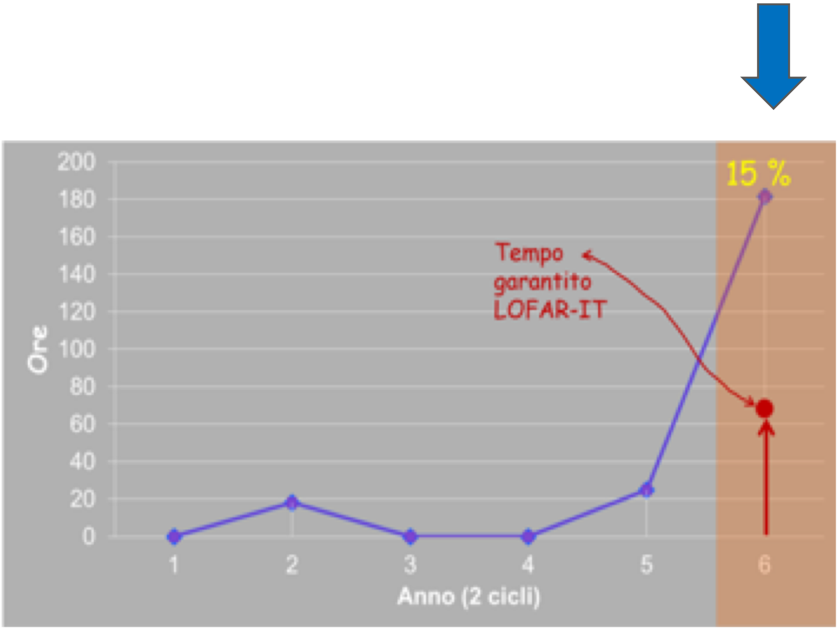
Survey KP → 17 full members + 1 Executive Body member (3rd contributing country)

Magnetism KP → 4 members + 1 core member

Transient KP → 3 members

Solar KP → 2 members

Proposals submission: about 90 h requested/cycle (last 3 cycles)



Science:

2018: 17 papers (2 IT PI)

2019: more than 25 (4-6 IT PI)

Technology :

Build a distributed data reduction and analysis infrastructure and user support.

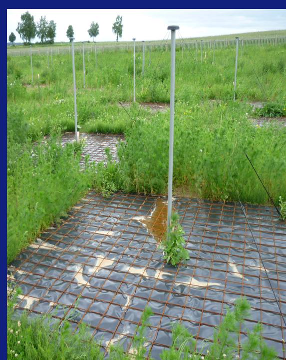
Involvement in LOFAR 2.0 RCU

pipeline and software for calibration and imaging optimization/development

LOFAR - Computing Model - Data flow

Central Processor system

Stations



Realtime system

Offline processing

Long Term Archive

240 Gbit/s → 110 Gbit/s → 2 Gbit/s

LOFAR COMPUTING MODEL.
Central Processor System:The Correlator

- 9 Dell T620 nodes
  - Dual Intel Xeon E5-26xx
  - 2 Nvidia Tesla K10
  - 2 Dual port 10GbE
  - 2 FDR Infiniband HCA

COBALT (**CO**rrelator and **B**eamformer **A**pplication for the **L**OFAR **T**elescope)

Both the F-stage (Fourier transfom) and the X-stage (cross-correlation) are implemented in GPUs

LOFAR COMPUTING MODEL. Central Processor System:Post-Processing

Two central processing (CEP) clusters in Groningen (i.e. near the correlator). Pipelines use locally-developed generic framework.

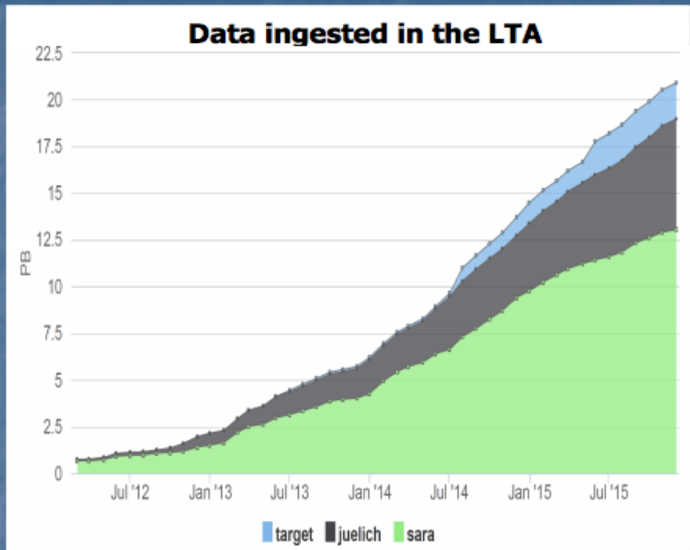
Distributed system built using a *co-design approach* (we know the algorithms and we design the HW)

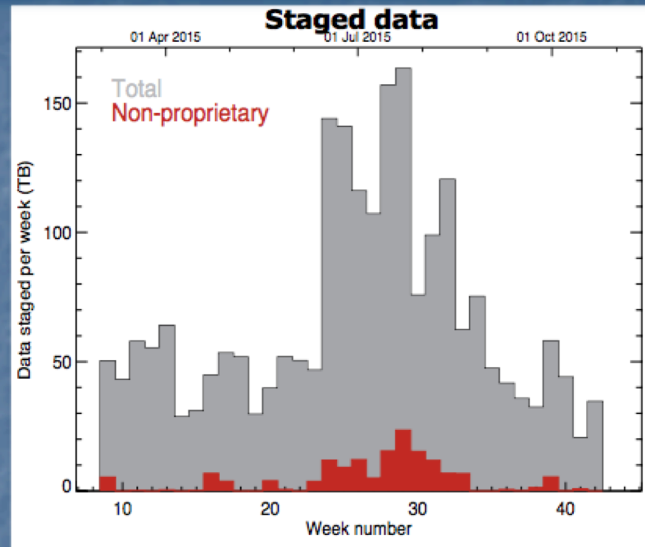
CEP4
Running observatory pipelines: reduction pipelines are used to further process the data into the relevant scientific data products depending on the specific type of observation
Strictly limited to the Radio Observatory -> ingest into LTA
50 compute nodes (+ 4 with GPU; not used in production).
Per node:
12 cores (Intel Xeon E5-2680v3 2.5 GHz), 256 GB memory
2 PB LustreFS storage
Software stack deployed using Docker.

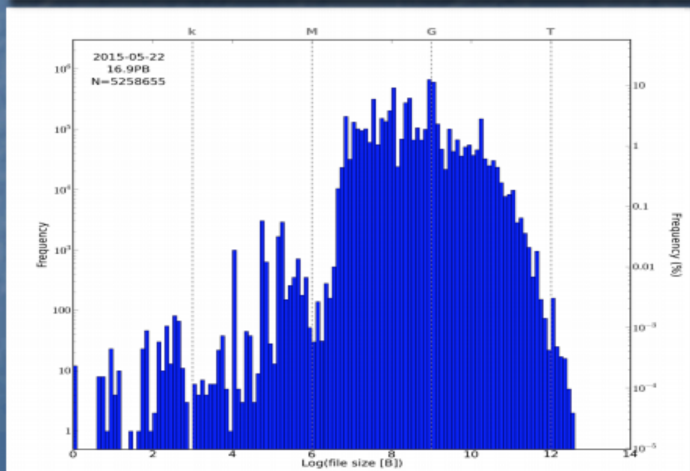
Cobalt
30 TB
CEP4
16TB
CEP3 or **Local facilities**
500GB
Pre-factor (calibration)
580GB
Initial sub (sky-model subtraction from vis: FoV im. of the station)
Factor (low-noise, hr wide-field images)
3 TB op. and **300 GB final**

CEP3: time allocation to PIs based on proposal to do post-processing: 20 nodes
Per node: 20 cores (2 x Xeon e5 2660v2) 128 GB memory
2 x 10Gbps Ethernet interface
22 TB space
Standard LOFAR software

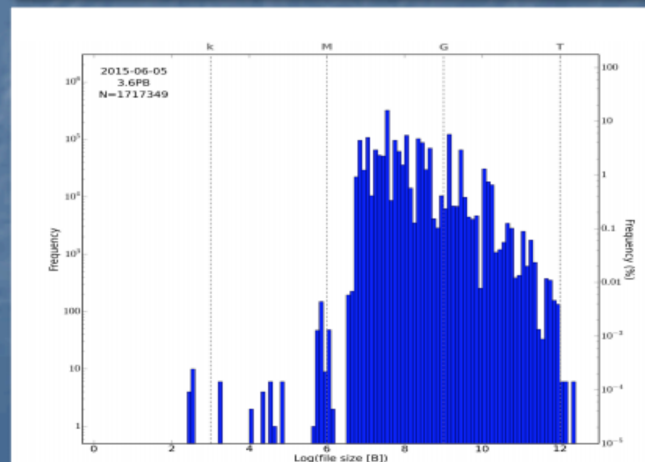
Technological Challenges: LTA







File size distribution ingested



File size distribution staged

- Exceeded 20 PB of data in the LTA!
- Current growth per year: 6 PB (and increasing!!)
- 5.5 million data products
- > 1 billion files

Courtesy of LOFAR LTA team: L. Cerrigone, J. Schaap, H. Holties, W. J. Vriend, Y. Grange

LOFAR.IT: Data Working Group

- **provide the design** of the hardware and software infrastructure for calibration and **data reduction in Italian LOFAR nodes** and coordination of the infrastructure itself;
- coordinate the **installation, configuration and management** of specific software and pipelines for the reduction of LOFAR data;
- provide **technical support** to users belonging to LOFAR IT through testing, verification, **optimization and development of pipelines** for LOFAR data reduction;
- collaborate with LOFAR developers for further code testing and optimization/parallelization of codes and data reduction pipelines (e.g DDFacet pipeline);

LOFAR.IT: Data Working Group
New organization

Data WG

- **Task 1:** User support and Training
  - Documents Manuals best practice
  - Help desk and support
- **Task 2:** Infrastructure
  - infrastructure
  - SW deployment
- **Task 3:** pipelines and software
  - Pipelines optimization
  - Software development

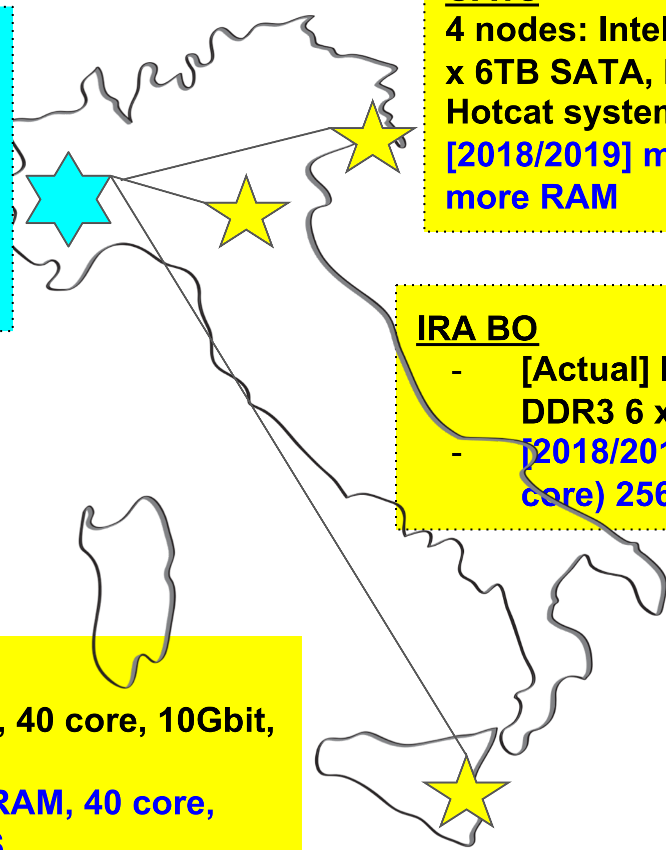
LOFAR.IT e-Infrastructure

UNITO

3 FAT node on OCCAM

4 x Intel Xeon (12 core)

RAM 768 GB DDR4, 1 SSD 800GB, 1 HDD 2TB, 2x10Gb



OATs

4 nodes: Intel Xeon, 256 GB RAM DDR3 6 x 6TB SATA, Infiniband ConnectX®-3 - Hotcat system. 40 TB parallel FS

[2018/2019] more nodes more storage more RAM

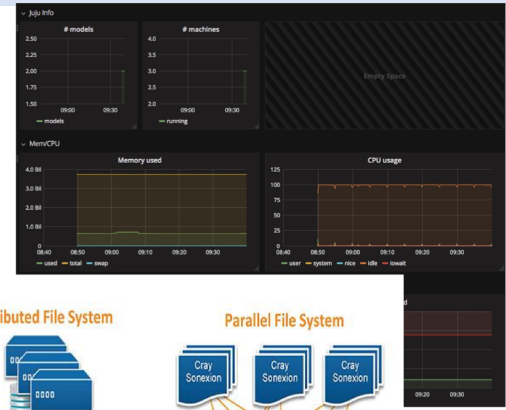
IRA BO

- **[Actual] Intel Xeon 512 GB RAM DDR3 6 x 6TB SATA**
- **[2018/2019] 2 nodes: 2 socket (40 core) 256 GB RAM, 10gbit ethernet**

OACT

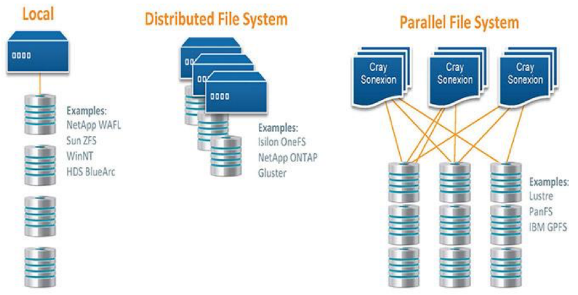
- **[Actual] 2 nodes: 256GB RAM, 40 core, 10Gbit, 20 TB Storage**
- **[2018/2019] 2 nodes: 256 GB RAM, 40 core, 10gbit 20 TB storage Bee-GFS.**

How to build the infrastructure



Monitoring and Operations (telgraf+grafana)

Simplified access to resources and Uniform use: **containers and schedulers**



Role of parallel filesystem to share data between nodes and to speed up IO

Identify or train **“support” groups** for data reduction and for HW/SW

Produce Manuals and HowTOs

Software and data access

- **Software containerisation**
  - **Platform independence**: Build it once, run it anywhere
  - Resource efficiency and density
  - Effective isolation and resource sharing
  - **Operational simplicity and minimal overhead**
  - Orchestration
- **HPC Integration: Singularity containers**
  - parallel computing (MPI)
- **Container repository**
  - Versioning
  - Fine grained Access control
  - Usability
  - Integrated with user software.

LOFAR main pipelines/containers

Pre-factor does the first calibration of LOFAR data

phase 1.a: flag - calibration - transfer of solutions to the target and initial calibration of the target - averaging. Computing time: 3 - 4 days, RAM (core) at least 64GB.

phase 1.b: high and low resolution images and models for auto-calibration. Computing time: 4 -5 days, RAM (core) at least 64GB.

Factor produces low-noise, high-resolution wide-field images from LOFAR HBA data. Computing time: 1 month

DDF alternative direction dependent pipeline for SK (different versions according to the user capabilities). Multithreaded and faster.

more….

LOFAR data reduction tests

INAF ISTITUTO NAZIONALE DI ASTROFISICA
OSSERVATORIO ASTROFISICO DI CATANIA

Tested **Factor** pipeline on OATS and OACT sites employing Singularity containers on single node.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	start:	17:15											
2	end:	21:20											
3		1	1bis	2	3	4	5	6	7	8	9	10	11
4	ncpu	24	40	40	40	40	20	20	20	20	10	10	10
5	nthread_io												
6	wsclean_fmem	0.9	0.9	0.6	0.45	0.3	0.9	0.6	0.45	0.3	0.9	0.6	0.45
7	ndir_per_node												
8	nbands	6		6	6	6							
9	GRAN TOTALE	141	141	131	131	137	140	137	136				
10													
11													
12				ATTENZIONE SONO SCRITTI 123 GB in FACTOR_OUT									
13													
14													
15	NON CI SONO GROSSE VARIAZIONI: Quindi al crescere del numero di core... non c'e' un'effetivo guadagno. Sembra prevalere la fase di I/O nel computo del total time. Poca influenza ha anche la memoria per il wsclean, a meno di non												
16			Anche in confronto con MUONI nodo 14 (storage remoto NFS) sembra prevalere la dase I/O infatti su tale wn-14 anche variando memoria e numero di core... il tempo è costante a meno di non scendere										
17			La differenza tra i 2 run in termini di tempo è da spiegarsi prevalentemenete con l'I/O anche se un certo "miglior effetto" ce l'ha anche il power del nodo, ma non determinante										
18			RESTA INSPIEGABILE LA DIFFENZA Nodo vs nodo CON OATS!! Forse gli RPM del DISCO? usano SSD?										
19													
20	DA FARE CON 6 BANDE SU STO		DA FARE CON 6 BANDE SU STORAGE ESTERNO										
21		1	1bis	2	3	4	5	6	7	8	9	10	11
22	ncpu	24	40	40	40	40	20	20	20	20	10	10	10
23	nthread_io												
24	wsclean_fmem	0.9	0.9	0.6	0.45	0.3	0.9	0.6	0.45	0.3	0.9	0.6	0.45
25	ndir_per_node												
26	nbands	6		6	6	6							

OACT-MUP | OACT-LOFAR | OATS

Factor parameters

- 6 bands (2 MHz)
- *#cpu@OACT*: from 6 to 24 and from 10 to 40
  #cpu@OATs: from 3 to 40
- *wsclean_mem@OACT*: from 30 to 90%
- *wsclean_mem@OATs*: from 30 to 90%

Conclusions: the pipeline is **I/O time dependent** therefore a parallel File Systems should be employed (e.g. BeeGFS). Nodes with at least **10 cores and 256 GB RAM.**

ExaScale: Towards the future

EuroEXA: towards a platform for exascale in Europe (H2020 20Meuro)

- hw-sw **codesign approach**
- **port LOFAR pipelines on Arm**
- Test and verification of system SW that enables the use of exascale capabilities (e.g. OmpSs, GPI etc.
- **BigData application for EuroEXA** (parallel filesystem and software development)

New tech skills → SKA

Role in ILT

Improve our capacity to satisfy the needs of scientific community

LOFAR.IT: Data Working Group. People

Alessandro Costa

Gianmarco Maggio

Sara Bertocco

Luca Tornatore

Eva Sciacca

Fabio Vitello

Simone Riggi

Francesco Cavallaro

Cristina Knapic

Francesco Bedosti

Annalisa Bonafede

Manuela Magliocchetti

Andrea Botteon

Marzia Rivi

Marisa Brienza

Etienne Bonnassieux