Macroarea 1 & ICT
("Galassie e Cosmologia" & "Information and Communication Technologies")

Carlo Burigana

INAF-IASF Bologna,

Univ. Ferrara Dip. Fisica & Scienze della Terra, INFN-Sezione di Bologna

on behalf of the MA1 Committee

with inputs from Meeting INAF - Macroarea 1 (MA1)

Galassie e Cosmologia, *Bologna, 16-17 giugno 2016*

& MA1 (official & unofficial) members

Outline

- **Input from MA1 meeting presentations and other contributions**
  - **… certainly not for describing so many science topics in 20 minutes**
  - **… but for empathising needs for computation resources**
- **Some considerations**
- **Main points in MA1 meeting reports**

CMB data analysis - I

Long list of science targets in cosmology and fundamental physics

(all with computation/simulation needs, often HPC!)

Courtesy
P. Natoli
(CMBday-ASI)

Wide range of angular scales:

- **Large datasets: full sky maps (Mpix)**

Signals ranking from faint to extremely faint

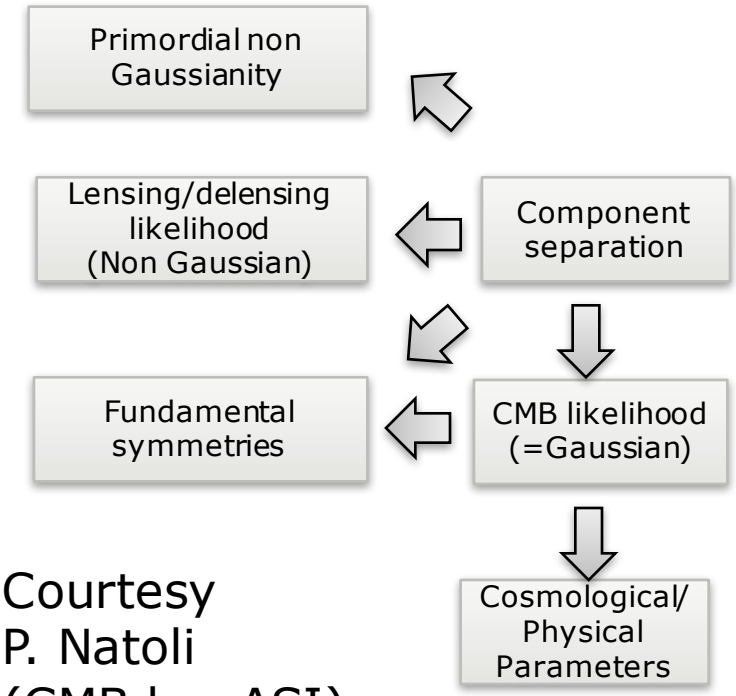
- **Large datasets: many detectors, long observations (Tb to Pb)**

Large not huge. But analysis is *extremely* challenging:

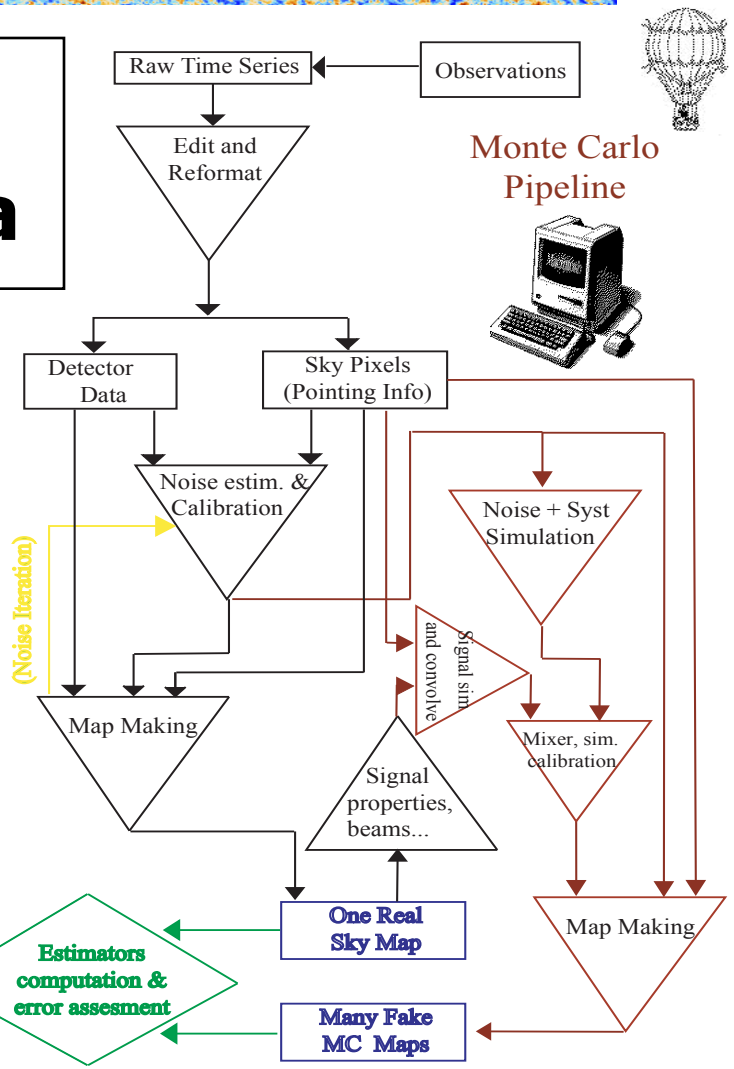
- **Statistically optimal techniques needed: dense problem**
- **Error budget dominated by systematics**

CMB data analysis - II

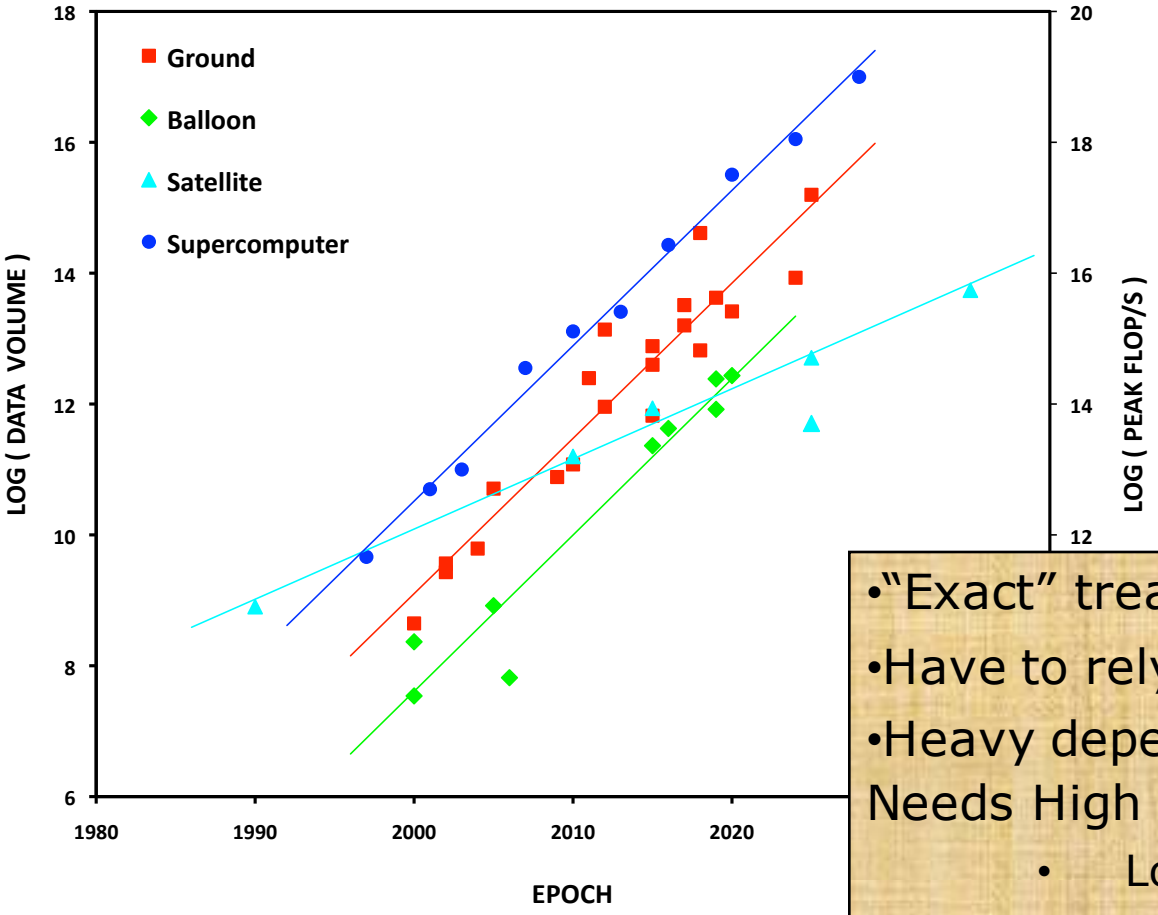
- **Complex pipeline**
- **Relies on simulated data**



Courtesy
P. Natoli
(CMBday-ASI)



CMB data analysis - III



Plot/data by J. Borrill

Courtesy
P. Natoli
(CMBday-ASI)

- "Exact" treatment totally unfeasible
- Have to rely on Monte Carlo methods
- Heavy dependence on supercomputers.

Needs High Performance Computing:

- Low latency, high bandwidth communication
- Significant storage, fast I/O
- No grid or share-at-home!

erc DARK LIGHT

Courtesy
L. Guzzo

"ILLUMINATING DARK ENERGY WITH THE NEXT GENERATION OF COSMOLOGICAL REDSHIFT SURVEYS"

- ERC Advanced Research grant, 5 years (1 May 2012 – 30 April 2017)
- Budget: 1.72 Meuro
- 6 postdoc + 2 PhD positions

GOALS:

- Improve modelling and estimators of clustering and redshift distortions, preparring for precision cosmology
- Test on numerical simulations
- Apply them to current and new surveys to fully exploit information content (e.g. VIPERS)
- Optimally combine with other probes (CMB, WL, clusters, …)

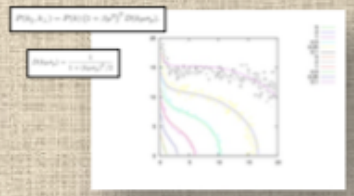
Building scientific leadership and tools for future surveys

Courtesy
L. Guzzo

erc DARK LIGHT

DEVELOPMENT OF NEW MODELS AND ESTIMATORS OF COSMOLOGICAL PARAMETERS FROM GALAXY CLUSTERING AND REDSHIFT-SPACE DISTORTIONS

J. Bel, B. Granett, A. Hawken, F. Mohammad (PhD), D. Bianchi (PhD)

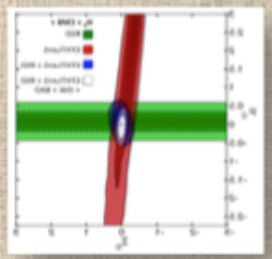


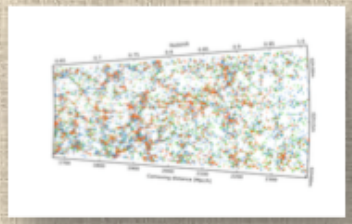
NUMERICAL SIMULATIONS AND MOCK SURVEYS

C. Carbone, Y. Koda, M. Zennaro (PhD)

TESTS AND VALIDATION ON MOCK SAMPLES

Y. Koda, A. Pezzotta (PhD)





APPLICATION TO REDSHIFT SURVEY DATA (VIPERS, etc.)

B. Granett, A. Pezzotta, S. Rota (PhD)

OPTIMAL COMBINATION WITH OTHER COSMOLOGICAL PROBES (WL, CMB,...)

J. Dossett, C. Carbone

COSMOLOGICAL PARAMETERS

"Dark Energy and Massive Neutrino Universe" (DEMNUni) simulations
(PI Carmelita Carbone)

- ISCRA/CINECA Class-A (Jan 2012): 5×10^6 cpu-hours on Tier-0 BGQ/FERMI
- ISCRA/CINECA Class-B (Nov 2015): 8×10^6 cpu-hours on Tier-0 BGQ/FERMI
- Repository at Pico@cineca: DRES_carbone (85 TB) + DRES_DEMNUni (110TB)
- Class-A: baseline Planck cosmology + M_ν=0, 0.17, 0.3, 0.53 eV
- Class-B: baseline Planck cosmology + (M_ν, w_0, w_a)=(0÷0.16,-0.9,±0.3),(0÷0.16,-1.1,±0.3)
- Gadget-3 with ν-particle component and dynamical dark energy background.
- Box-side size: 2 Gpc/h
- Particle number: 2×2048^3 (CDM+ν)
- CDM mass: ~ 8×10^{10} $M_\odot$/h (neutrino particle mass depends on M_ν)
- Softening length: 20 kpc/h starting redshift: z_{in}=99

Simulation outputs

- 62 temporary snapshots per simulation: ~0.54 TB/snap (CDM+ ν) (stored 5 snaps/sim + all the snaps for $M_\nu = 0, 0.16, 0.53$ eV)
- 62 halo/subhalo catalogs (M_{200}, M_{500}, M_{2500}, M_{vir})
- 62 Matter power-spectra

Courtesy
C. Carbone

- 62 temporary gravitational potential grids of size 4096^3 (for CMB weak-lensing)
- 62 temporary grids of size 4096^3 for the derivative of the gravitational potential (for ISW/Rees-Sciama)

DEMNUni collaboration

- **List of tools/activities/topics:**

Courtesy
C. Carbone

- **Initial conditions**
- **P-Gadget3 code with massive neutrinos**
- **Simulation runs**
- **CMB-lensing and ISW/Rees-Sciama**
- **Galaxy clustering**
- **HOD/SAM with massive neutrinos**
- **SZ-maps and cross-correlation with lensing**
- **Cross-correlation clusters/weak-lensing**
- **High-order statistics**
- **Voids with massive neutrinos**

Activities for the next future

Courtesy
F. Marulli

- The aim is to extend these works to analyse the huge datasets that will be provided by the next-generation missions, like Euclid

Euclid+SKA: huge synergies. → Scientific: smaller volume higher res. vs large volume low-res, complementarity constraints, lensing, multi-tracers, etc. → Programmatics: e.g. simulations, likelihood definitions and coding, etc.

- This represents a big challenge, both in terms of the implementation of statistical methods to analyse these large datasets, and for what concerns the modelling, with sufficient accuracy, of the measured quantities as a function of the cosmological model to be tested

Courtesy
S. Borgani

Courtesy
F. Marulli

Numerical issues

In the recent past:

- one single code developer, or few (often developer = user) → no need for any specific code language, numerical libraries, operating systems, etc.
- software: few lines of code → simple linear structure, no need for object-oriented languages (e.g. Fortran was ok)
- hardware: 1 CPU → scalar codes ; no need to parallelize or aggressively optimize the algorithms
- no big issues on validation and documentation → a Readme file was enough
- one or few specific tasks for a single code → one code language was ok, no need for code integrations (no wrapper needed to convert from one language to another)

Courtesy
F. Marulli

Numerical issues

Now and in the next future:

- many code developers working together → need of collaborative environments (e.g. CODEEN for Euclid) and common repositories to organise the activities (e.g. SVN, GitHub)
- software: huge number of code lines → object-oriented languages are now highly recommended (e.g. C++, Python)
- hardware: thousands of CPUs → codes have to be parallelized and optimized to run on large clusters in order to analyse increasingly large datasets
- significant efforts on validation and documentation (e.g. doxygen)
- many tasks to be accomplished → wrapper needed to connect codes written in different languages (e.g. SWIG)

What is required

Courtesy
F. Marulli

→ astronomers have to develop complex numerical codes, that cannot be implemented by software engineers (scientific knowledges are required)

→ high-level computing skills are required : object-oriented languages, collaborative environments, parallelization, optimization, documentation, etc.

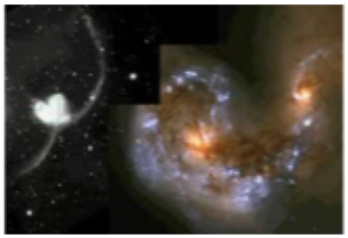
→ many work hours are needed for code implementation and validation

→ huge computing resources are mandatory to optimize, parallelize and validate the algorithms, and generate large mock catalogues

the FIRST team and collaborators

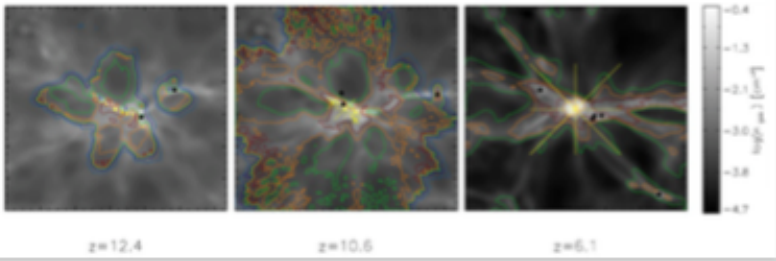
Current Cosmological structure formation simulations include feedback

Mechanical: galaxy dynamical Interactions → **GADGET/RAMSES .**



Chemical: metal/dust production + enrichment, interplay with CGM and IGM → **dustyGadget**.

Radiative: IGM reionisation, Star formation → **CRASH/GAMESH**



Courtesy
L. Graziani &
R. Schneider

Issue

PRACE offers Tier-0, Tier-1 resources required to perform large projects but their access is not so easy..

PRACE critical on global HPC competition
Requires time&commitment on project deadlines
Requires Skills& code
Pre- development and scaling CURIE:
French Tier-0 supercomputer

Limited support for project/code development/debug/testing

Proposal

Courtesy
L. Graziani &
R. Schneider

What is needed for a code development project?

- **Access to software technologies:**

- Commercial compilers (Intel/Portland)
- Standard queue system (SGE/IBM LoadLeveler)

- **Access to similar hardware with less resources:**

→ 512/ (max) 103 cores + accelerated nodes (GPU/Phi) → 1/(max 2) TB RAM Memory/node
→ Terabytes of storage

- **Non competitive/time limited access to resources**

→ Tier-1 cluster DEDICATED to INAF scientists

A theoretical approach to the formation of galaxies

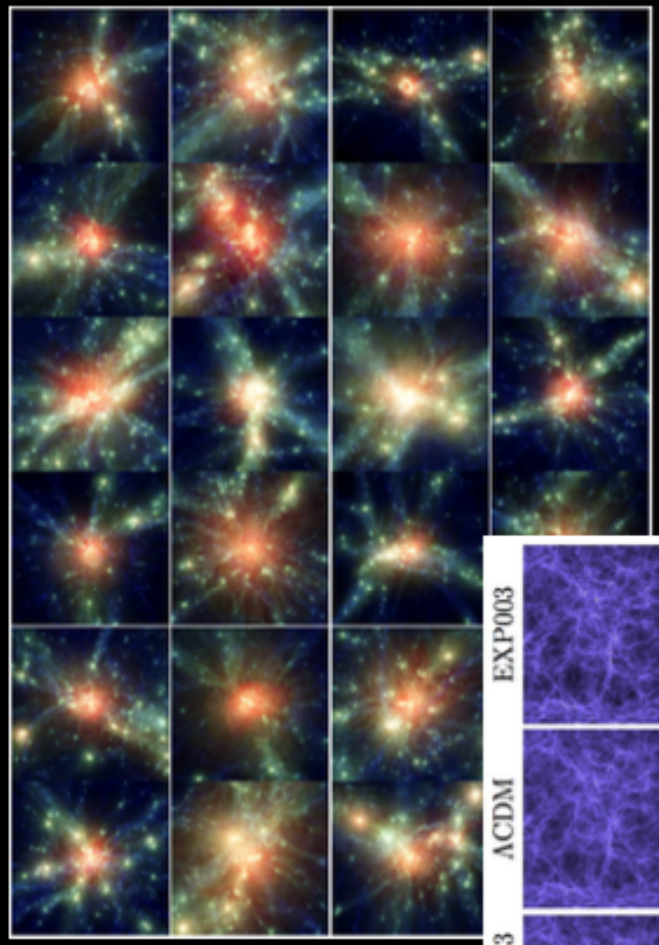
Tools

Courtesy
P. Monaco

- **Phenomenological models**: schematic description of flows in galaxies
  - pro: easy to handle
  - con: little insight on the physics
- **Semi-analytic models**: simplified description of the fate of baryons in the "skeleton" of dark matter halos
  - pro: very fast, easy to sample the parameter space
  - con: difficult to include some processes (e.g. reaccretion of ejected gas)
- **Cosmological hydro simulations**: numerical evolution of a set of initial conditions as predicted by the ΛCDM model
  - pro: much more realistic setting (...yes, they are expensive)
  - con: need of sub-resolution prescriptions (semi-analytic...)

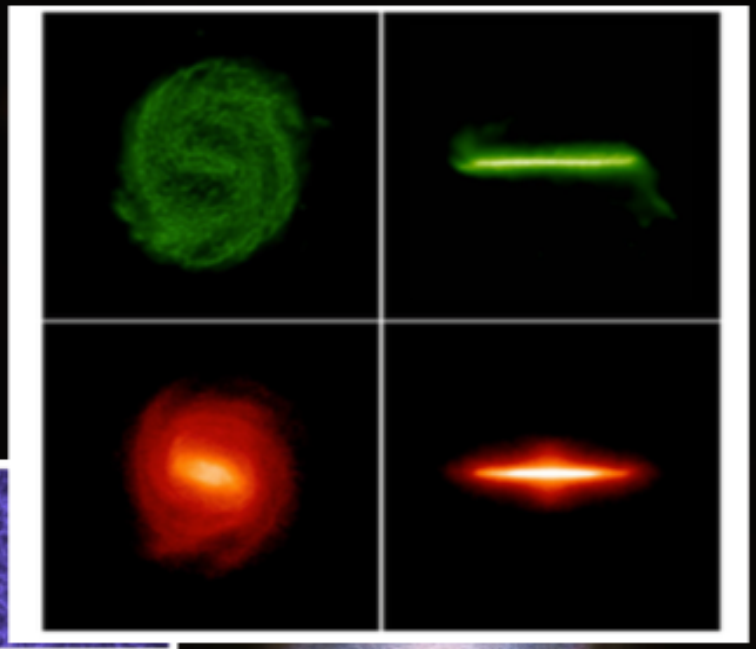
MA1 meeting, Bologna, June 2016

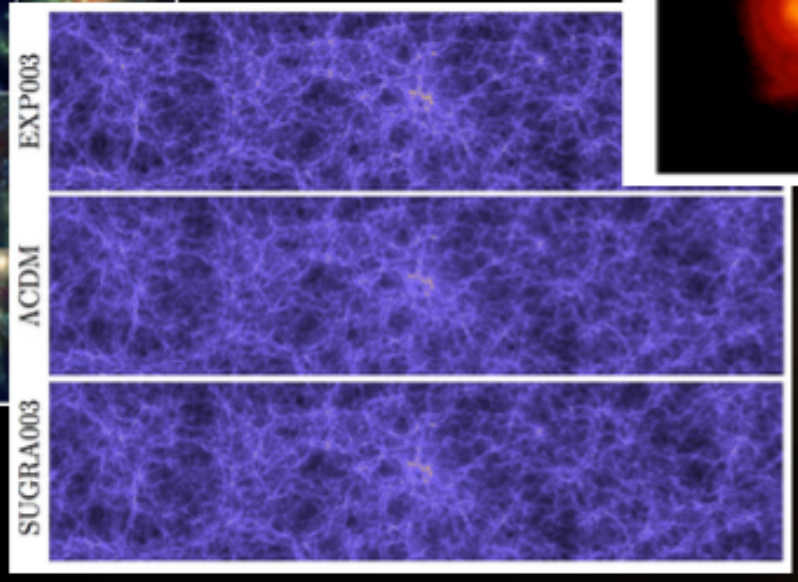
Cosmological (hydrodynamical) simulations



Courtesy
P. Monaco

Galaxies
Galaxy clusters
Large-scale structure





MA1 meeting, Bologna, June 2016

Towards exa-scale computing

Courtesy
P. Monaco

- Codes will be required to scale up to **millions** of cores
- Coding must change: we need specialized staff, software engineers, to port our codes on the new architectures
- It is difficult to hire such professionals: low salary, loose connection to software industry
- Some "tecnologo" staff should be devoted to theory
- Scientists should only be concerned in developing physics modules (and exploit the code, write papers, teach to students...)

MA1 meeting, Bologna, June 2016

The role of INAF in the support of HPC

- It is dangerous to continue running simulations without a "change of paradigm" in the organization of our work
- If INAF **WANTS** to support HPC, then we need:
  - Infrastructure: Tier-2 supercomputer (…in house??)
  - Planning: hiring technological staff in support of theory
  - Lobbying to have a stronger participation in the national contest
- These are the same needs of technological development (e.g. Euclid) -> MA5
- If INAF **DOES NOT** want to support HPC, then please tell us... claiming to support HPC and not providing what listed above is like claiming to support optical astronomy and not to build telescopes

Courtesy
P. Monaco

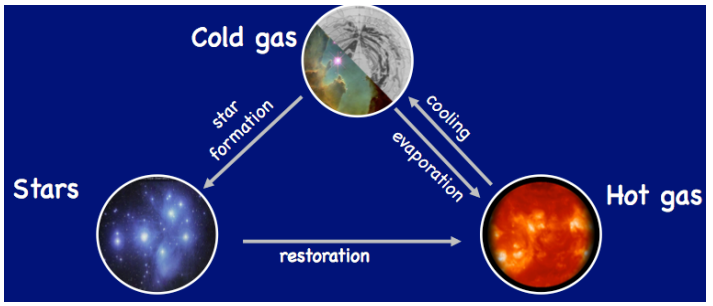
MA1 meeting, Bologna, June 2016

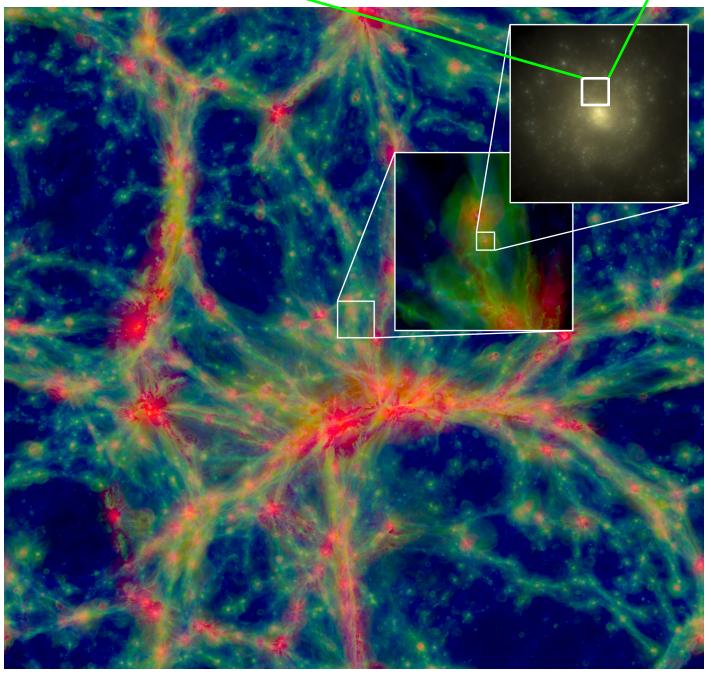
Why are cosmological simulations so complex?

- Gravity: long range interaction, no screening
- Large (spatial and temporal) dynamic ranges:
  - ➔ From ~100 Mpc of cosmological environment to sub-pc scale, relevant for astrophysical processes: > 8 decades
- Resolve down to ~100 pc scales and describe the rest through sub-resolution models
- Cross-talk between resolved and unresolved scales

➔ Codes for computational cosmology: intensive and tough to parallelize

➔ Apps well suited for for co-design of exascale-oriented architectures (**ExaNeSt**)





Courtesy S. Borgani

Facilities are not only telescopes

A policy of INAF is needed (and long overdue !) for:

- High-performance computing (HPC)
- High-throughput computing (HTC)
- Ultra-wide band connectivity

Courtesy
S. Borgani

Crucial for scientific exploitation
of a variety of observational data !!!!!

Data storage and preservation infrastructure in place (IA2 service), BUT:

➔INAF doesn't even have a Tier-2.5 machine !

➔Fragmentation in a number of (often obsolete) small clusters

➔No expertise on HW configuration & middleware

➔No collaboration with HW companies to develop HPC/HTC facilities tailored on our needs

Which solution for INAF

A INAF computing center needs far more than "just" a Tier-N machine:

- A hosting infrastructure
- Mid- and long-term data storage (~1/2 of the cost)
- Personnel
- Commitment for a long-term policy: a machine becomes obsolete in 4 years!
- Shall we rather make a deal with other institutes or HPC centre?

Big splash simulation: apply to PRACE or make special deals with a national supercomputing centre

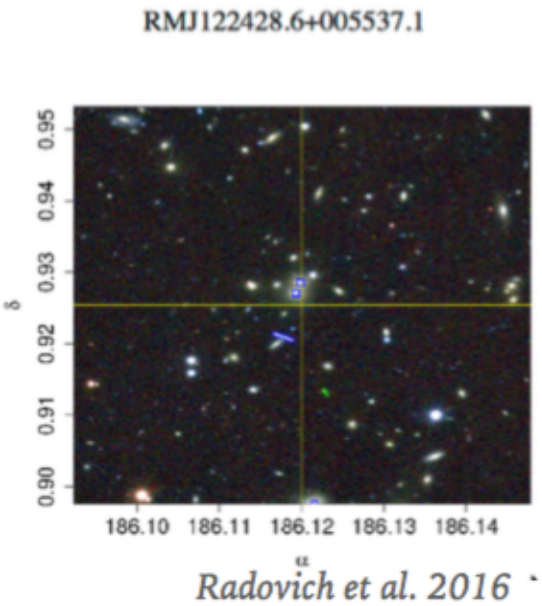
Development phase (including development of a "culture of computing"): flexible and continuous access to a Tier-2/2.5 INAF machine

What is NOT a solution for our simulations: grid/cloud computing

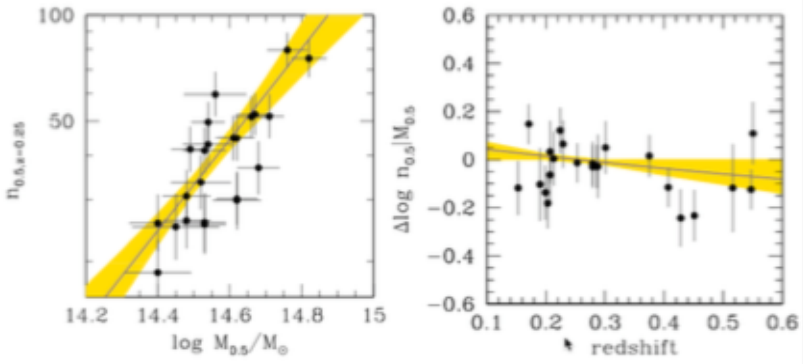
Courtesy S. Borgani

CLUSTER COSMOLOGY: CALIBRATION OF SCALING RELATIONS

- **Euclid and Athena are ahead of us [+other large surveys like XXL, eROSITA...]**
- tens of thousands of clusters available for cosmology, provided we measure their masses
- not yet feasible to think to measure lensing masses for all of them, but we can use lensing to calibrate scaling relations, e.g. **Mass-Richness, M-(X-ray observables)** [e.g. Andreon & Congdon 2014; Sereno & Ettori] (see A. Biviano's talk)
- groups in **Padova, Bologna and Napoli** (*M. Radovich, L. Moscardini, F. Bellagamba, M. Roncarelli, G. Covone,...*) are working on efficient algorithms for detecting photometrically galaxy clusters
- algorithms being tested and used on **KiDS** data, being challenged by others for implementation in Euclid
- currently working on an area of **~100 sq. deg.** from the KiDS survey
- shortly moving to **~450 sq. deg.** (including the GAMA areas). KiDS covers 1500 sq. deg. of sky
- **KiDS provides accurate imaging for the WL analysis:** ideal for measuring WL masses and calibrate scaling relations



Radovich et al. 2016



Andreon & Congdon, 2014

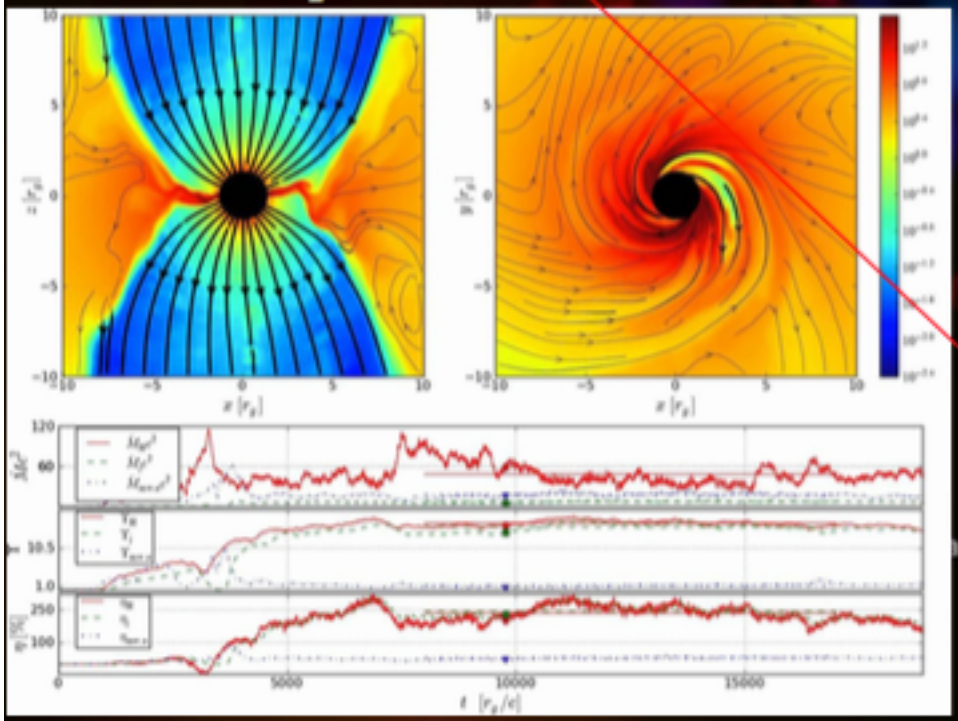
Courtesy M. Meneghetti

Galaxy-SMBH connection: the numerical gap

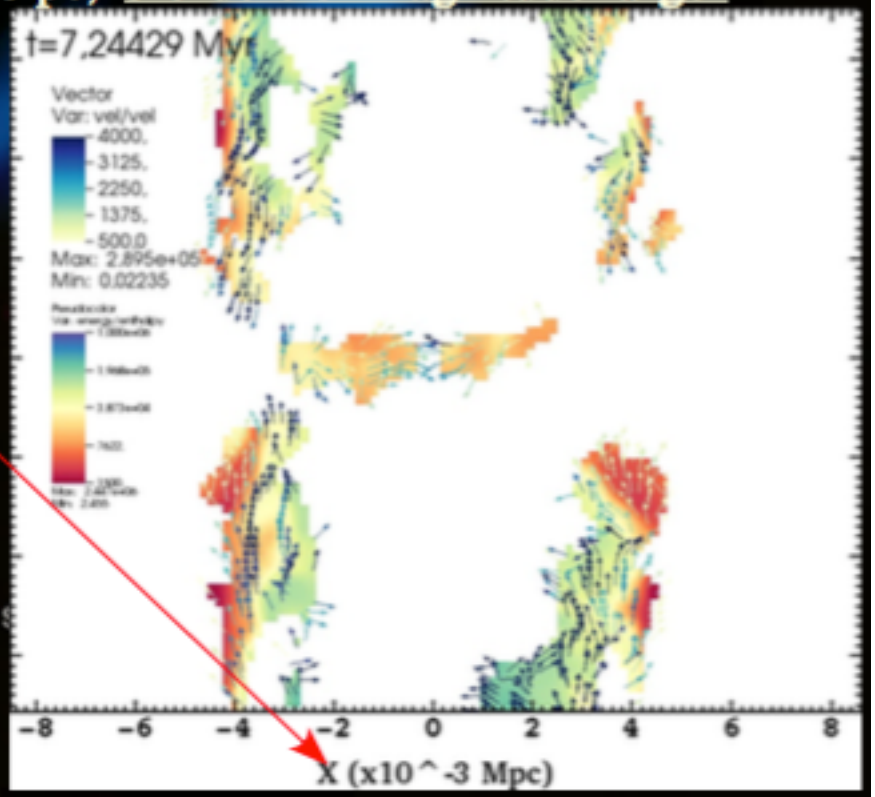
Courtesy V. Antonuccio-Delogu

- Accretion/jet production region: $R_d \sim 10^3\ r_g$ ($\sim 4.78*10^{-3}$ pc)

AMR (Gaspari, Brüggen, Fong, Bicknell, Wagner, Cielo+V.A.-D.,..):
l_{max}=10-15 → $R = L_{box}/2^{lmax} \simeq$ 1.52 - 48.8 pc, 3 orders of magnitude larger than R_d production



Tchekhovskoy et al., 2015



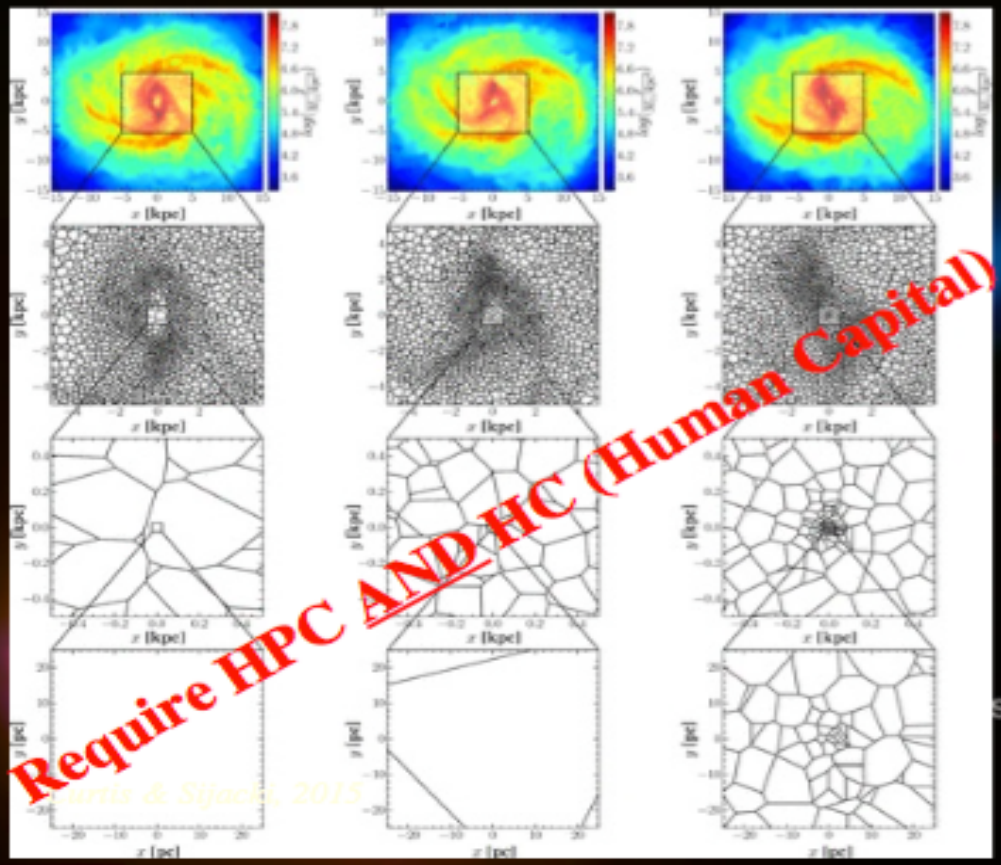
Cielo et al et al., 2015

Galaxy-SMBH connection: HOW TO FILL the numerical gap?

- AREPO (TreeSPH, *Springel*) + HPC parallel *simulations*

Courtesy V. Antonuccio-Delogu

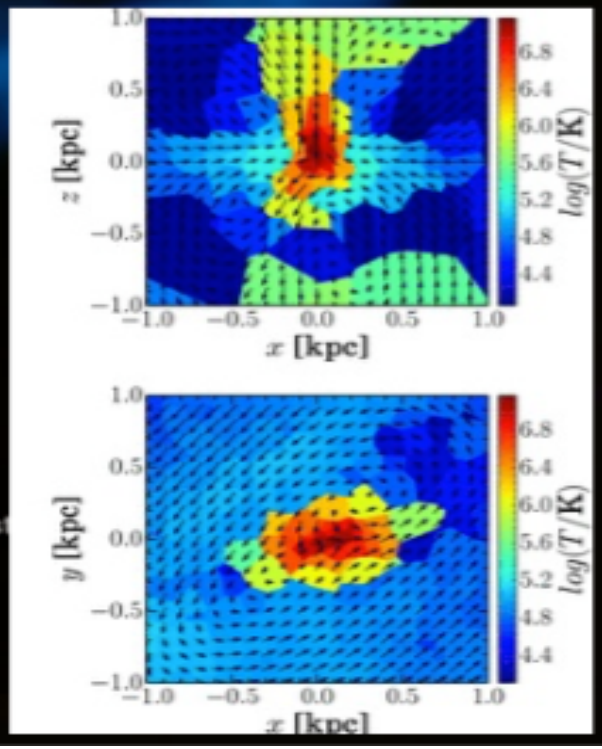
➔ Resolves accretion region (~10^4 r_g) within a single galaxy



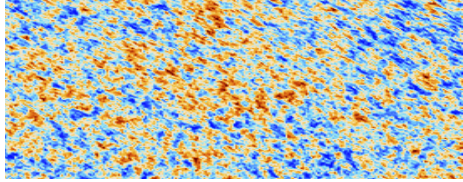
Require HPC AND HC (Human Capital)

Curtis & Sijacki., 2015

- High resolution → backflows from the large-scale jet are resolved



[How environment affects] galaxy evolution

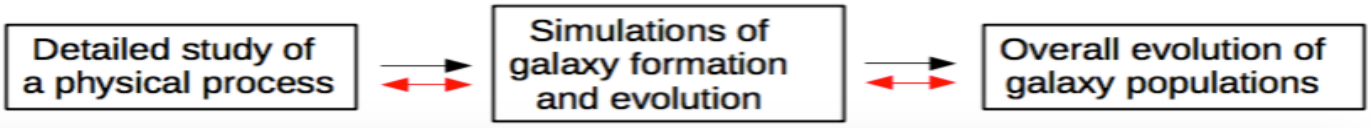


Beyond SDSS (very simplistic view):

- In-depth study of single objects or relatively small samples of galaxies
- Large (and deep) galaxy surveys, to analyse global trends

→ complementary approaches, linked by models of galaxy evolution:

Courtesy
O. Cucciati



Courtesy
L. Pozzetti

➢ *Key future observations & approaches and instruments:*

- *Medium/high Resolution spectroscopy with high multiplexing to analysis individual galaxies*
- *Spatially resolved stellar population inside galaxies with IFU*
- *Archeological approach and lookback reconstruction at z>0.3*
- *Population Synthesis models + Photoionization codes*
- *Constrains to galaxy formation models and sub-kpc scale hydro simulations*

Common aspects

HPC is common to so many area of MA1

- ✓ **Prototype stages "can survive" on moderate computational clusters**
- ✓ **But subsequent steps in both simulations for forecasting and data analysis and scientific exploitation of real data need HPC (needed also for code finalization/porting/testing)**
- ✓ **Detailed theoretical computations require HPC in many circumstances**

Even beyond MA1 … of course

these are aspects common to (likely) all MA: codes, algorithms, computation facilities, …

MA5 has specific expertise for computation facilities management and coding skills

Need for versatility / Commonalities

Many & very different projects
Many & very different types of data analysis & pipeline

Many & very different theoretical codes
Many & very different scales

Many & very different libraries
Many & very different languages

Many & very different interfaces between theoretical modeling & data

Other? ☺ ☺ → GREAT VERSATILIY OF "RESOURCES"

- **Sinergies between codes, tools, databases, projects, MAs**
- **Common HPC infrastructure with personnels dedicated also to link resources to scientific tools/needs &**
- **Common repository for tools/databases etc.**
- **→ represent in principle a strong driver for increasing scientific collaboration, sharing of expertise, improving INAF research quality**

Education

Currently:

- ✓ **limited knowledge about computational aspects and use of facilities**

Need for medium/long term efforts:

- ❖ **dedicated courses at degree level (in physics/astronomy)**
- ❖ **collaboration with informatics/mathematics departments to get skills where they are "naturally" developed**
- ❖ **collaboration with computation centres (e.g. CINECA) … we don't need only to pay only for cpu time**
- ❖ **collaboration with software & hardware companies to support grants for PhD, "borse di studio", "AdR" on projects of common interests**

Courtesy S. Borgani

Da report esteso

Da "2.3 Galassie e ISM"

Un nuovo approccio teorico alla formazione delle galassie tenta di unire i risultati dei modelli fenomenologici con quelli dei modelli semianalitici e delle simulazioni idrodinamiche.
→ Necessità di capacità di calcolo (high performance computing, HPC) attualmente non disponibili: calcolatori con milioni di cores e personale staff ed ingegneri specializzati dedicati ai codici e al software specifico.

Da "3 Infrastrutture"

Riguarda HPC: necessità di nuove competenze in *software engineering* e di facilities di calcolo di alto livello (Tier2) - problema comune a molti ambiti della MA1 (e non solo).

Da 3.3 HPC – sezione dedicata – I

Necessità di infrastrutture e personale dedicato allo high-performance computing: sempre più impellente e cruciale in tutti i settori della MA1.

- **Facilities di calcolo richieste per simulazioni numeriche e modelli semianalitici (indispensabili per interpretare i processi fisici caratteristici di ammassi di galassie e galassie)**
- **Mole di dati delle survey cosmologiche (CMB, LSS) → necessità di facilities di HPC anche per l'elaborazione statistica dei dati**
- **(Riduzione ed elaborazione dei dati osservativi spesso proibitiva su singole workstation)**
- **Elaborazione di diversi tipi di *forecast* per le missioni future**
- **Organizzazione e analisi di grandi moli di dati (*data management*)**

Da 3.3 HPC – sezione dedicata – II

- **Passato/presente: epoca in cui il singolo ricercatore poteva scrivere il suo codice personale, nel linguaggio di programmazione preferito, ed eseguirlo sulla sua workstation**

  →

- **Situazione presente/futura: la programmazione deve gestire calcoli ed interfacce a database così complessi da richiedere più sviluppatori (spesso che programmino in linguaggi *object oriented)* a cui deve seguire una procedura di validazione e documentazione**

- **Molto spesso tali codici devono essere parallelizzati, in modo da poter essere eseguiti su grandi cluster e distribuiti tra diverse CPU**

Da 3.3 HPC – sezione dedicata – III

L'ottimizzazione di questi codici non può quindi venire realizzata dai ricercatori con sole competenze, pur ottime, in cosmologia e astrofisica, ma sono diventate necessarie figure con

- **solide competenze in *software engineering,* che abbiano anche una formazione in cosmologia e astrofisica, tali da consentire un proficuo sostegno alle attività di programmazione in collegamento con gli obiettivi scientifici**

Dunque, se INAF vorrà supportare HPC, necessiterà di:

- ✓ **infrastrutture (ad es. Tier2 supercomputer)**
- ✓ **pianificazione (assunzione di personale in supporto al calcolo dedicato ai modelli teorici e pipelines)**
- ✓ **costruire una maggior e più forte partecipazione nel contesto nazionale**

Summary-Conclusion / da Report sintetico

4. Calcolo: sia per la parte cosmologica che per quella relativa alle strutture barioniche si è evidenziato:

- **necessità di un investimento nel calcolo, non solo riguardo CPU e infrastrutture, ma anche in personale**
- con l'aumento delle dimensioni dei progetti e delle simulazioni, non è più pensabile che il singolo scienziato scriva anche il software finale, non avendo, per la maggior parte, abbastanza competenze di ottimizzazione o parallelizzazione
- **servono figure con solide competenze in *software engineering* che abbiano anche una buona formazione in cosmologia e astrofisica.**