Simulations of cosmic structures: a challenge for coding and HPC architectures

G. Murante, INAF-OATs

P. Barai – INAF OATs

S. Borgani – INAF OATs & Univ. Ts

A. Biviano – INAF OATS

E. Contini – Univ. Ts

G. De Lucia - INAF OATs

S. Farris – INAF OATs

M. Girardi – Univ. Ts

D. Goz – Univ. Ts

G. Granato – INAF OATs

U. Maio – INAF OATs

P. Monaco – Univ. Ts

E. Munari – INAF OATs

S. Planelles – INAF OATs

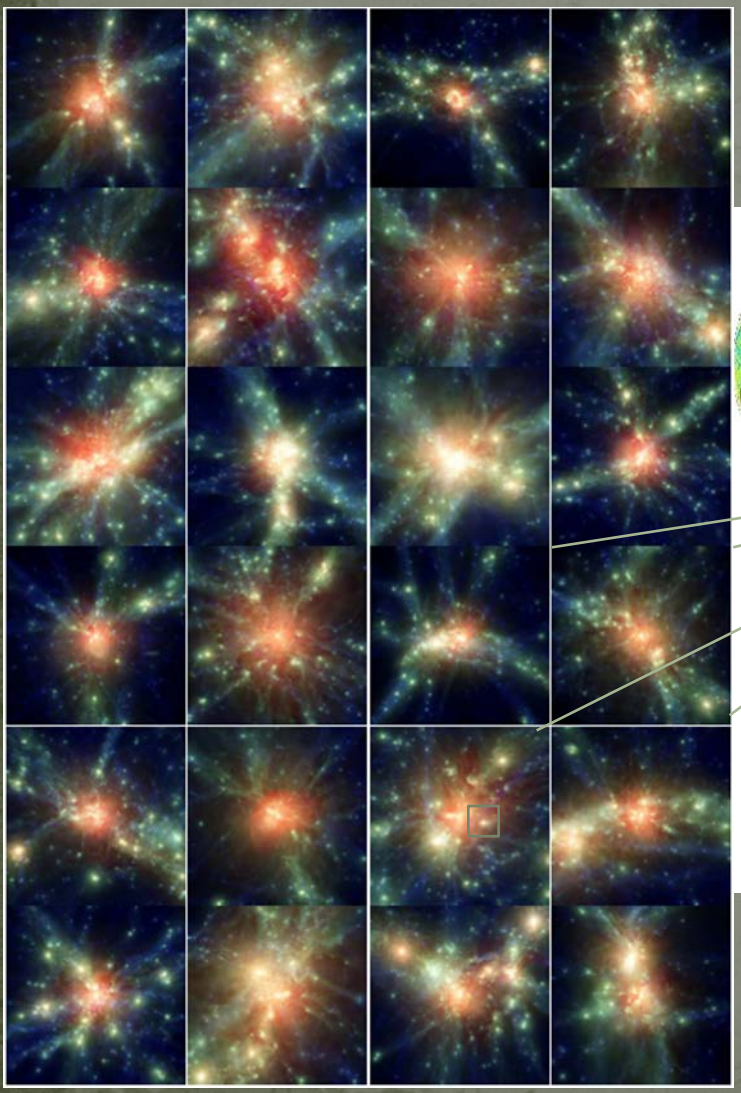
L. Tornatore – INAF OATs

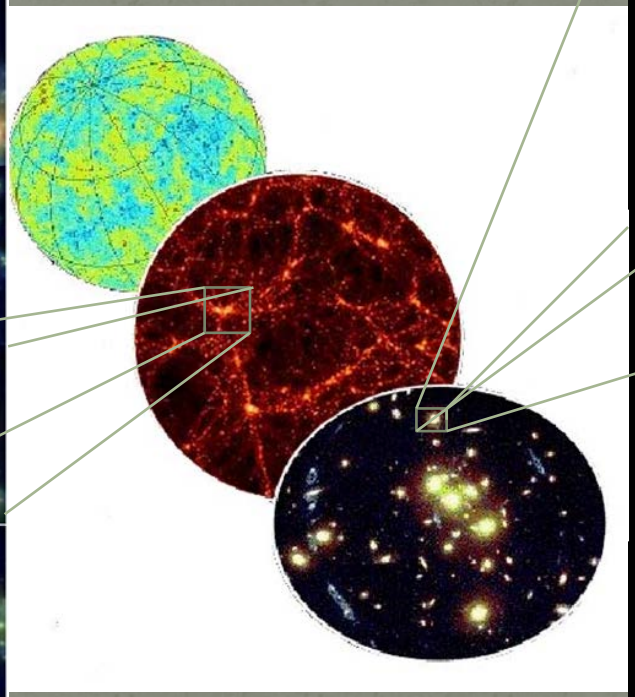
M. Viel – INAF OATs

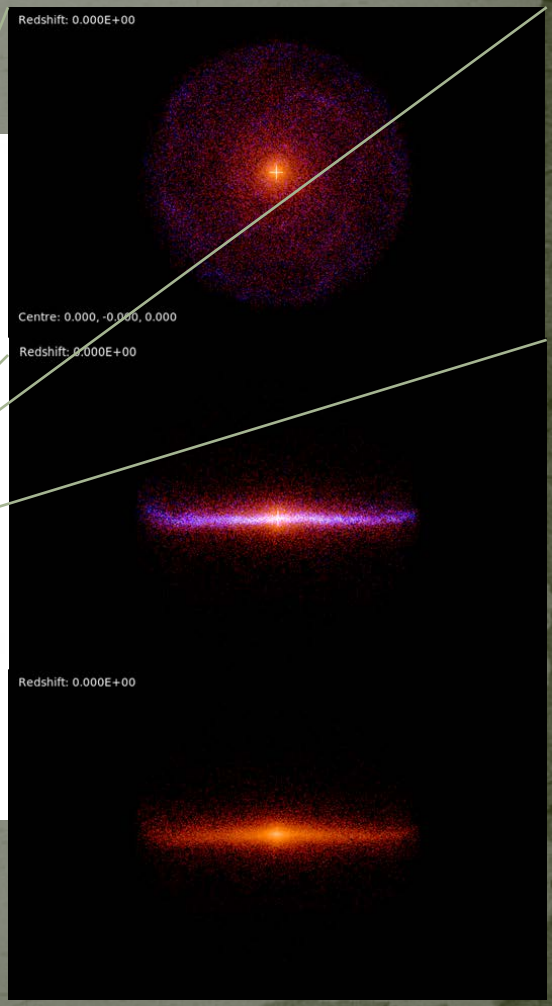
Numerical techniques and codes

- TreePM+SPH GADGET3 (GADGET2: Springel 2005)
  - Gravity: PP small scales, Multipoles intermediate scales, PM large scales
  - Hydro: SPH
  - Physics: chemical evolution & metal dependent cooling (Tornatore+ 2007), star formation & supernovae feedback (Tornatore+ 2007, Murante+ 2010,2014), uniform UV background, SMBH evolution, AGN feedback (Fabjan+ 2010; Ragone-Figureoa+ 2013)
  - In collaboration with USM-Muenchen (K. Dolag et al)
- PINOCCHIO: LPT formation and evolution of DM haloes (Monaco+ 2013)

Our activity fields



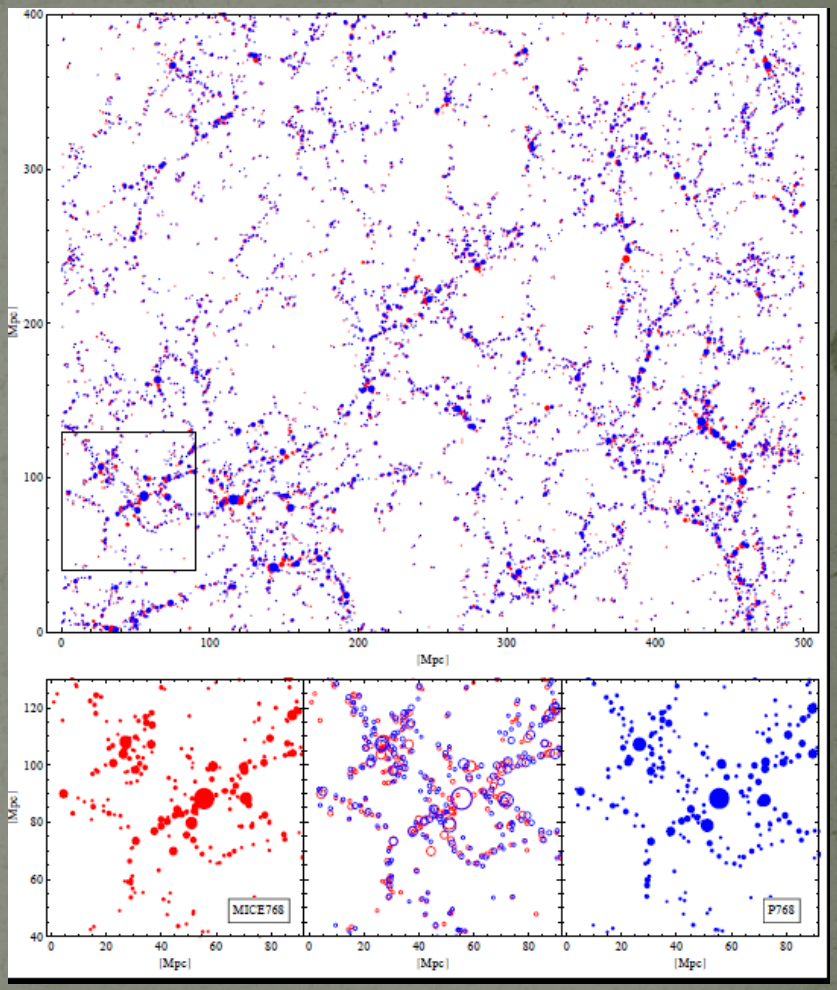




10^6 – 10^{10} calculus elements (particles)

PINOCCHIO

- PINpointing Orbit Crossing-Collapsed HIerarchical Objects: a code for the fast generation of catalogs of DM halos, with known mass, position, velocity and merger history, based on Lagrangian Perturbation Theory, excursion set theory and ellipsoidal collapse. The code is distributed under GNU GPL license.
- http://adlibitum.oats.inaf.it/monaco/Homepage/Pinocchio/index.html
- Very good performances also on Fermi
- Fast generations of thousands of catalogues… well suited for e.g. covariance matrix

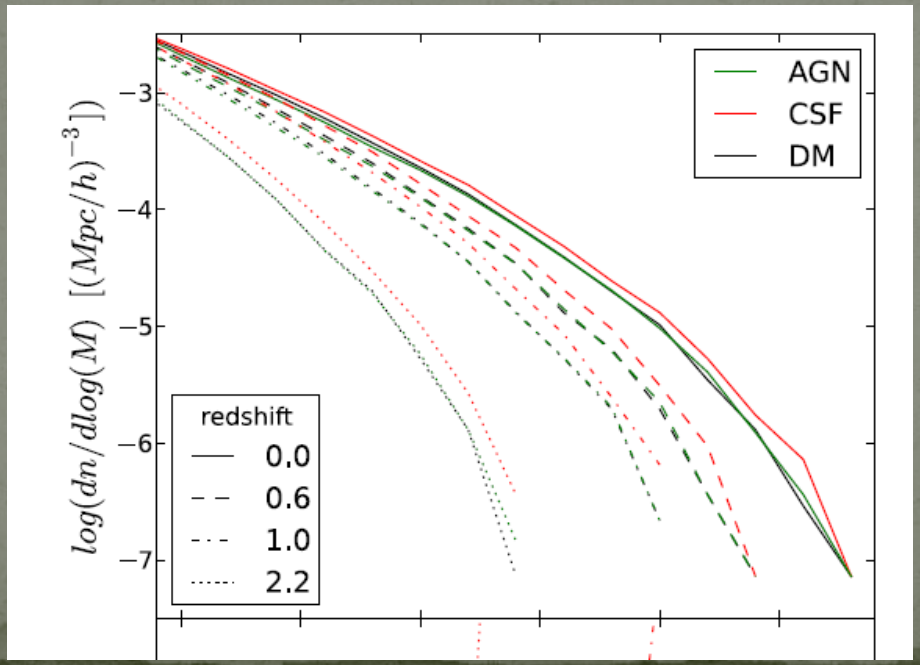


Blue: N-Body: Red: PINOCCHIO

Numerical Cosmology

Large-scale, multi-billion particle numerical simulations ($2 \cdot 1024^3$ particles, box size 410 Mpc/h)

EG: Cui+ 2014, effect of baryons and AGN on cluster mass function



REQUIRES: RAM,
Storage
CPU not a big issue

Galaxy clusters

- Zoomed-in single object simulations taken from a low resolution cosmological run
- Current state-of-the-art: Dianoga set (27 Lagrangian regions, ≈120 clusters)
- MUSIC-AGN set (in collaboration with G. Yepes, UAM)

EG: Planelles+ 2014, role of AGN feedback on ICM properties

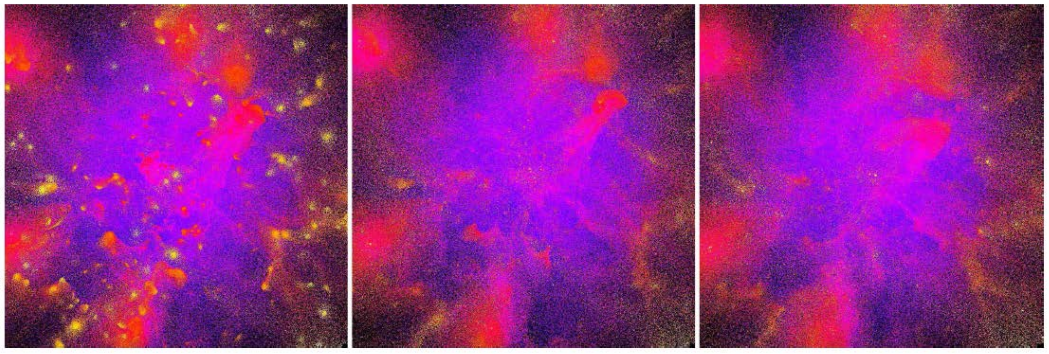
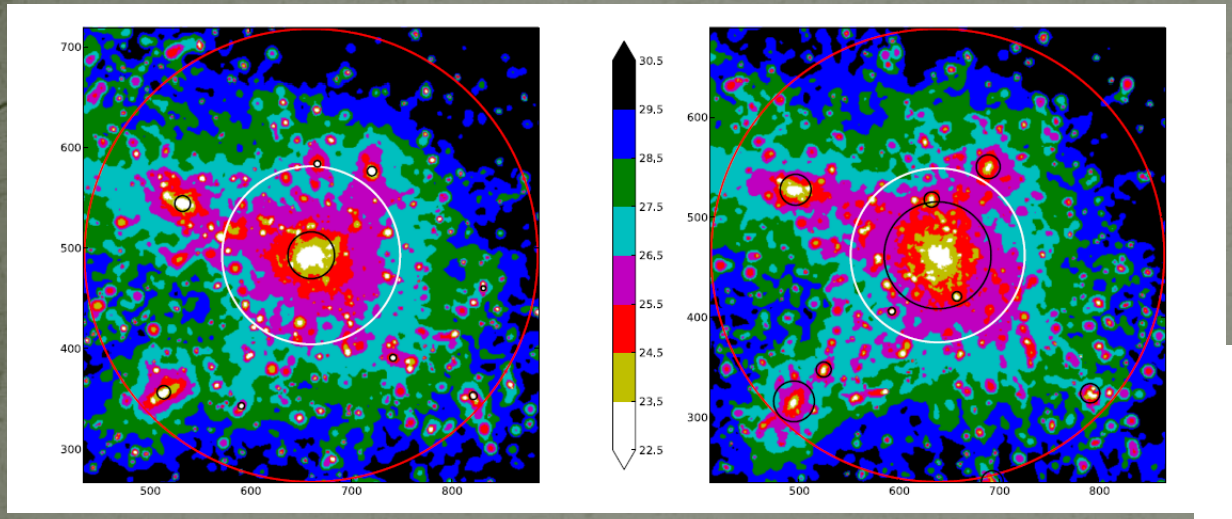


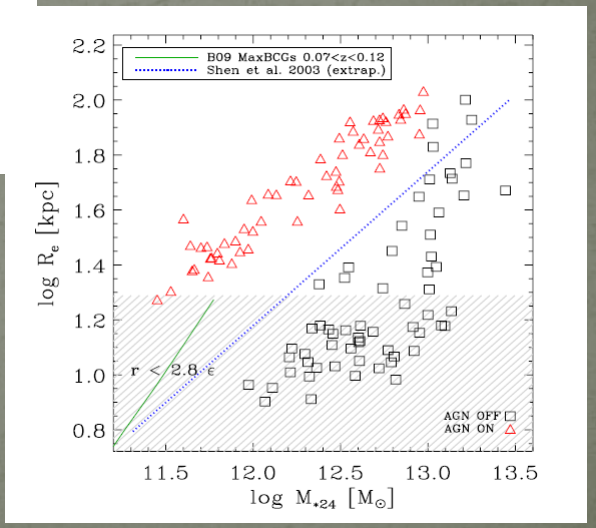
Figure 3. Temperature maps centred on a massive cluster having $M_{200} \simeq 1.3 \times 10^{15}\, h^{-1}\, M_{\odot}$. From left to right, panels show maps for the NR, CSF and AGN simulations. Each panel encompasses a physical scale of $6.25\, h^{-1}$ Mpc a side. Colder regions are shown with brighter colours.

REQUIRES: CPU
RAM not an issue at these resolutions
Storage not an issue

Properties of galaxies and diffuse stars in galaxy clusters

EG: Cui+ 2014, Ragone-Figueroa+ 2013

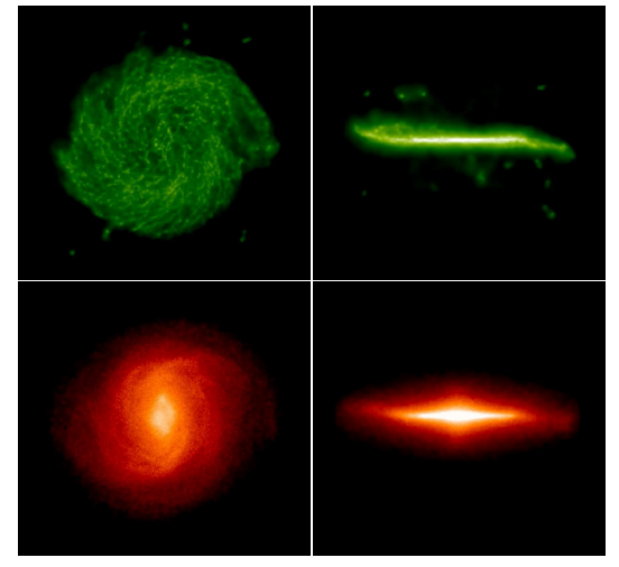


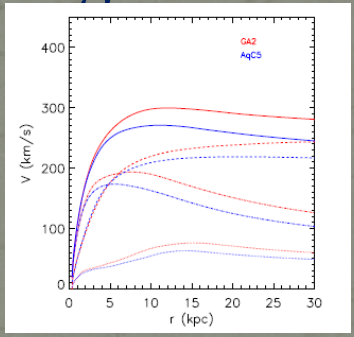


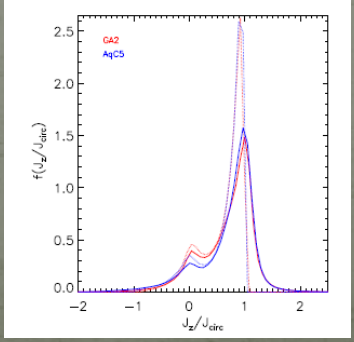
Isolated cosmological disk galaxies

- Zoomed-in simulations of carefully selected DM halos
- Current state-of-the-art: two main galaxies at several resolution, a set of ≈10 galaxies at intermediate resolution

EG: Murante+2014(submitted), Goz+2014 (submitted), properties of the simulated galaxies





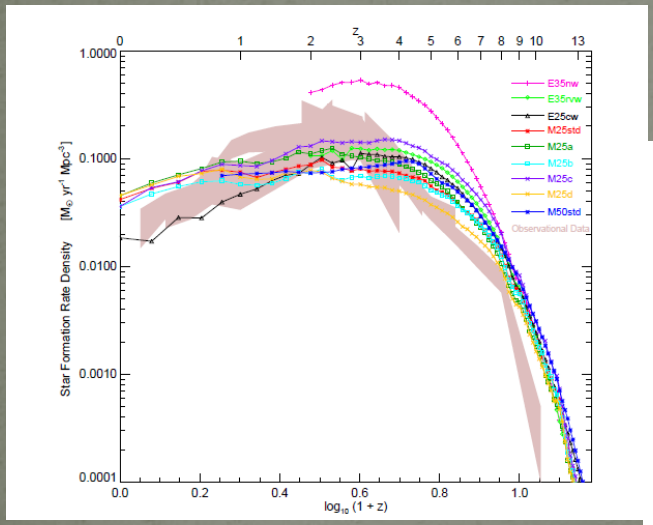


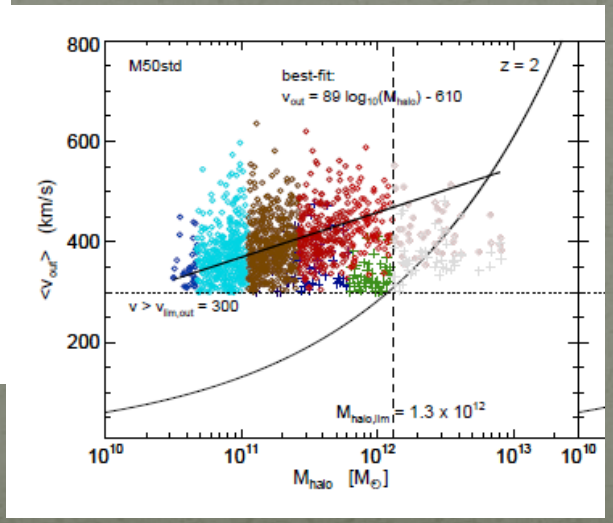
REQUIRES: CPU
RAM not an issue
Storage not an issue

Cosmological HR boxes

- Relatively small cosmological boxes, high resolution
- Needed to study simulated galaxy population properties

EG: Barai+, MNRAS submitted





REQUIRES: CPU, RAM, Storage

(Illustris, EAGLE)

HW resources

- SIMULATION EXECUTION:
  - National competitive grants (ISCRA). SP6, Eurora, PLX, Fermi. Overall CPU/core hours in the last 5 years: sp6, 645,000; Fermi, 6,300,000
  - European competitive grants (DECI). NIIF (750,000 hours, ongoing)
  - International agreements (IFT, USM, CAASTRO). MareNostrum, Raijin
- STORAGE:
  - Long term: CINECA tapes
  - Short term/analysis: local server (50 TB)
- CODE DEVELOPMENT/DEBUGGING:
  - ***Only local 16cores server***
- SMALL PROJECT/TEST RUNS:
  - CINECA/UniTS convention (but PLX overcrowded)
  - Local 16cores server
  - Incoming: OATs/SISSA/ICTP agreement. OATs will have access at reasonable prizs to the new 1000cores-scale cluster installed at ICTP and will have a member in the steering committee
- POST PROCESSING:
  - Local 32 cores, 256 GB server

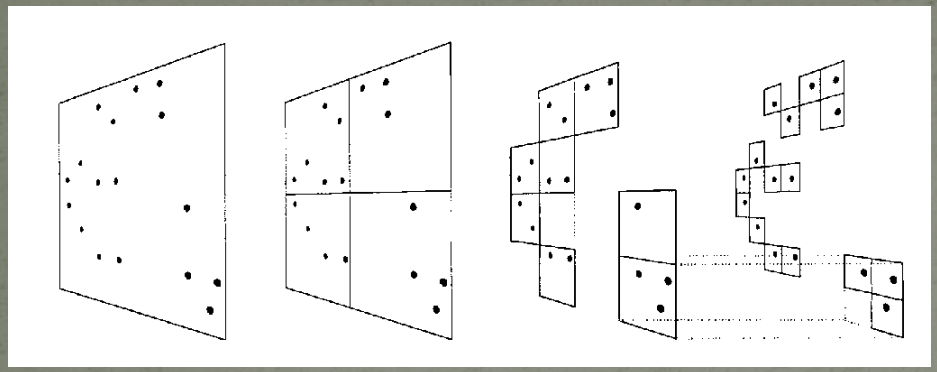
HW resources criticalities

- CODE DEVELOPMENT/DEBUGGING requires a tiny amount of resources (however more than 16cores would be better) but immediate availability
- SIMULATION TESTS/SMALL PRODUCTION RUNS require a small amount of resources , quick availability and simple access (*I mean: no competitive grant for these!*)
- We don't have issues with CPU time for LARGE PRODUCTION RUNS (but: see HW architectures problems…) since we usually get time in competitive grants calls
- STORAGE is a problem.

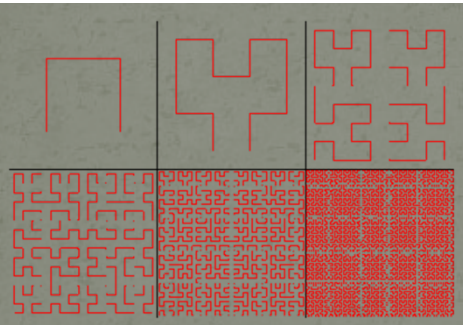
Funding (numerical sector)

- 3 PRIN MIUR (P.I. S. Borgani); 1 PRIN INAF (P.I. GM)
- Partecipation in European projects (ITN): CosmoComp
- 2 ERC grants: G. De Lucia, M. Viel
- These sources of funds allowed us to set up our local system – fundamental for the postprocessing (otherwise difficult/impossible)
- **No money explicitly dedicated to numerics was ever granted by INAF or MIUR**. But: INAF/CINECA convention till 2009; MIUR contribution to Fermi

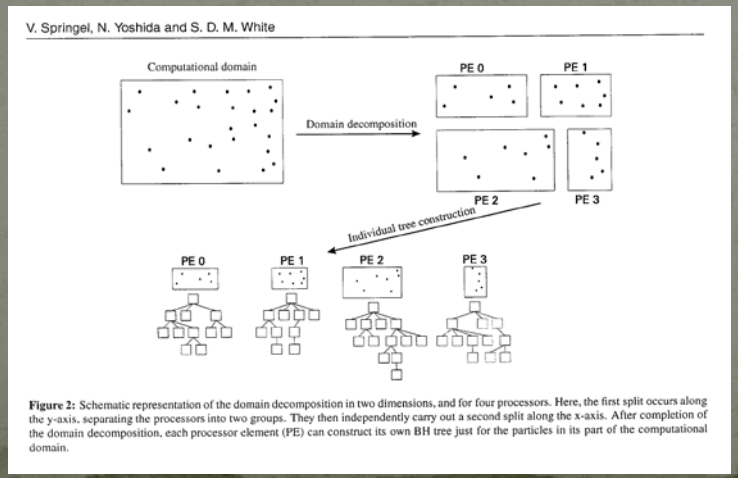
Code & architectures (GADGET3)



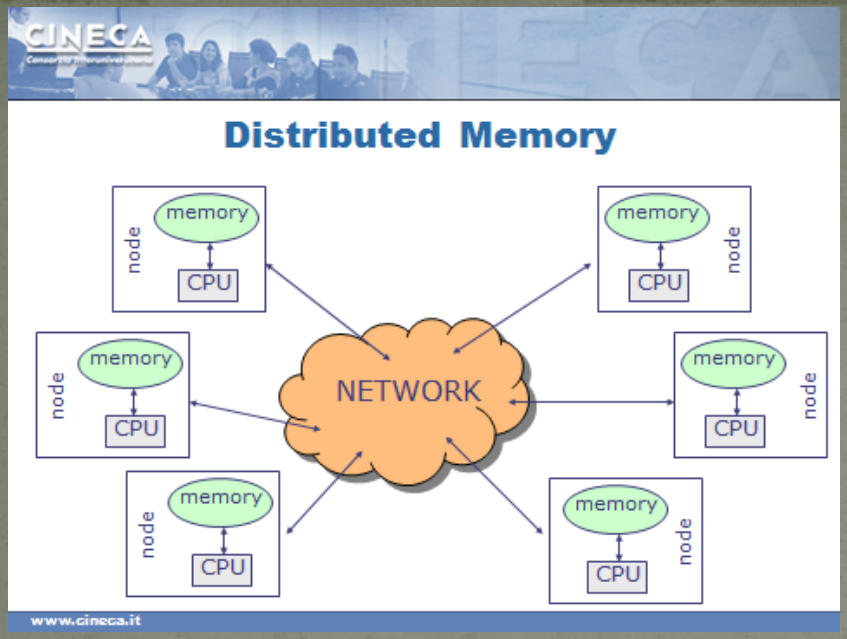
A tree is used for gravity computation
(approximate, but less communications)



DOMAIN DECOMPOSITION using a Peano space-filling curve: work-load balance at the cost of memory unbalance



Computation assigned at single MPI tasks.
Inside them, OpenMP for shared memory parallelization
Multiple timestep problem: # of active tasks limited by # of active particles



YESTERDAY architecture:

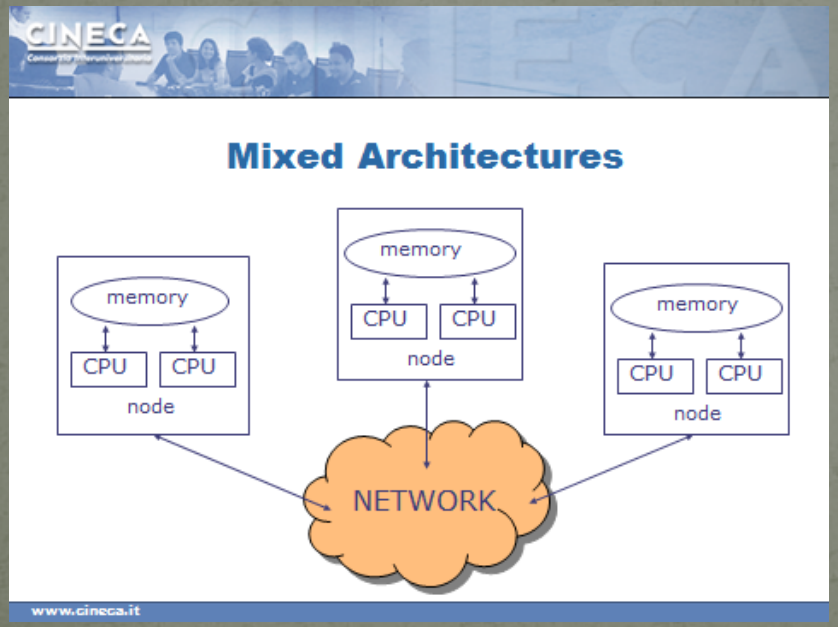
POWERFUL chips

LOTS of memory

Memory is distributed

O(100) CPUs

OK!



TODAY architecture:

LESS powerful chips

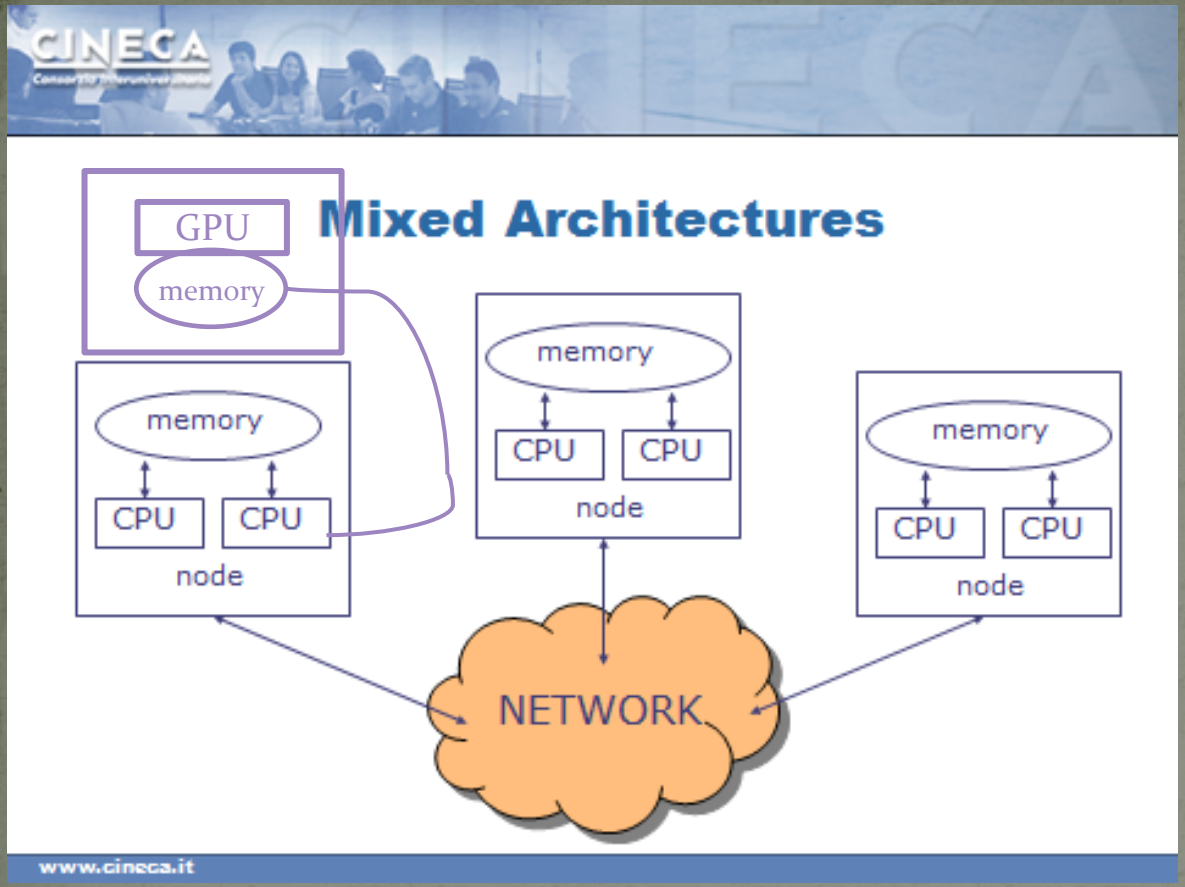
1GB of memory/core

O(10) chips per computing unit

O(1000) computing units

MIXED memory paradigm

NOT SO BAD....



TOMORROW architecture:

Add accellerators and cores, reduce RAM/core (RAM is energivorous)

SLOW connection accel/CPU RAM (now improving)

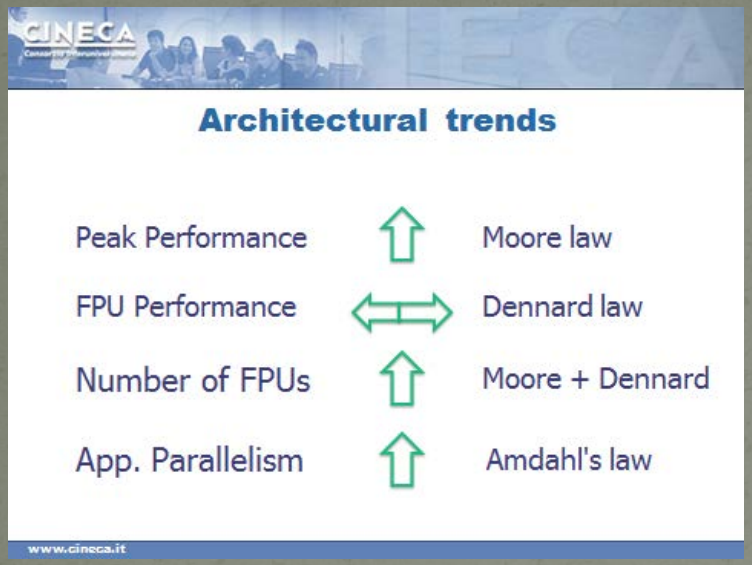
OK for doing lots of calculations always on the same dataset

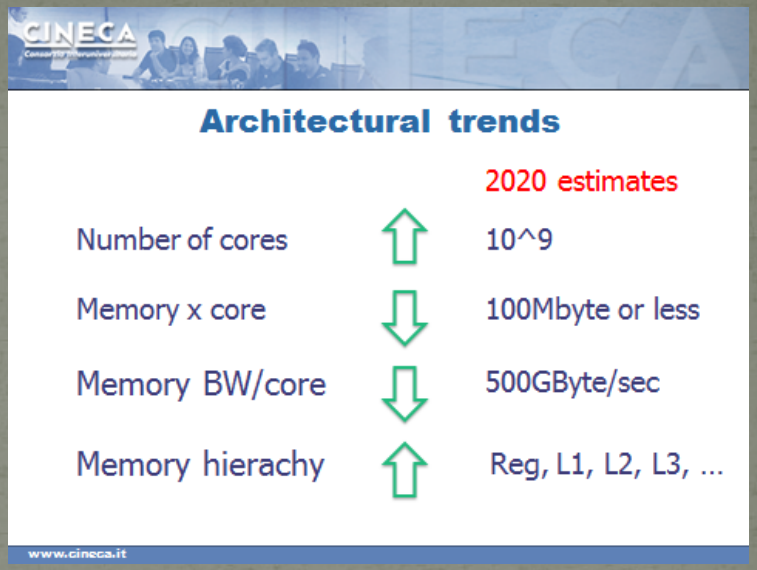
BAD!

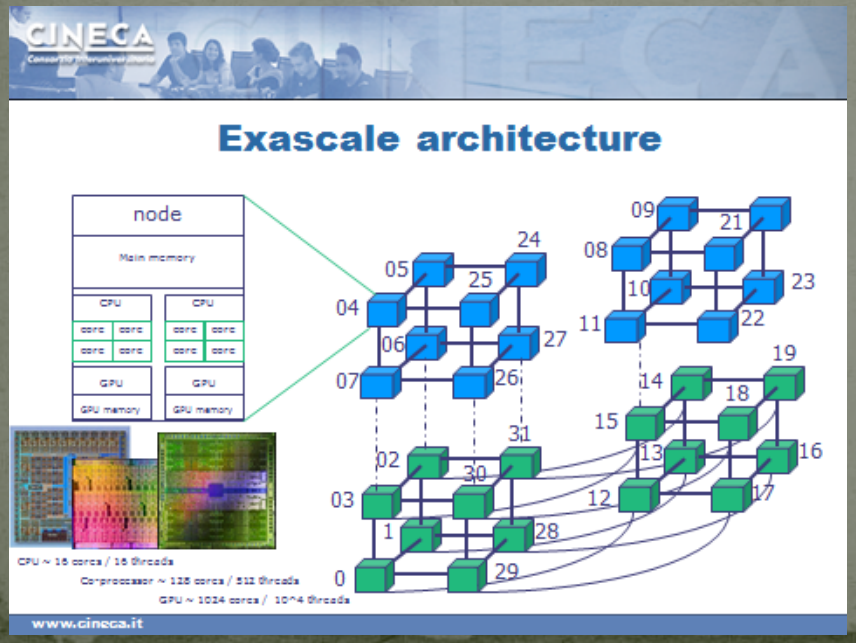
Accelerators:

FAST floating point operations on a vector of numbers

Towards exascale



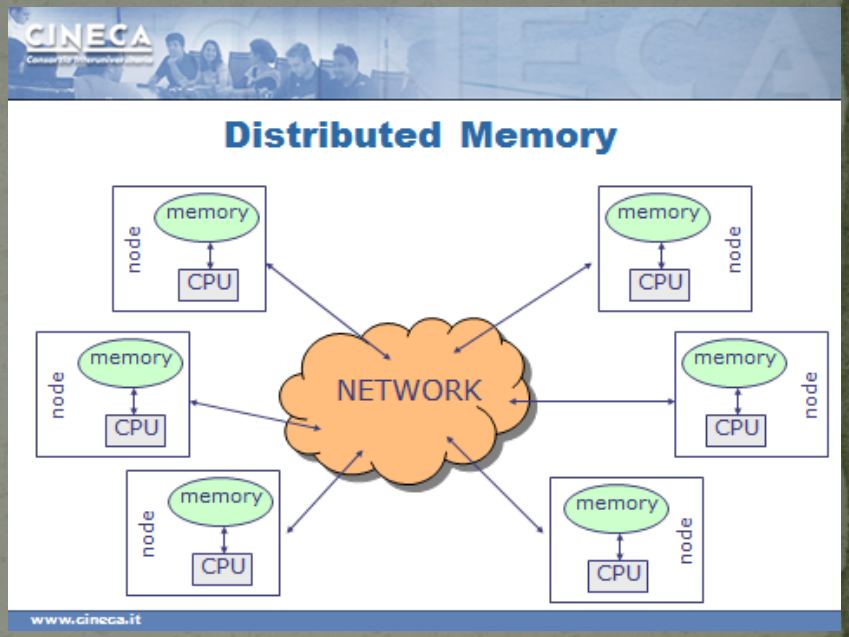




- Exascale computers: blocks of «weakly connected» petascale units
- NO substantial increase of available CPU power for a «single» simulation
- Very limited memory per core
- Well suited for *high throughput* sets of simulations
- I/O-storage problem
- Technological breakthroughs possible?

HW requirements for current codes

- OLD-STYLE machines still best suited to our kind of codes (Exemplum: ERIS run 9 months on an SP6!)
- Currently GADGET3 performs best on clusters
- BlueGene kind of computers are a **full disaster**
- *GRID/CLOUD/ DISTRIBUTED COMPUTING is **totally** useless for the kind of problems we deal with.*



(this problem is worsened when it results to be **impossible** to tune the code, due to queue managing choices)

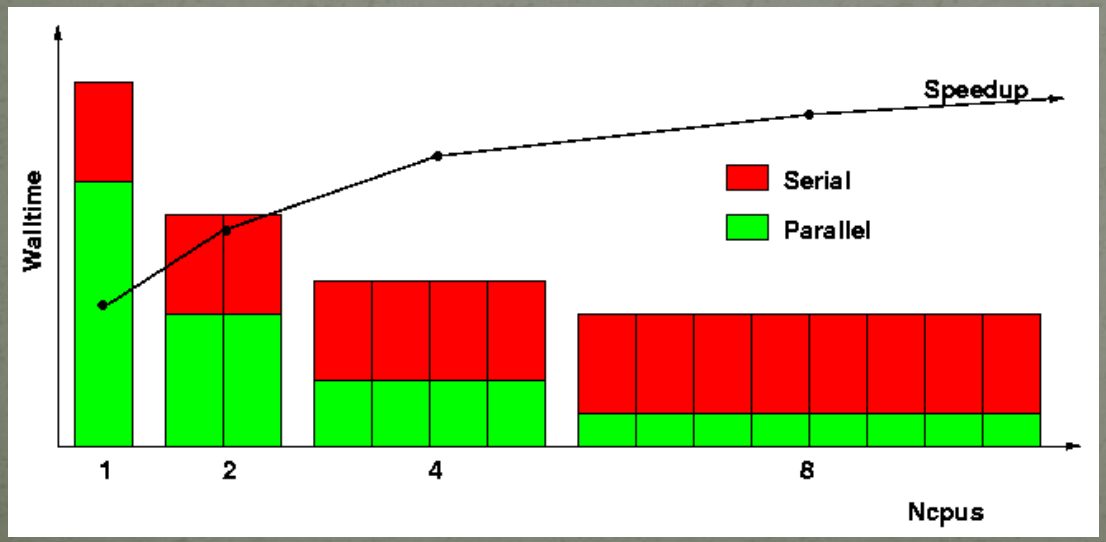
Programming paradigms

- Yesterday: MPI, distributed memory
- Today: MPI+Threads, mixed memory
- Tomorrow: need careful code architecture

  we can't allow to have low parallelization fractions

  ….need interaction with computer scientists?

  need to detach portion of physics to accelerators



maximum speedup tends to
$1 / (1 - P)$
P= parallel fraction

1000000 core

P = 0.999999

serial fraction= 0.000001

Future code developement

- Asynchronous parallelization. Analysis of code atomic tasks and their dependencies, server-client task manager, independent tasks taking care of IO and communications. *Needs dedicated personnel*.
- Asynchronous physics module managing. We need to perform all possible additional physics while gravity and hydro are being computed, and possibly to perform it using accelerators.
- A collaboration on this with data center/industrial partners very appreciated

Conclusions

- **Good** news:
  - our codes are currently state-of-the-art and are able to run on most of the current HPC systems
  - we get enough CPU time for large programs via competitive grants
  - we get some money for our numerical projects via competitive grants, Italian and European, and thank to this…
  - …we have a barely sufficient postprocessing set of servers

  ➡ we are currently able to produce internationally competitive numerical science.
- **Bad** news:
  - our codes are **not ready** for NG **HPC/Exascale**
  - they will **not** be, unless we hire dedicated personnel for coding
  - we need easy access to Tier 1,2 systems for developing, debugging and testing codes and for small projects

  ➡ without this, the risk to lose competitivity is very strong.