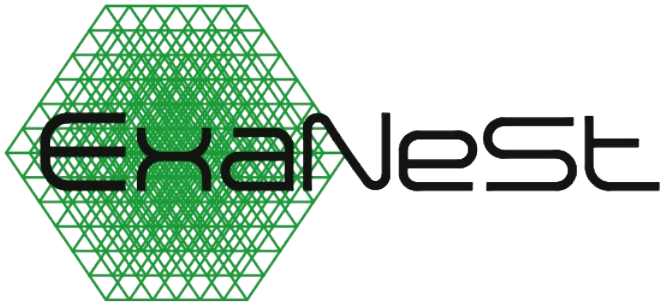
N-body codes
(migration into Exascale Era)

David Goz

With

S. Bertocco, L. Tornatore,

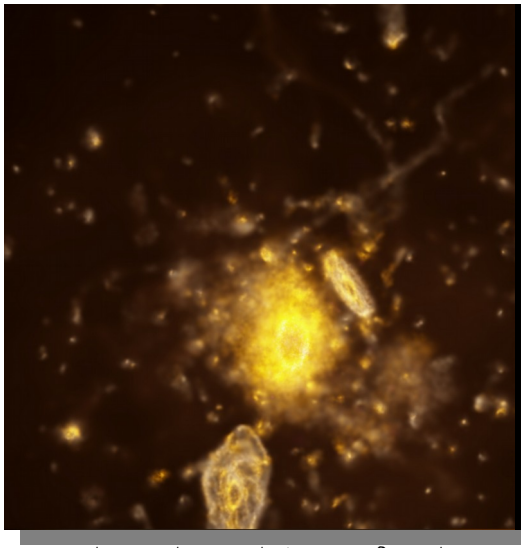
G. Taffoni, and M. Molinaro



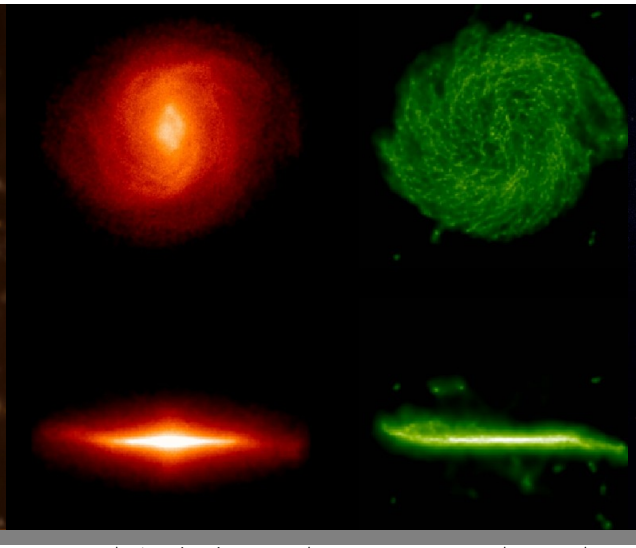
Horizon 2020

Next generation computing roadmap

Most of us rely on numerical codes to perform calculations.



Cosmological simulation of galaxy formation using GADGET code (Springel 2005).



Simulated disk galaxy in cosmological environment at present epoch (Goz+2015).

NGC 4414, a typical spiral galaxy in the constellation of Coma Berenice, is about 60 million light-years away from Earth (Credit HST).

- HPC numerical simulations are one of the more effective instrument to compare observations with theoretical models;
- the new generation of observational facilities also implies high performance data reduction and analysis tools.

Why Exa-scale?

" Crucial problems that we can only hope to address computationally require us to deliver effective computing power orders-of-magnitude greater than we can deploy today ".

DOE's Office of Science, 2012

"EXA-scale" is the necessary upscale step that HPC needs to achieve in the next years.

It is defined as the frontier of a sustained performance around 10^{18} flop/s

There are deep consequences in the way we design, write, and optimize scientific codes.

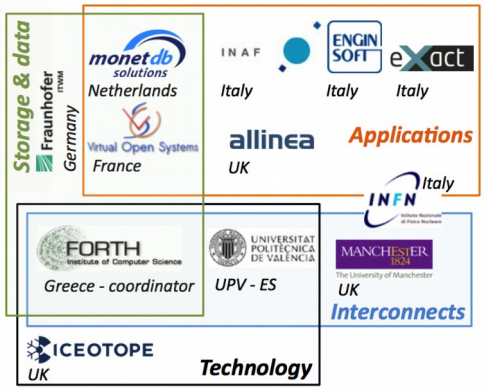
ExaNeSt European project

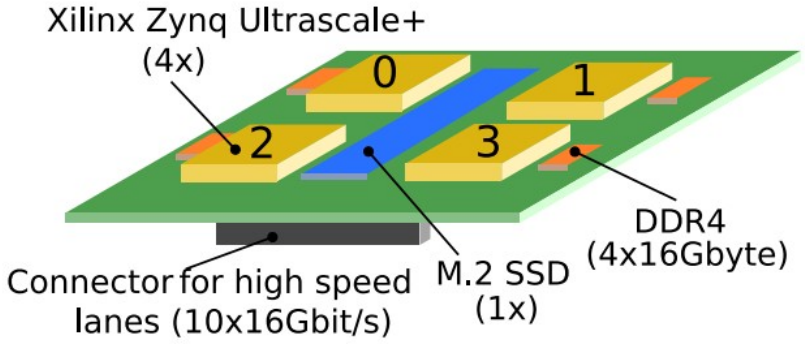
The **Horizon2020 ExaNeSt project** aims to demonstrate the feasibility of a European technology based ExaScale HPC system.

Who we are: the ExaNeSt consortium combines industrial and academic research expertise.

How we do it: following a **co-design** approach,

- applications drive the HW development and test it;
- applications are **re-designed to develop new HPC SW** able to exploit exascale HW.





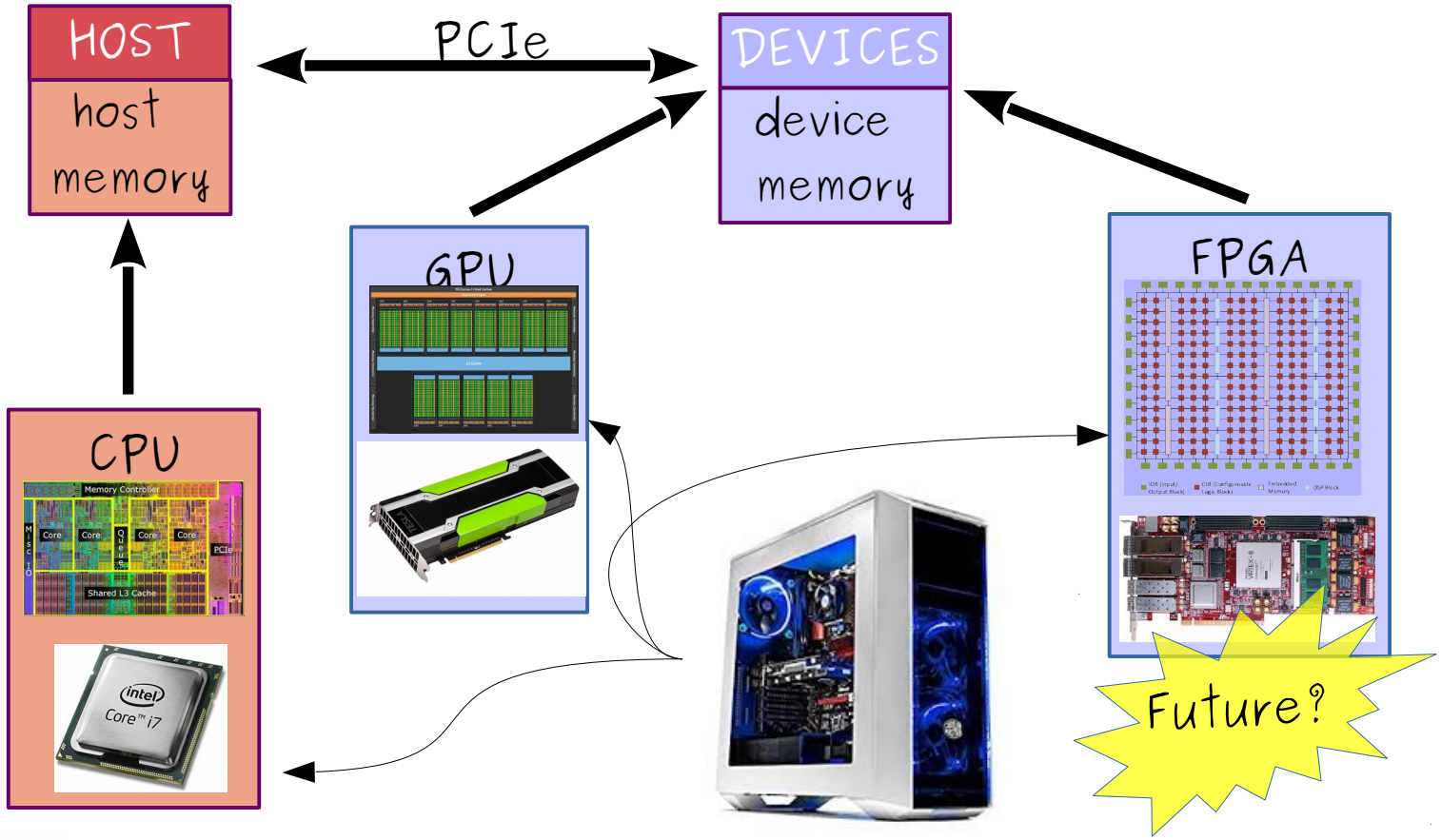
The ExaNeSt compute Quad-FPGA daughter-board.

ExaNeSt compute unit:

- 4 Xilinx Zynq Ultrascale+ FPGAs;
- 4 ARMv8 cores @1.5GHz per FPGA;
- 16 GB of DDR4 memory per FPGA;
- one NVM SSD storage device.

Heterogeneous hardware

Node level heterogeneous architectures compared to traditional CPUs offer **high peak performance.**



Embedded & mobile hardware

System-on-Chip (SoC) heterogeneous hardware compared to traditional hardware is more **energy and cost efficient.**

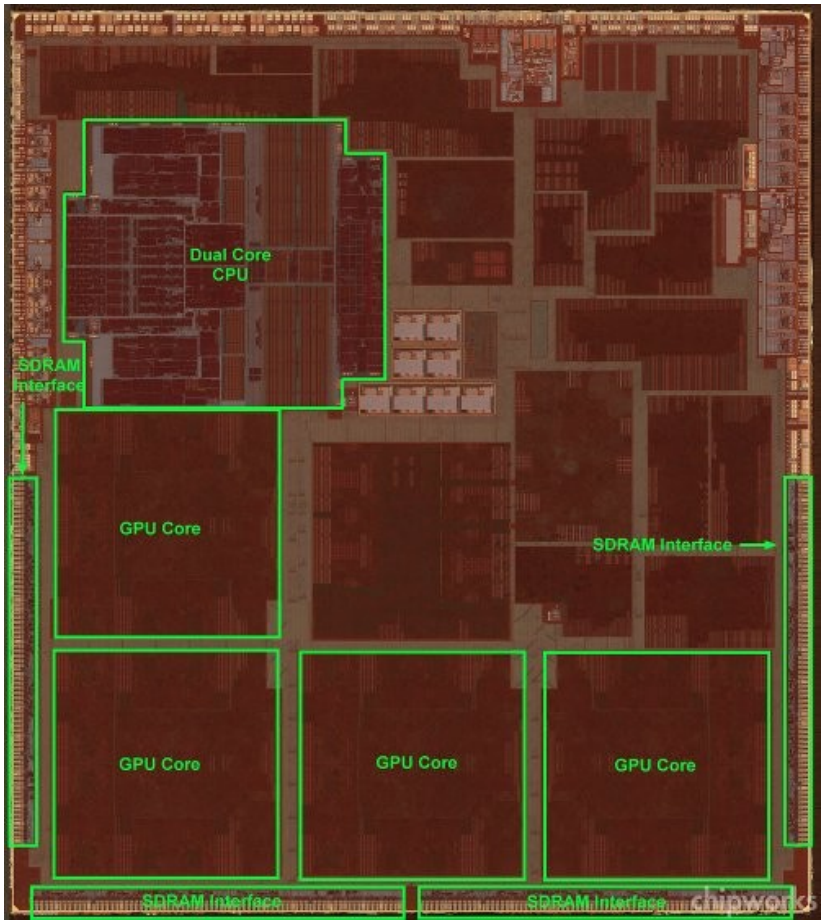
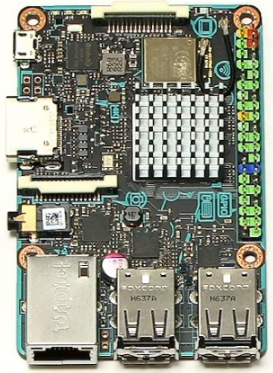


Photo from ChipWorks

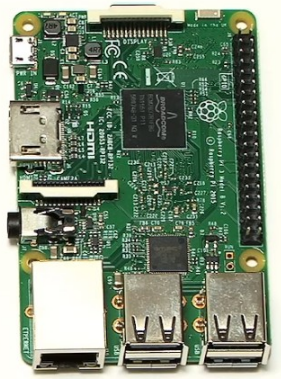
- SoC are in contrast to the motherboard-based PC architecture;
- SoC integrates CPU/GPU/memory interfaces into a single chip;
- SoC has reduced modularity and replaceability of components;
- energy-efficiency is the main concern;
- ARM is the *de facto* SoC technology.

ODROID-XU4

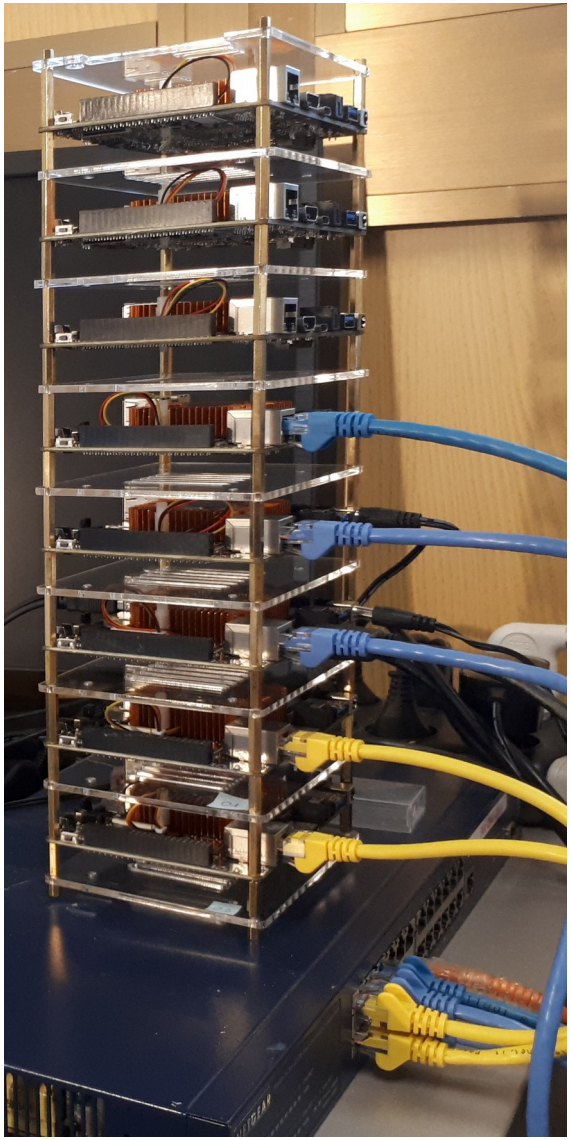
Tinker Board



Rasberry Pi 3



INCAS (INtensive Clustered Arm-SoC)



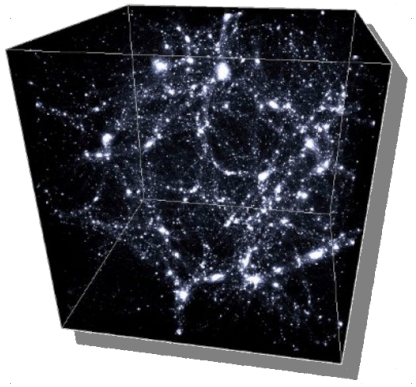
Cluster components.

Nodes available	8
SoC	Rockchip – RK3399
CPU	Six-Core ARM 64-bit (Dual-Core Cortex-A72 and Quad-Core Cortex-A53)
GPU	ARM Mali-T864 MP4 Quad-Core
Ram memory	4GB dual-channel DDR3 (per node)
Network	1000 Mbps Ethernet
Power	DC12V – 2A (per node)
OS	Ubuntu 16.04 LTS
Compiler	gcc version 7.3.0
MPI	OpenMPI version 3.0.1
OpenCL	OpenCL version 2.2
Job scheduler	SLURM version 17.11

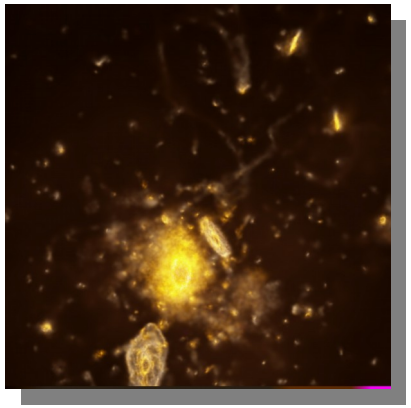
The INAF astrophysical codes in ExaNeSt

Hy-Nbody:
direct N-Body code to simulate cluster dynamics and close encounters.

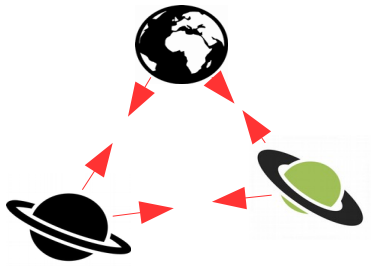
PINOCCHIO (P. Monaco, T. Theuns & G. Taffoni, 2002):
a fast code, based on Lagrangian perturbation theory, to generate catalogues of cosmological dark matter halos and their merger history.



GADGET (V. Springel 2005):
is an N-body and hydrodynamical code for large-scale, high-resolution numerical simulations of cosmic structure formation and evolution.



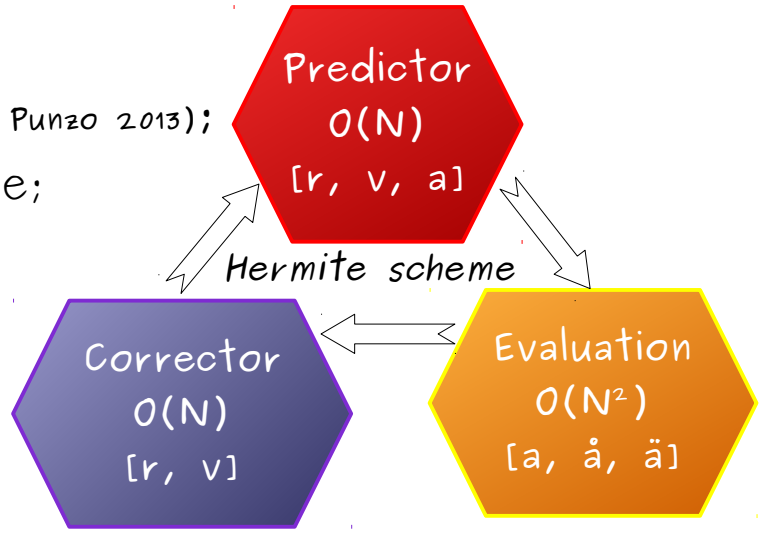
Hy-Nbody: direct N-body code



Hy-Nbody (hybrid N-body) is a direct N-body code suitable for studying the dynamical evolution of stellar systems.

Features:

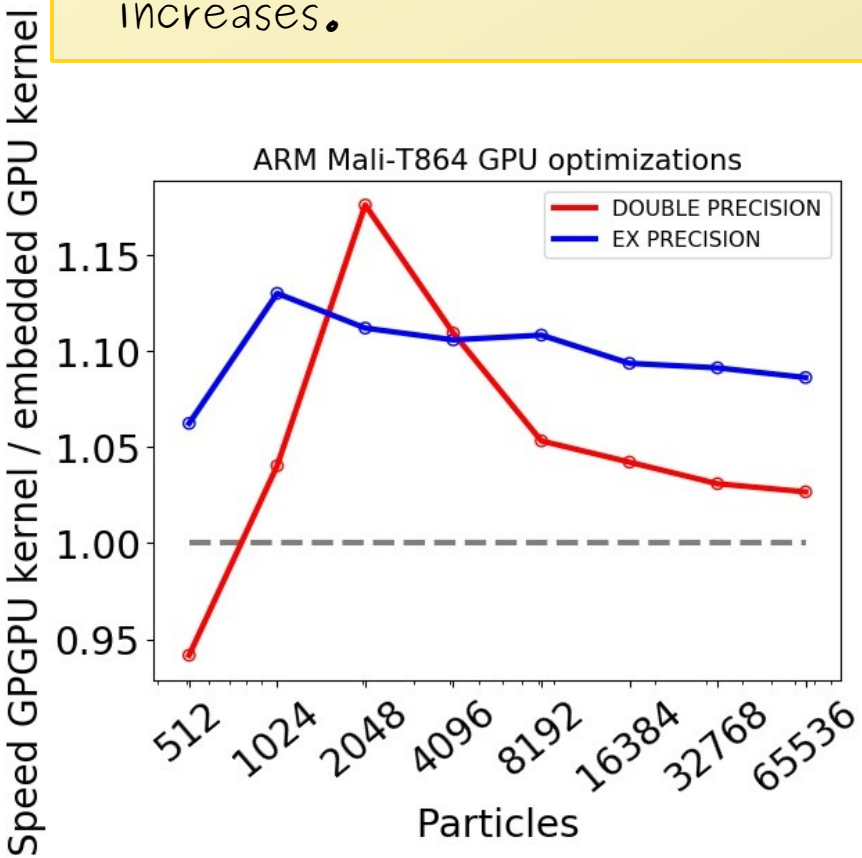
- *based on* **HiGPUs** (R. Capuzzo-Dolcetta, M. Spera, D. Punzo 2013);
- *Hermite 6th order* time integration scheme;
- exploitable devices: CPU/GPU/FPGA;
- parallelization schema:
  - host code : MPI + OpenMP;
  - device code: OpenCL.

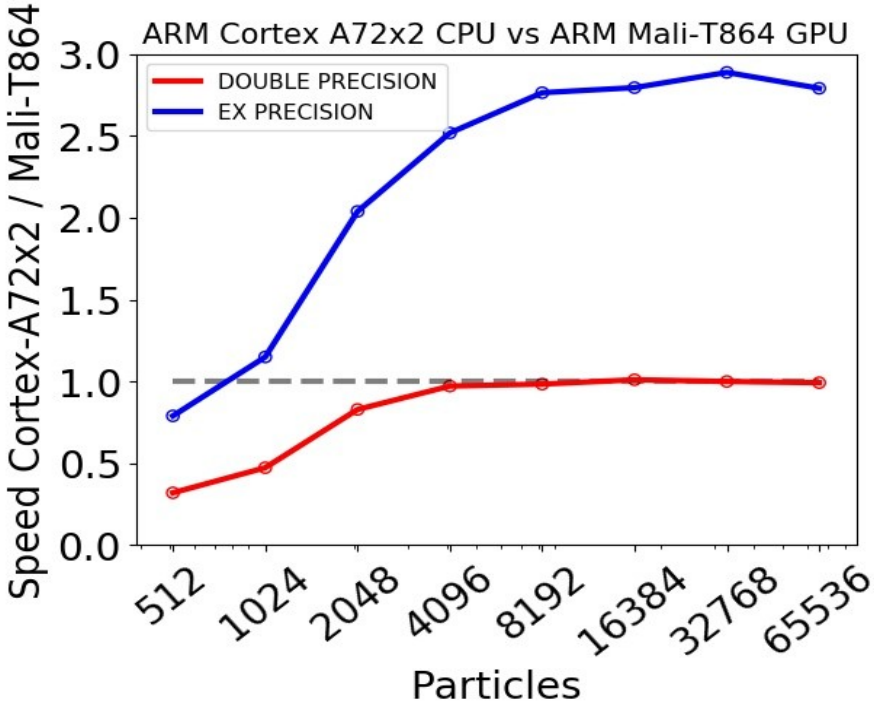


ADOPTED STRATEGIES	MOTIVATION
vectorization	increase the number of ops
exploitation of local memory	enable memory burst mode
extended-precision (EX) arithmetic	trade-off between accuracy and resource usage
host-device communication vs memory-mapping	e.g. discrete vs embedded GPUs

Hy-Nbody: results on ARM SoC (INCAS)

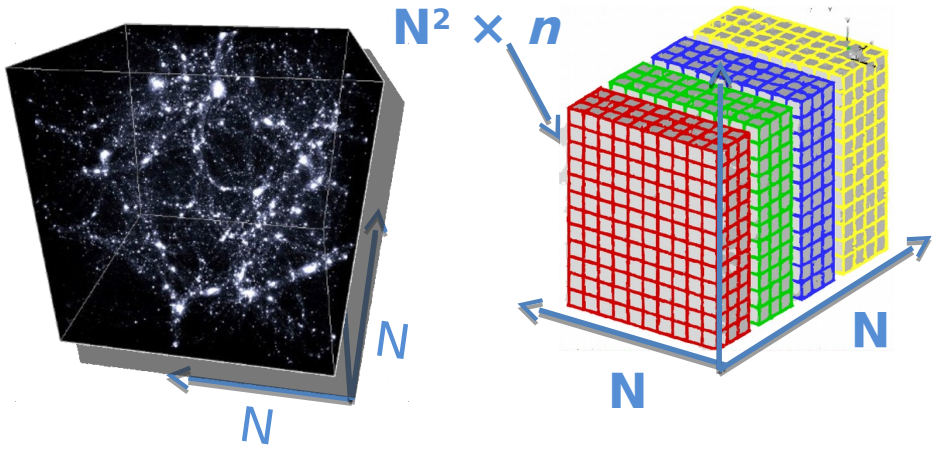
- Optimization strategies for high-end GPGPU computing lead to worse performance on embedded GPUs;
- embedded GPUs appear to be attractive from a performance perspective as soon as their double-precision compute capability increases.

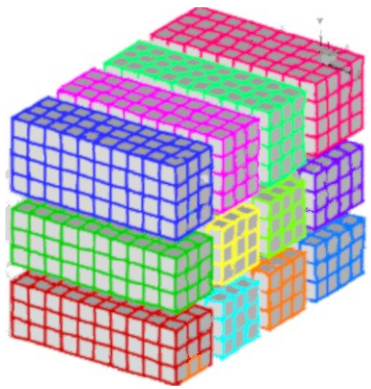




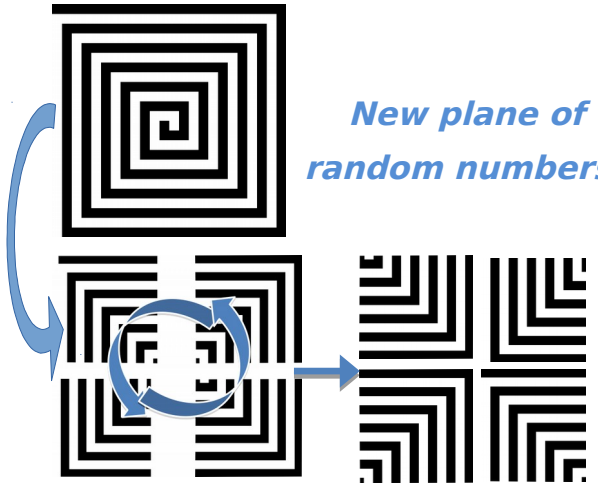
Re-engineering of PINOCCHIO code

Pinocchio is being used in **EUCLID** EU project to produce a large sample of realizations of the Universe.





2D FFT decomposition



New plane of random numbers

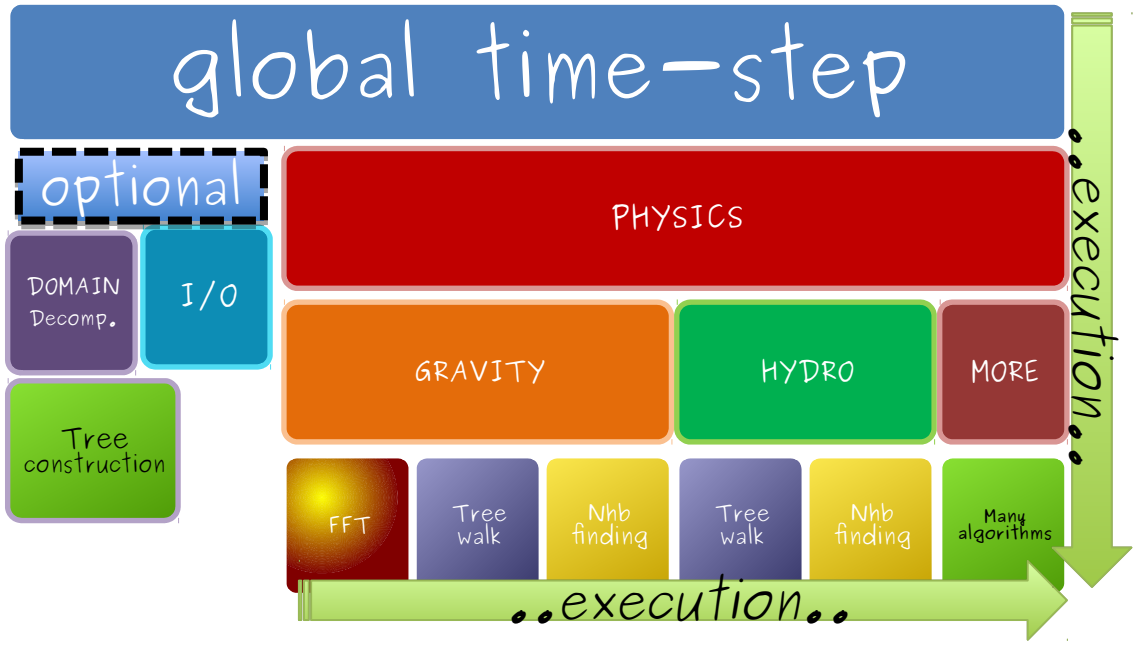
Old version of PINOCCHIO:

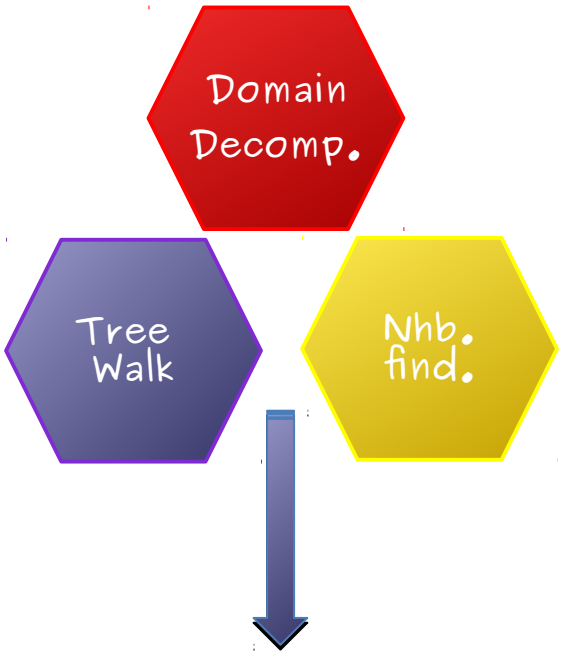
- use of FFT(W) with 1D spatial decomposition:
  - N calculating task at most;
  - memory limitation when $N \geq 10^4$;
- initial power spectrum is entirely replicated among MPI tasks and not distributed.

New version of PINOCCHIO:

- use of FFT(W) with 2D-3D spatial decomposition;
- re-designed algorithm to generate power-spectrum;
  - it has the same properties and symmetries;
  - each MPI task has only its portion of the pseudo-random field.

GADGET and exa-scale





Extremely complex code:

- many physical processes and diverse algorithms;
- rigid procedural design (MPI tasks handle global operation blocks).

Jumping to exa-scale codes requires:

- take into account NUMA-hierarchy;
- re-design algorithm in task-based, data-driven perspective.

[ongoing] working on kernels for:

- tree walk;
- neighbors finding;
- domain decomposition.

GADGET application requirements

Requirement	Platform-independent	Use case 1: single galaxy cluster at low resolution ~6 x 10^6 particles	Use case 2: single galaxy cluster at medium resolution ~19 x 10^6 particles
Language/ProgModel	C + MPI/OMP	NA	NA
Use-case platform	NA	Linux Cluster, Intel Xeon E5v3 processors, InfiniBand, 250 GB per node, 6.25 GB per core	Linux Cluster, Intel Xeon E5v3 processors, InfiniBand, 250 GB per node, 6.25 GB per core
No of Nodes	NA	160	240
Kernel	Tree+multipole expansion methods, SPH	NA	NA
Storage footprint	**~100-150 Bytes per particle per snapshot; for production cases ~100 snapshots. Storage 1-15 TB**	**~ 70 GB**	**~ 200 GB**
Memory footprint	Depending on the physics implemented and the number of cores used, from 0.004 to 0.01 MB per particle	~ 50–60 GB	~ 150-160 GB
Time to solution	NA	~ 7 x 10^6 sec	~ 50 x 10^6 sec
Throughput	Estimated using Performance Counter	~ 10 PFlop	~ 70-100 PFlop

Outputs

- Simulation outputs:
  i. ~ 100 snapshots with ~33 blocks (GADGET format);
  ii. additional informations (e.g. SFR, metals, BHs, …);
  iii. outputs from Friend-of-friends (Fof) algorithm.

- PostProcessing outputs:
  i. merger trees (dark matter and galaxies);
  ii. galaxy properties (global and profiles) extracted with suitable postprocessing codes;
  iii. intracluster medium properties computed and stored locally for all physics schemes included in hydro simulations.

Outputs, postprocessing, …

- Other future outputs to add:
  i. Infrared maps of dust in galaxies – e.g. GRASIL3D code (R. Domínguez-Tenreiro et al. 2013);
  ii. ray-tracer software tool that supports the effective visualization of cosmological simulations data – e.g. Splotch code (Dolag et al. 2008);
  iii. …

Potential users

- **Researchers already using the same simulations:**
  i. quick comparison of simulated data;
  ii. new models/algorithms validation;
  iii. learning developed VO services.
- **Wider numerical community:**
  i. comparison with their set of simulations/models.
- **Observational astronomers:**
  i. comparison with observational data;
  ii. postprocessed data free to be downloaded.
- **Students:**
  i. projects for bachelor students to familiarize with database;
  ii. development of new VO tools.
- **Outreach:**
  i. VO services to be used for visualizing data (also on the fly) and to search through different simulations/models.

Conclusions

- The usage of heterogeneous computing in scientific research (not only HPC) appears to be inevitable;
- we will be forced to re-engineer our applications (not only for HPC) in order to exploit new exascale computing facilities;

- A guarantee for a deep scientific impact means:
  i. results have to be shared with the wider scientific community and with any "potential user";
  ii. provide not just an archive of raw data, but:
    1) all information because simulations are not "black boxes";
    2) detailed information of the database content;
    3) pre-processed data (e.g. merger trees, data of interesting simulated regions,…);
    4) interactive tools for quick analysis and a flexible visualization.
  iii. continuous collaboration.