# Shall numerical astrophysics step into era of exascale computing?

Dr. Giuliano Taffoni

National Institute for Astrophysics – OATs

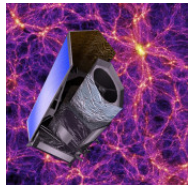
# Exascale

- Why
- What
- First ”experimental” results
- Final thoughts

# Future Challenges of Cosmology: the Role of Numerical Astrophysics

Fundamental laws of physics at cosmological scales (~$10^9$ light years): Dark Matter, Dark Energy, Nature of Gravity, ....

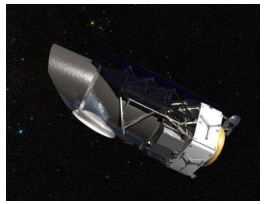
Euclid



SKA

LSST

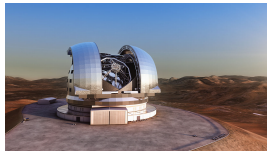
WFIRST



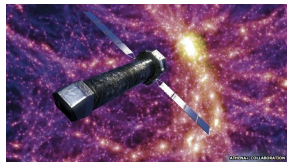
Formation and evolution of cosmic structures: from first stars and BHs, to the large-scale structures in the nearby Universe

JWST

E-ELT

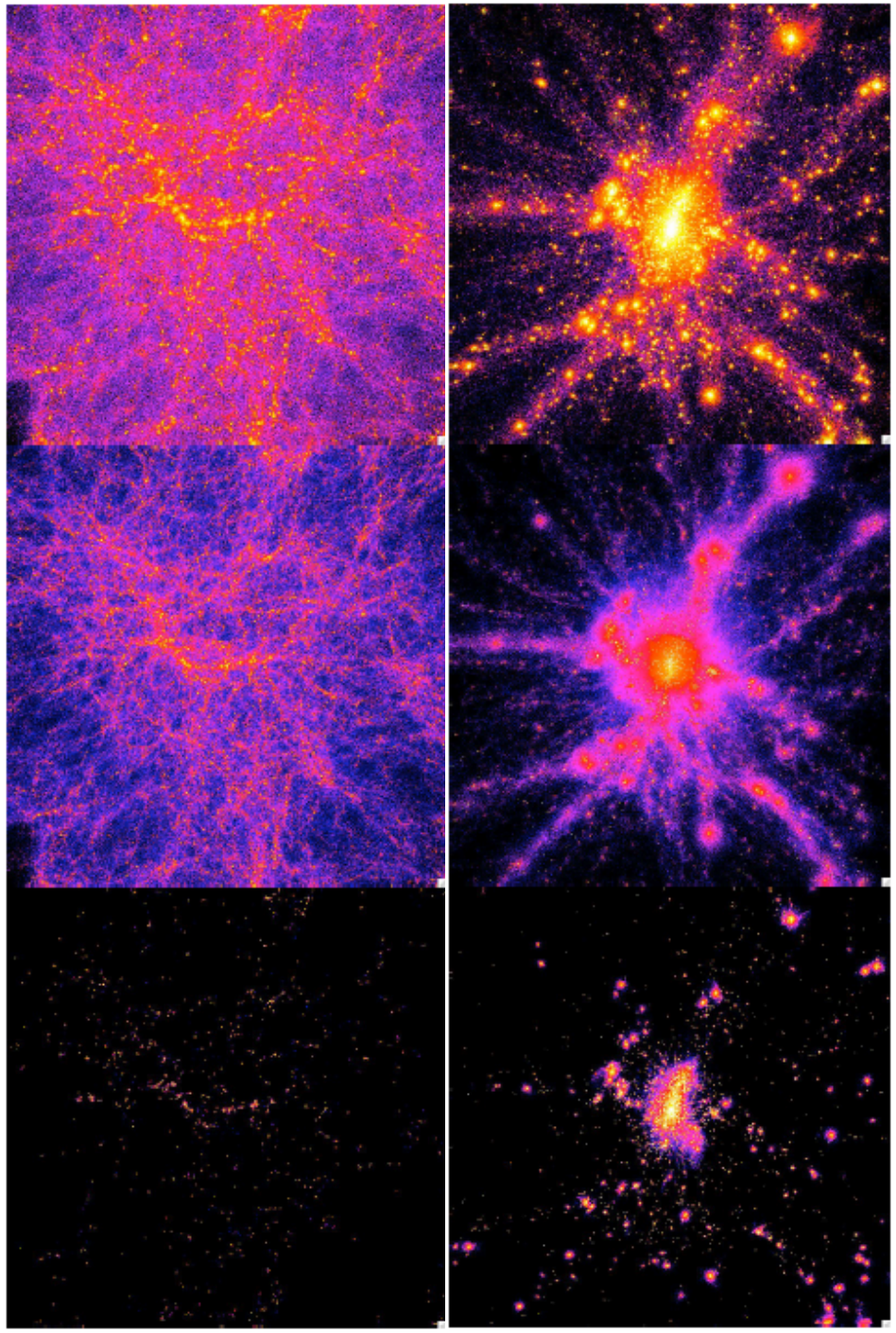


Athena



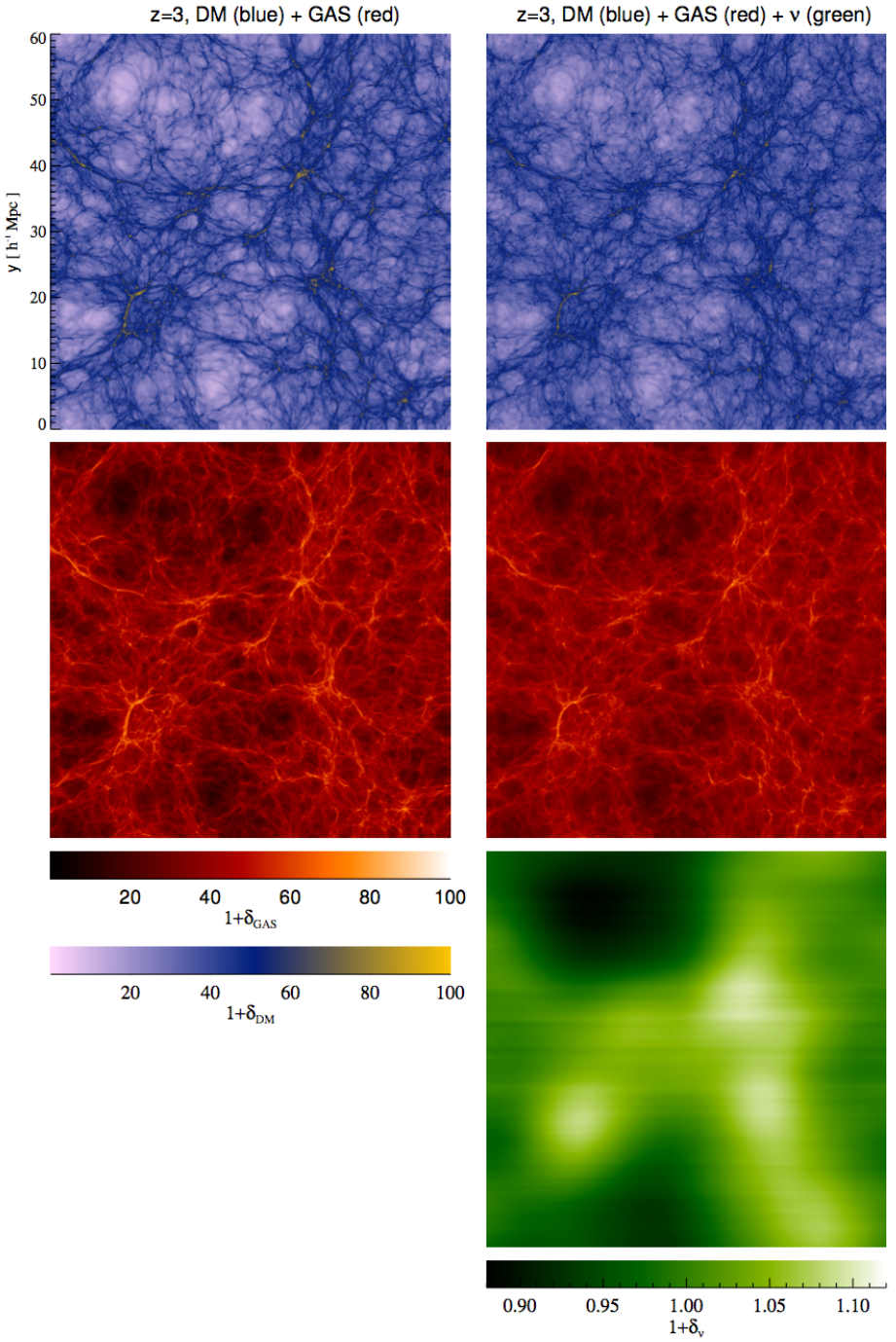
# What is the role of Numerical Astrophysics?

- to capture the complexity of the formation of cosmic structures
- as interpretative framework for the “tsunami” of observational data
- to optimize the design and operation of such large facilities
- to prepare methods and tools of analysis



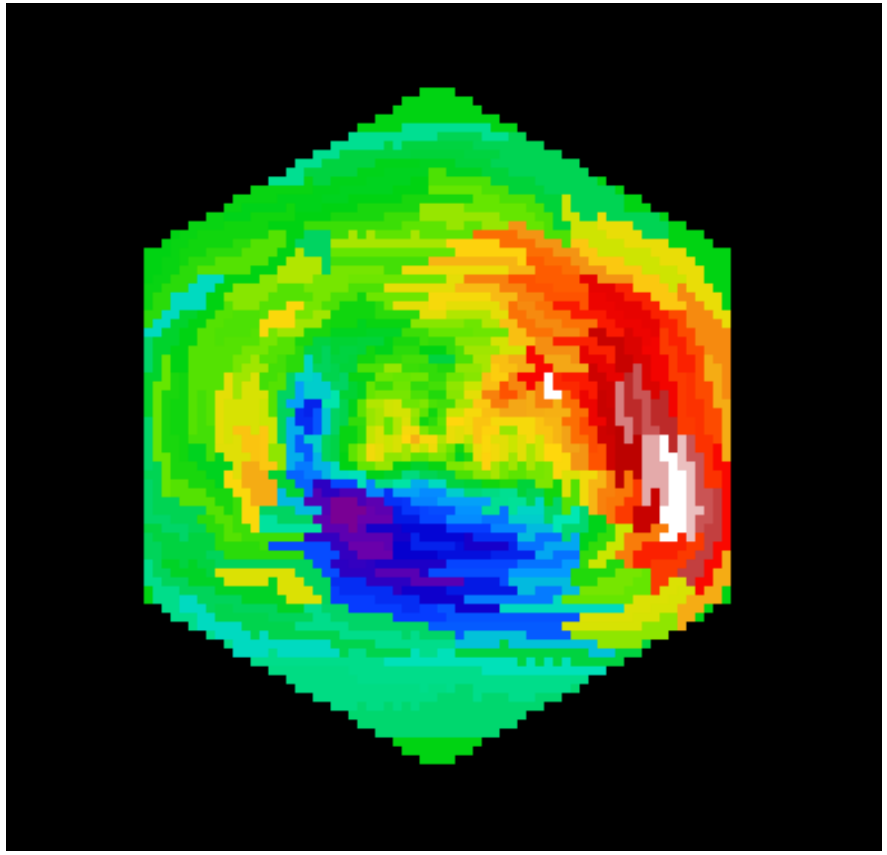
# What is the role of Numerical Astrophysics?

- to capture the complexity of the formation of cosmic structures
- as interpretative framework for the “tsunami” of observational data
- to optimize the design and operation of such large facilities
- to prepare methods and tools of analysis



# What is the role of Numerical Astrophysics?

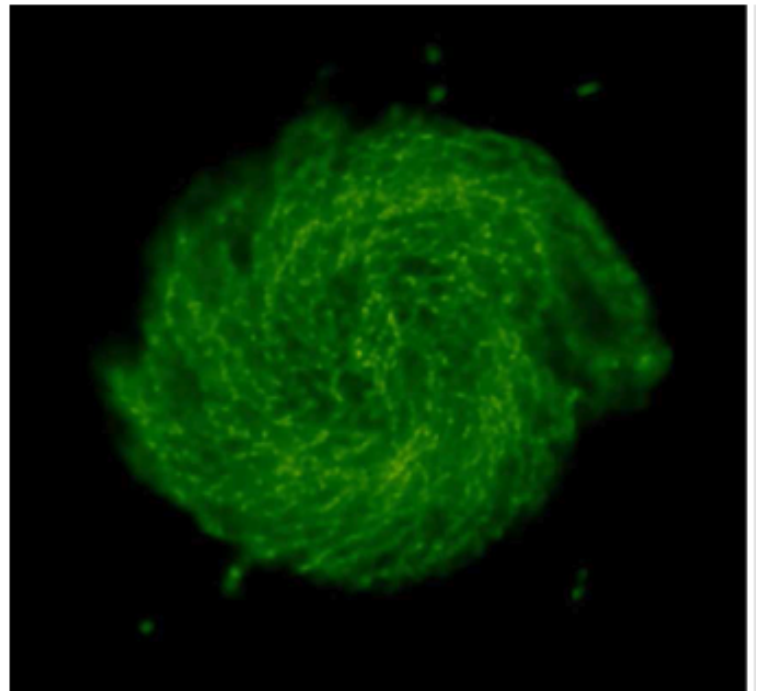
- to capture the complexity of the formation of cosmic structures
- as interpretative framework for the “tsunami” of observational data
- to optimize the design and operation of such large facilities
- to prepare methods and tools of analysis

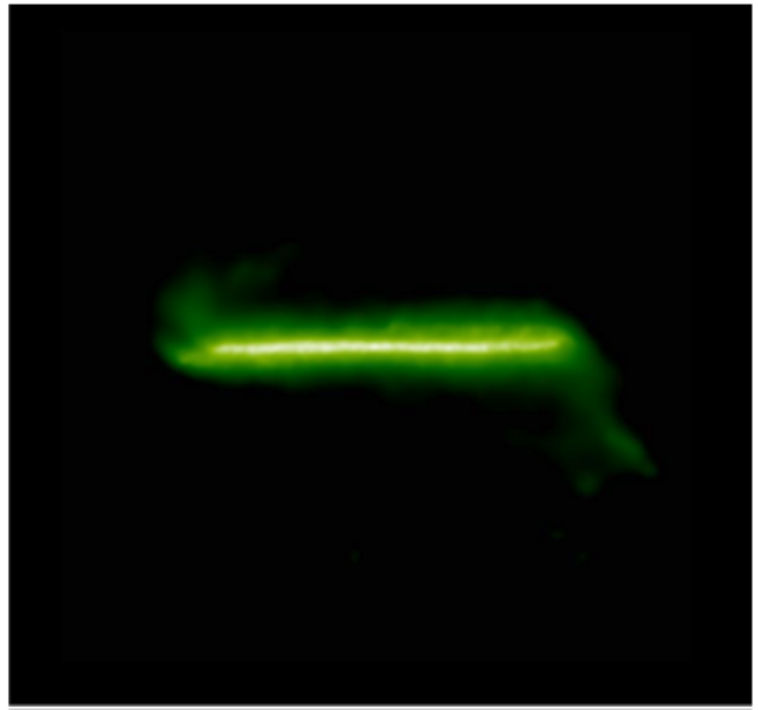


ICM bulk velocities of a simulated Perseus-like clusters with a 100 ksec exposure with Athena

# What is the role of Numerical Astrophysics?

- to capture the complexity of the formation of cosmic structures
- as interpretative framework for the “tsunami” of observational data
- to optimize the design and operation of such large facilities
- to prepare methods and tools of analysis





# Facilities are not only ”telescopes”

New exascale capable laboratories allow to increase dramatically the dynamical range (from cosmological scales down into galaxies)

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

New experiments require new exascale capable laboratories!

*Are Astrophysical (HPC) codes ready for that?*

# What is exascale?

**$10^{18}$ floating-point ops / sec**

10 to 100 times faster than today's fastest machines
(Human Brain estimated to α ~$10^{15}$ops/sec)

more than just a peak rate of sustained arithmetic ops
→ 1000-fold better "capability" than that at peta-scale

Such a system cannot be produced today.
Its realization is challenging and requires significant advances in a variety of technologies. It is uncertain whether exascale is achievable without disruptive changes in the way we build and use computers and write scientific applications.

# Exascale challenges

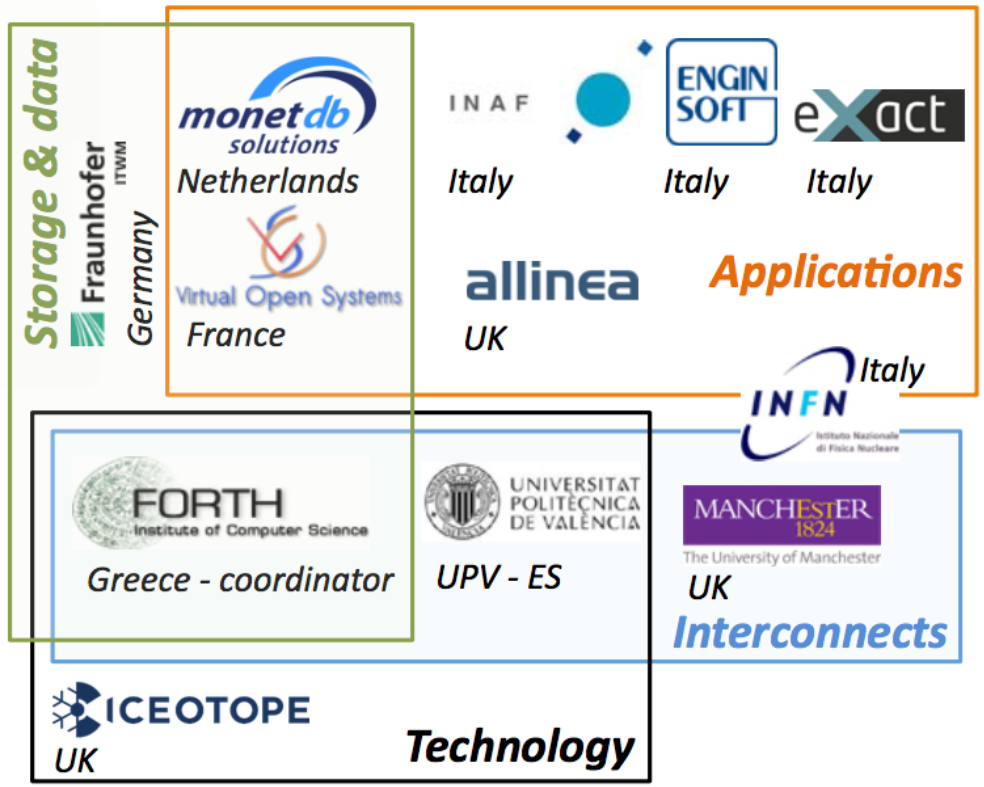
- Interconnect technology
- Memory technology
- Scalable system software
- Programming systems
- Data management (Filesystems)
- Algorithms
- Resilience

Energy Efficiency !!!!!

*The goal (constraint) is to sustain an exascale machine  with less then **50MW** of power.*

# EU Exascale projects: the “Exa” Projects

- European Exascale System Interconnect  & Storage
- EU Funded project H2020-FETHPC-1-2014
- Overall budget about 8 MEuro
- 12 Partners (6 industrial partners)




ExaNeSt

# Ecosystem for Exascale

- EuroServer: Green Computing Node for European microservers
  - UNIMEM address space model among ARM compute nodes
  - Storage and I/O shared among multiple compute nodes
- ExaNoDe: European Exascale processor-memory Node Design
  - ARM-based Chiplets on silicon Interposer
- ECOSCALE: Energy-efficient Heterogeneous Computing at exaSCALE
  - Heterogeneous infrastructure (ARM + FPGAs), programming, runEmes
- Kaleao: Energy-efficient μServers for Scalable Cloud Datacenters
  - Startup company, interested to commercialize many of the results

ExaNeSt

# Moving to exascale

“Free lunch” picture

cluster more CPUs

✚

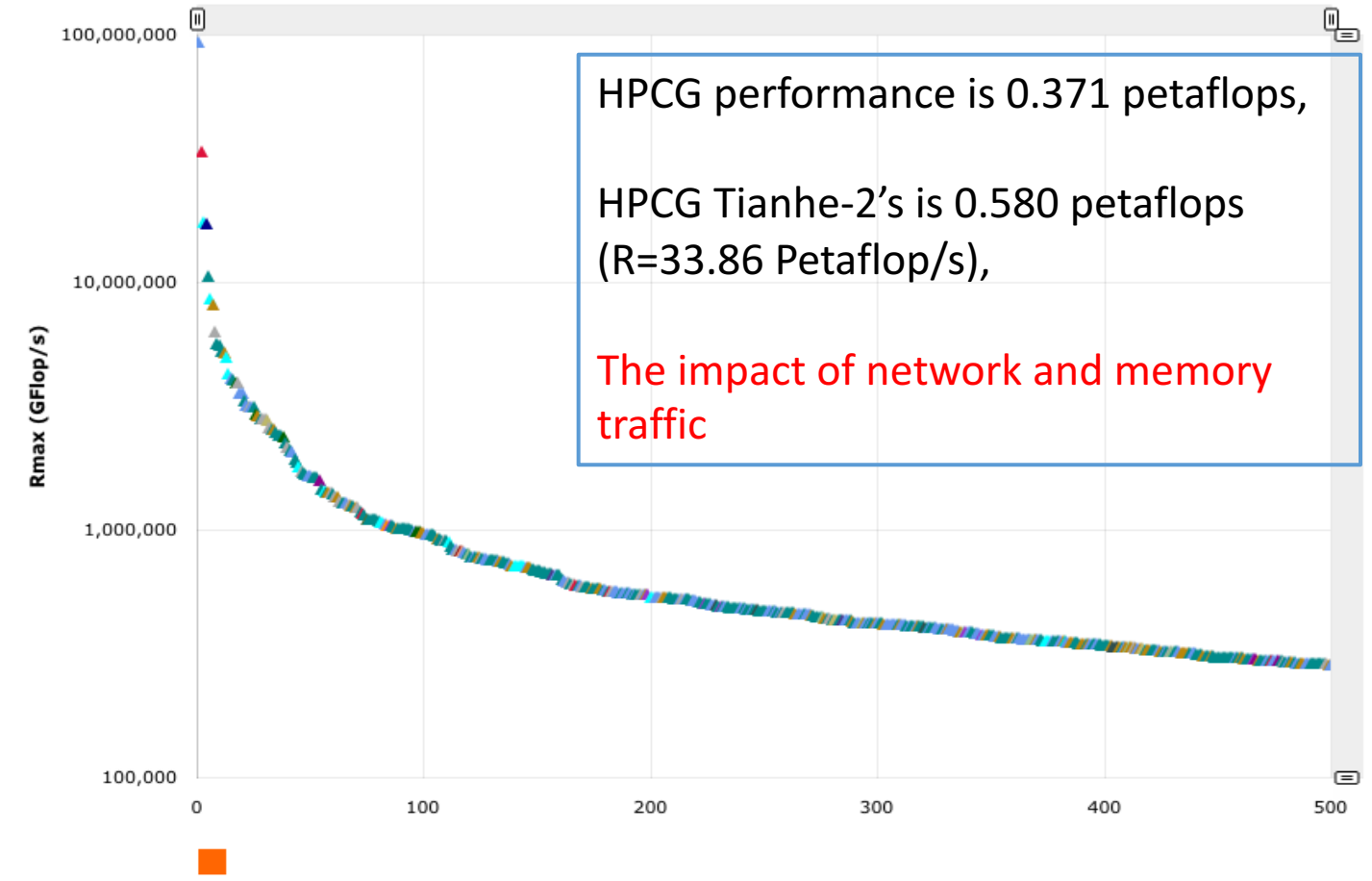
some amazingly fast inter-connection

✚

a ton of memory

# Super computing facilities today

- Sunway (MPP) is 2.75 times as powerful as the former Tianhe-2
- It has 10.6 million cores, with a theor. peak perf. of **125 Pflops**.
- Running LINPACK at **93 Pflops** means it performed at 74% of its theoretical peak



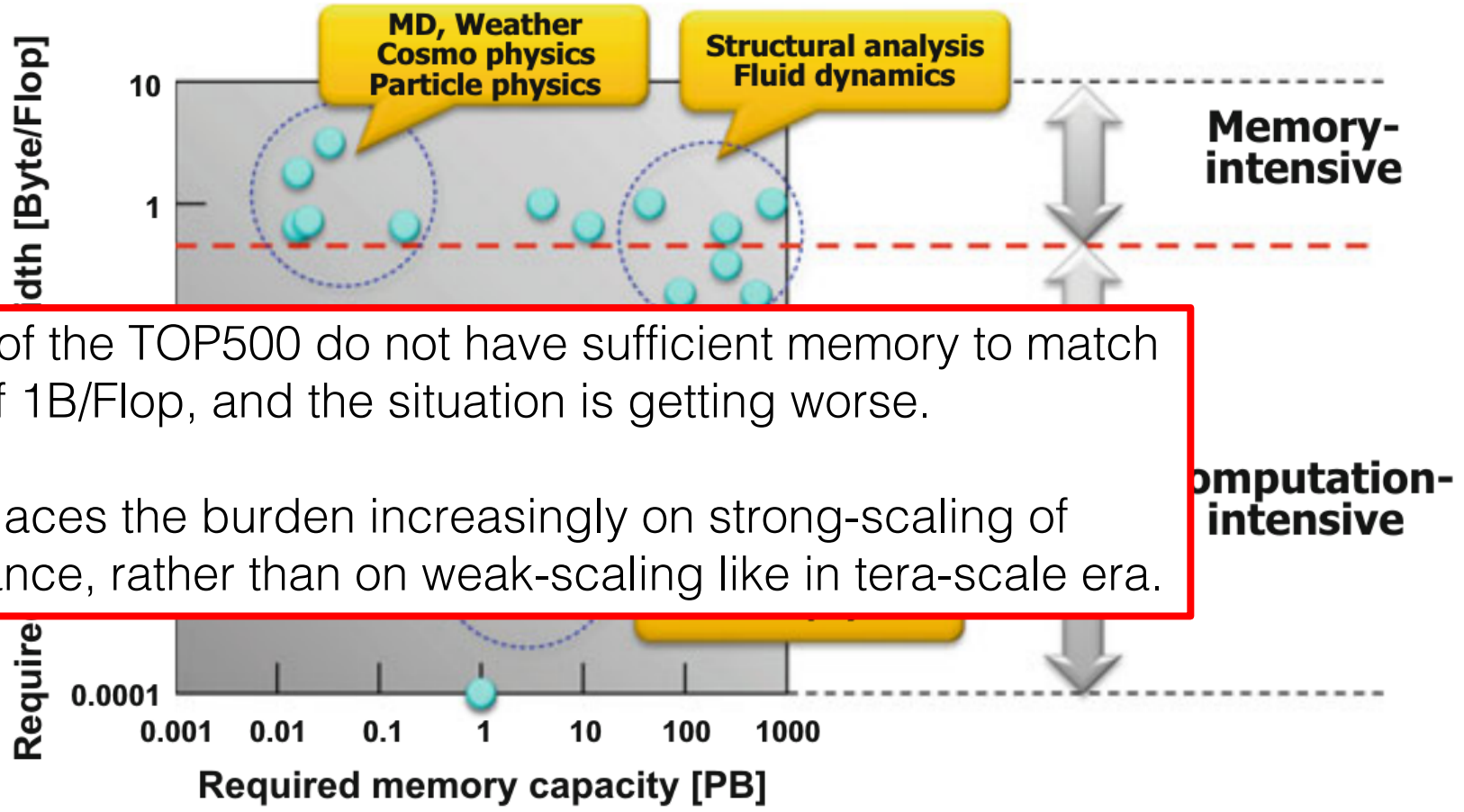

Legend:

Intel Xeon E5 (Broadwell), Power BQC, Xeon E7 (Westmere-EX), SPARC64 VIIIfx, Intel Xeon E7 (IvyBridge), Intel Xeon E5 (Haswell), Xeon 5500-series (Nehalem-EP), Xeon 5600-series (Westmere-EP), Intel Xeon E5 (IvyBridge), SPARC64 XIfx, Sunway, Intel Xeon E5 (SandyBridge), Xeon 5500-series (Nehalem-EX), Opteron 6300 Series "Abu Dhabi", Opteron 6100-series "Magny-Cours" , POWER7, Intel Xeon Phi, ShenWei, SPARC64 IXfx, Opteron 6200 Series "Interlagos", Intel Xeon E7 (Haswell-Ex),

# Sustained vs Peak performance

## Algorithms : memory intensive vs computation-intensive

We can engineer far more floating point capability onto a chip than can reasonably be used by an application. Today.



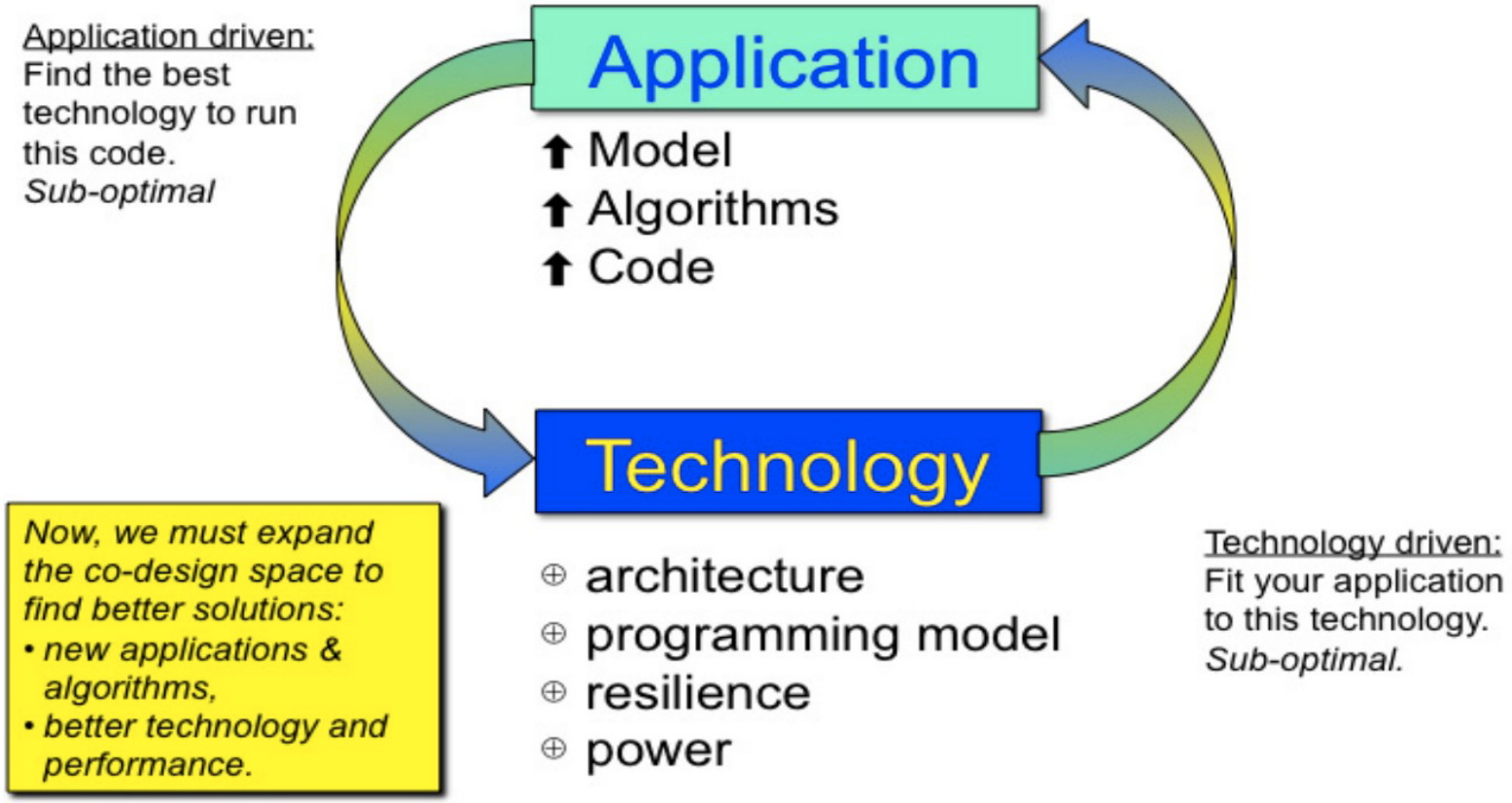

The machines at the top of the TOP500 do not have sufficient memory to match historical requirements of 1B/Flop, and the situation is getting worse.

This is a big change: it places the burden increasingly on strong-scaling of applications for performance, rather than on weak-scaling like in tera-scale era.

Credits: Hiroaki Kobayashi 2014

# Hardware Software co-design

Applications define the requirements for the system (network, IO, QoS, interconnect, resilience, and more)

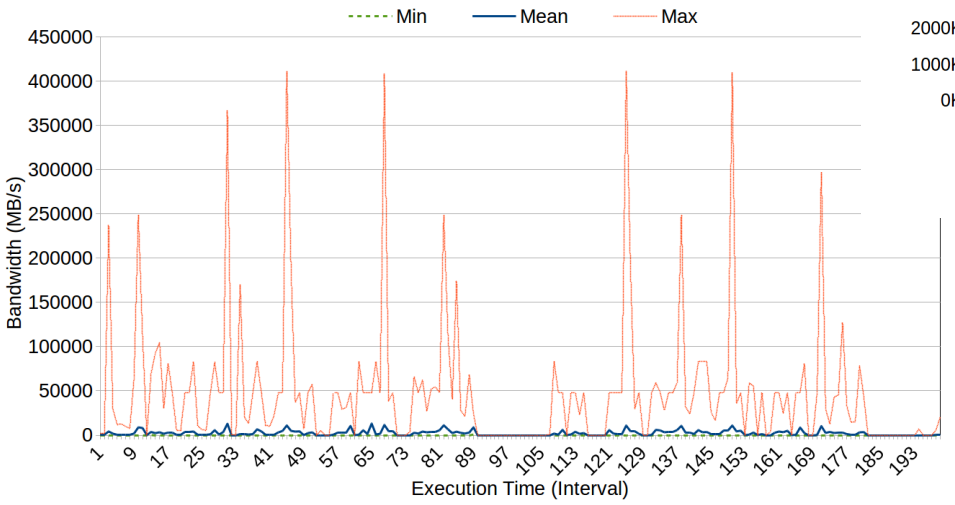



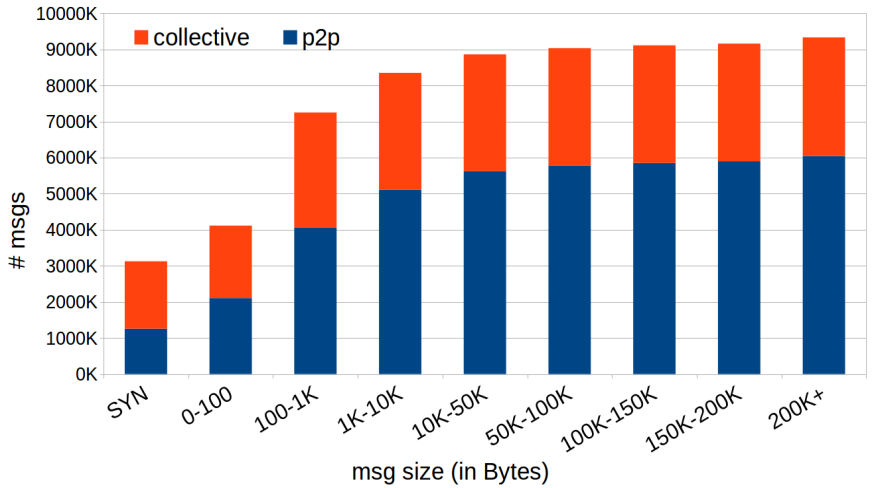
# Applications domains

- Cosmological n-Body and hydrodynamic code(s) suited to perform large-scale, high-resolution numerical simulations of cosmic structures formation and evolution (INAF).
- Brain Simulation.  Generate spiking behaviors and synaptic connectivity that do not change when the number of hardware processing nodes is varied (INFN)
- Weather and climate simulation (ExactLab)
- Material science simulations (ExactLab and EngineSoft)
- Workloads for database management on the platform and initial assessment against competing approaches in the market (MonetDB)
- Virtualization Systems (Virtual Open systems)

# Co-Design examples

## Network patterns (GADGET)




## Memory intensive vs computation-intensive estimate on PINOCCHIO code

1. the simple way → using profiler tools to account for a “global” B/F
2. the hard way → instrumenting the code using the PAPI library to account for precise B/F

→ initial condition generation: B/F ~ 0.01
→ first and second derivatives (using FFTWs): B/F ~ 0.05

# ExaNeSt approach

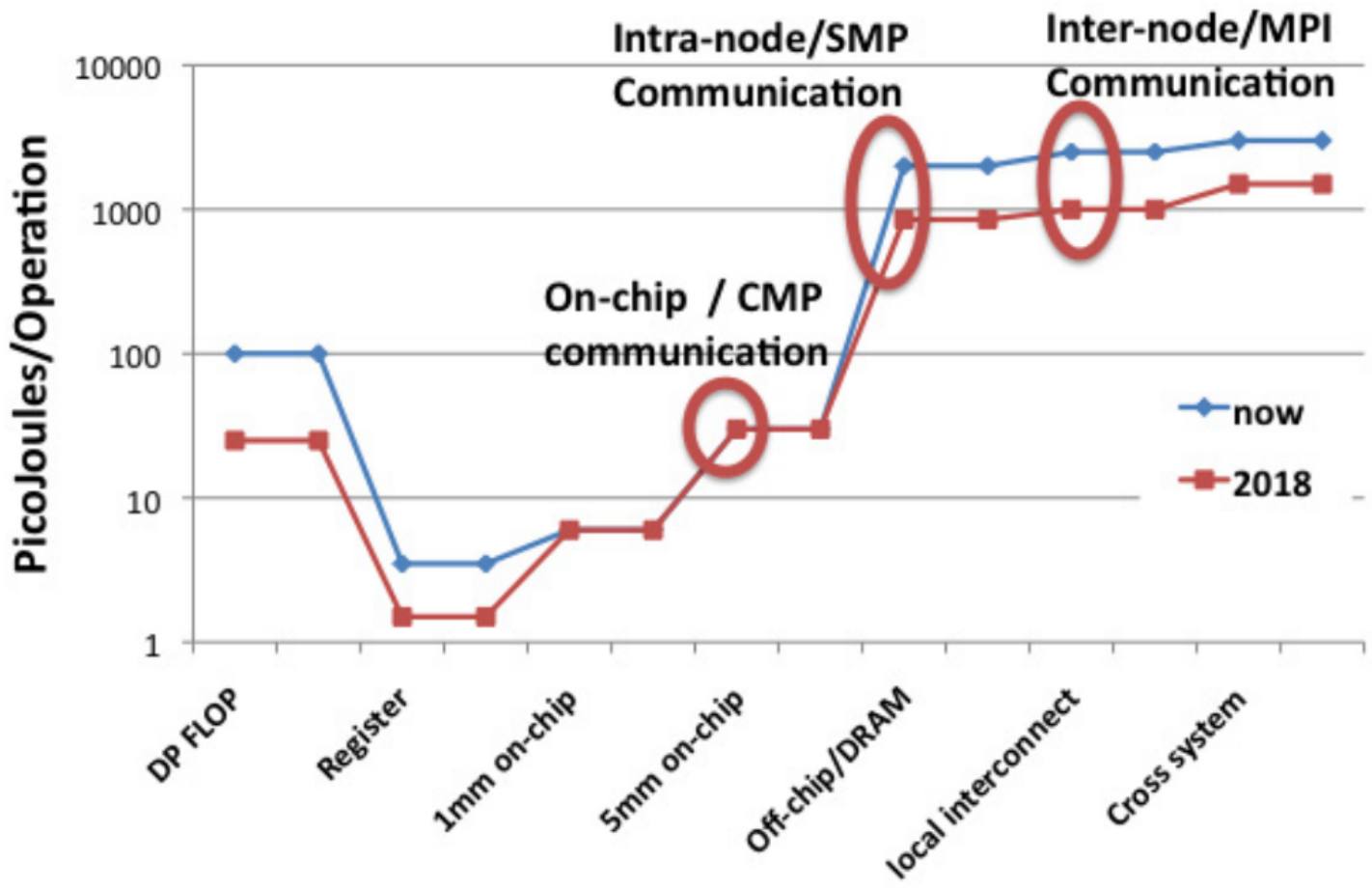
"Data movement presents the most daunting engineering and computer architecture challenge"

- Multi-Tiered scalable interconnect for Unified approach: merge inter-processor traffic with major storage traffic (photonic technologies)
- Packaging and network topology analyzed together.
- Support Quality of Service
  - Isolate flows with different requirements (low latency, high throughput)
- Support for queue, flow-control, congestion control, scheduling, monitoring
- Improve data locality.

# “Make” computing close to data

The closer it is, the less it “costs” (in terms of latency and power).

“Memory” Capacity is critical to applications.
It allows a powerful form of weak scaling, in-memory checkpoints and message logging/replay for resilience; it enables algorithms that buy performance by using data structures that may not be minimal in their “memory” footprint, improve data locality.



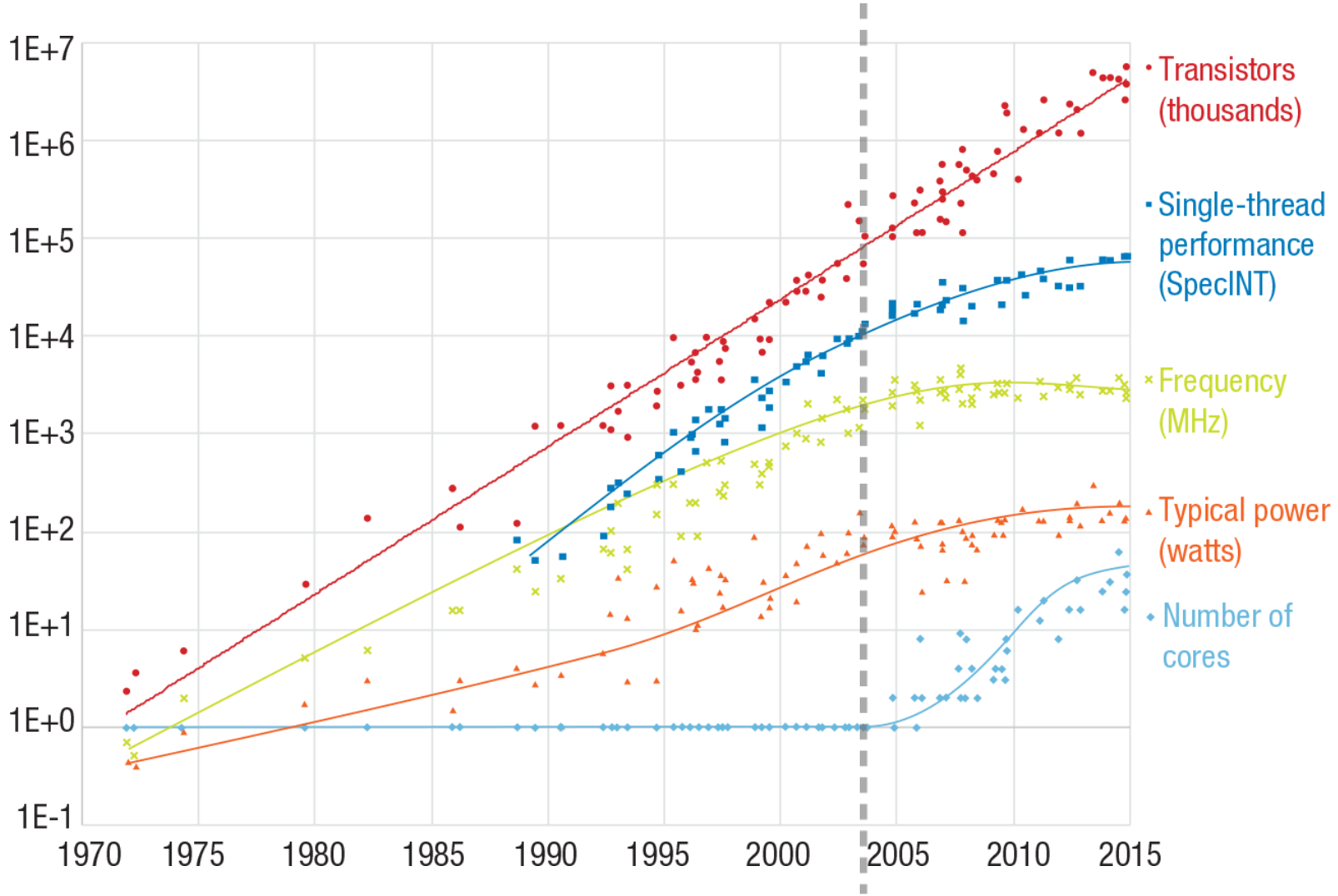
# Computing power

Moore's law enforced not by "more powerful" single-core CPU but by multi/many-cores CPUs + additional technologies:

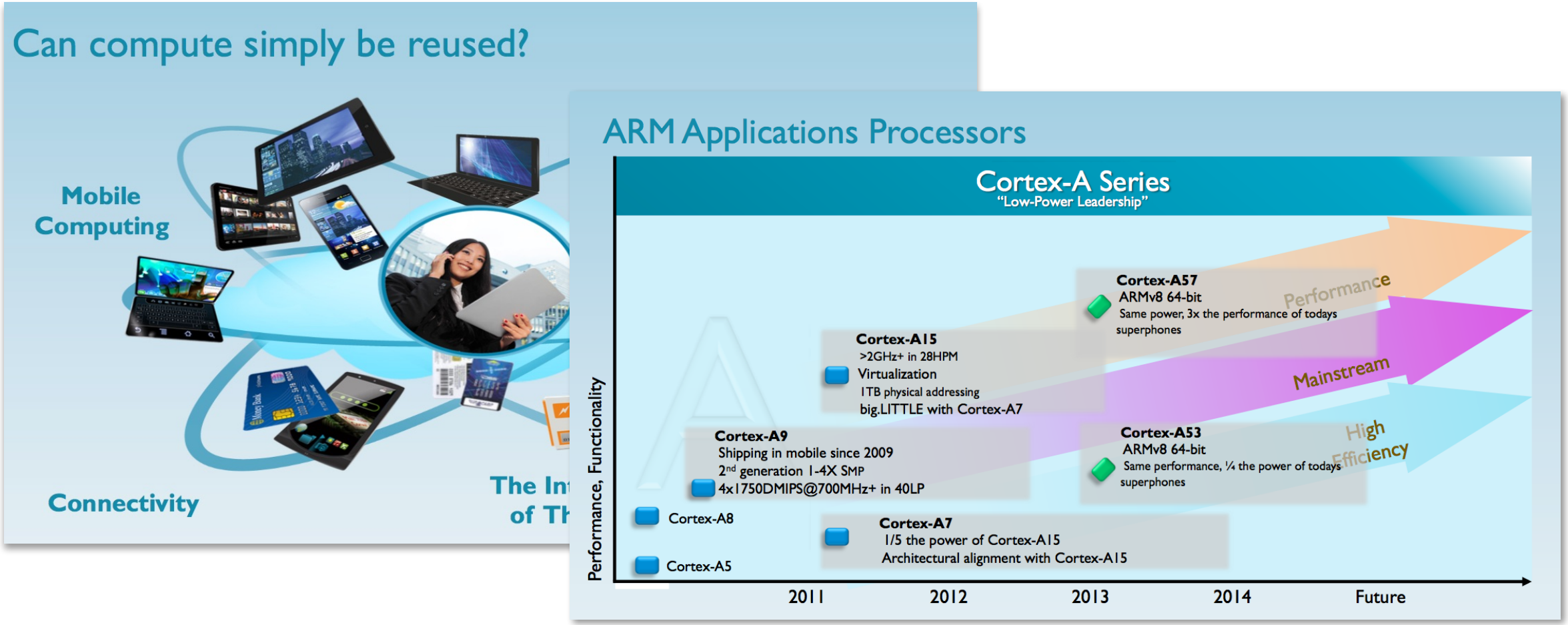
- Threading
- pipelining
- data parallelism (SIMD instructions)
- GPUs, PCIe, …

Many-cores CPUs are here to stay and their number of cores will increase: thousand CPUs/Core per chip.

Shall we expect "simpler" cores (do you remember RISC…)?



# ExaNeST approach: “Let’s ride the tiger”




Credits: John Goodacre (ARM) 2014

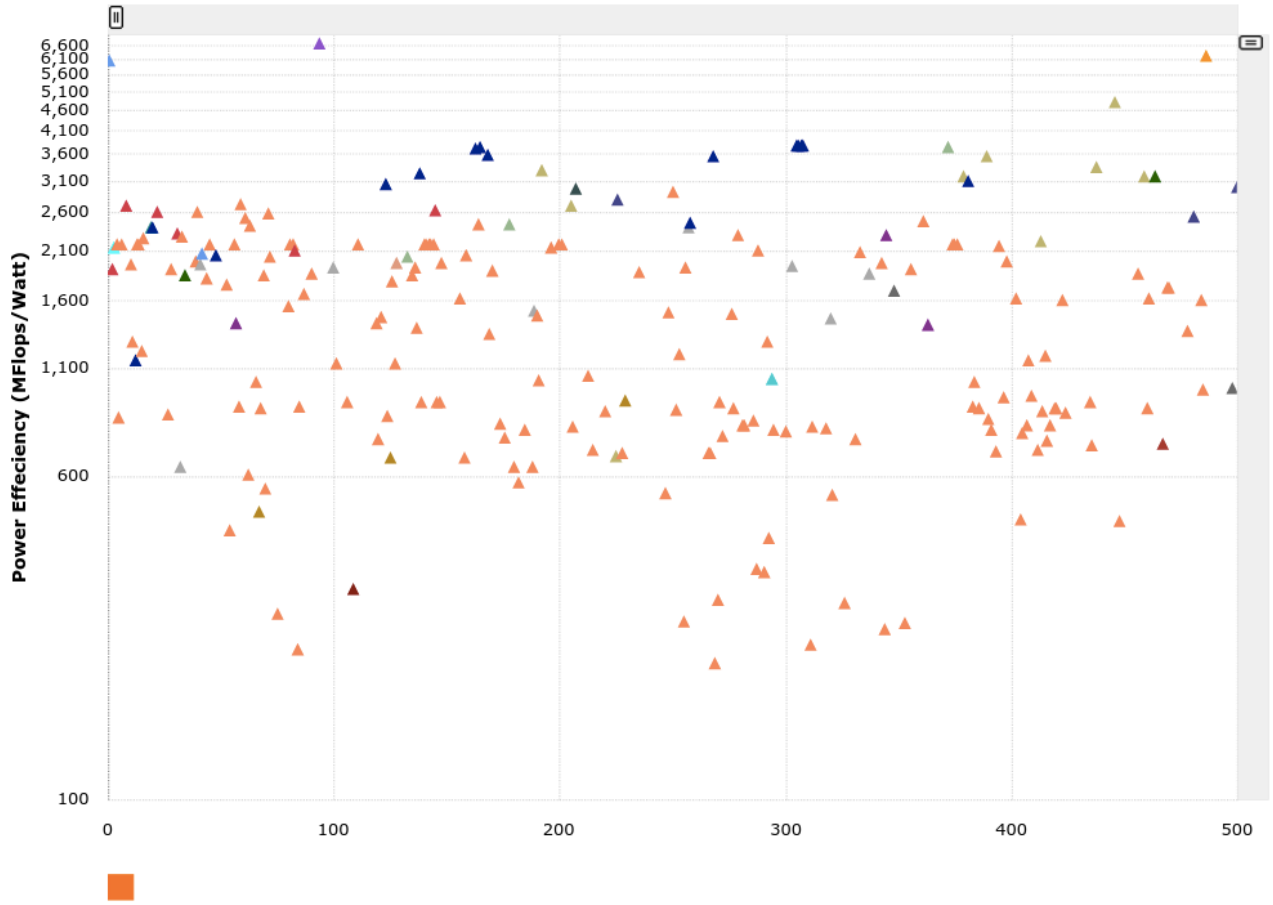
# Energy consumption

The Chinese Sunway is consuming ~**18MW** (RISC processors).

However, even re-scaling it to Eflops it would reach **~1.8GW** which is clearly prohibitive.

The exa-scale goal is to reach 1Eflops at less than 50MW of electric power,
i.e. about 50Gflops/W

Computing, data movement (memory and interconnect), storage, cooling contribute to energy consumption




Legend:

Intel Xeon Phi SE10P, Intel Xeon Phi 5120D, Intel Xeon Phi 31S1P, Intel Xeon Phi SE10X, Intel Xeon Phi 7120P, NVIDIA Tesla K40m, Intel Xeon Phi 7110P, NVIDIA Tesla K20, None, NVIDIA K20/K20x, Xeon Phi 5110P, NVIDIA 2090, NVIDIA Tesla K40, AMD FirePro S9150, NVIDIA 2050, AMD FirePro S10000, NVIDIA Tesla K20m, Nvidia Titan Black, NVIDIA 2075, ATI HD 5870, PEZY-SCnp, NVIDIA 2070, NVIDIA Tesla K20x, NVIDIA Tesla K40/Intel Xeon Phi 7120P, NVIDIA Tesla K80, Intel Xeon Phi 7110, Xeon Phi 5120D/Nvidia K40, Intel Xeon Phi 5110P,

# Memory (not so) hidden energy costs

| *Operation* | *pJoules* |
|---|---|
| 64bits FP 28nm CMOS | 12 |
| 32bits integer operations on 28nm CMOS | 3 |
| 64bits FP single-issue in-order core | 200 |
| 64bits multiple-issue out-of-order core | 1000 |
| reading 32bits instruction from 32KB cache | 20 |
| reading 64bits operands from DRAM | 2000 |

Today's (extreme) Samsung's GDDR5: 4.3 W / 64 Gbs

To feed a modest 0.2 B/flop for a sustained $10^{18}$ flops the requirement is about 12MW of power only for memory.

# ExaNeSt approach

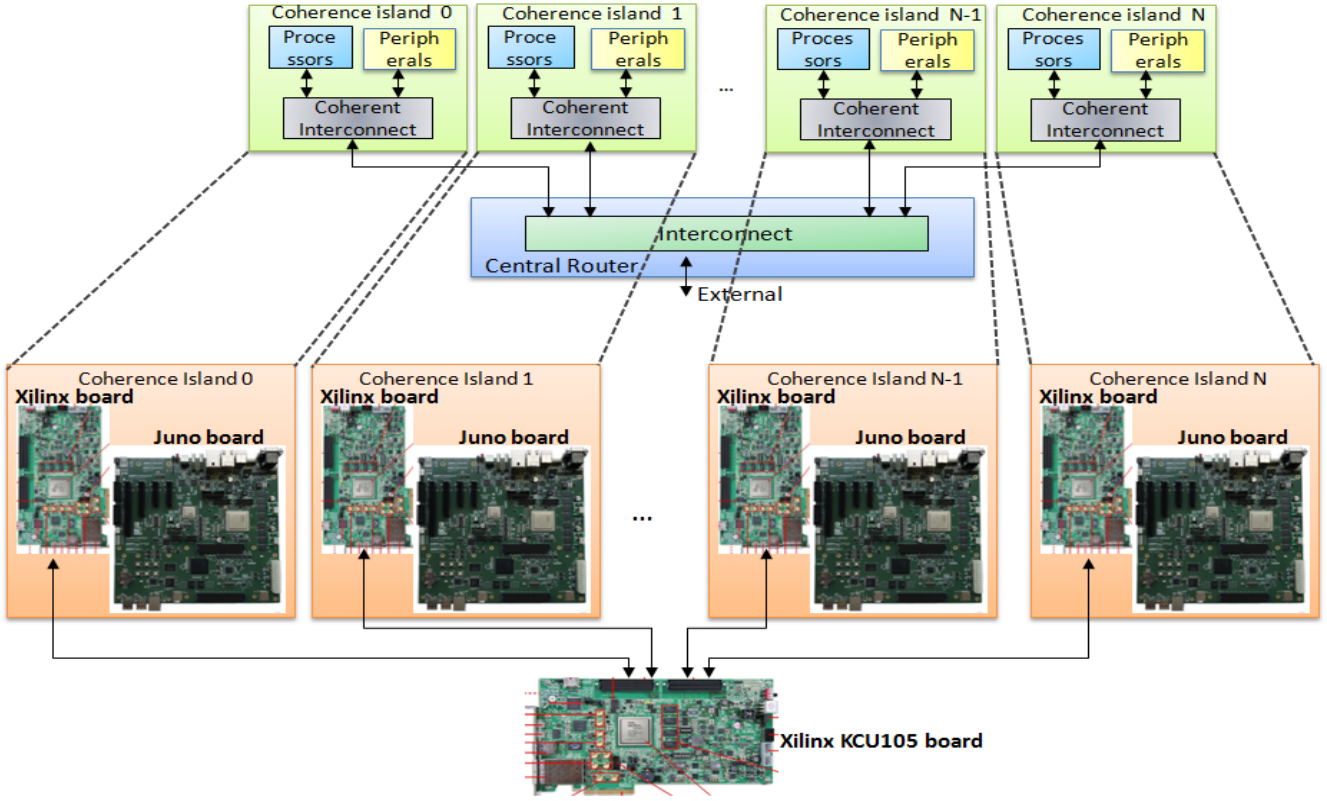
- Realistic rack-level shared-memory based on UNIMEM (shared memory at RACK level)
- Use of NVM on RAM sockets (keep storage close to computing)
- Accelerators that share RAM with CPUs (FPGA accelerators through OpenCL kernels)

# Unimem technology

- Single owner per page: every memory page in at most one node's cache
    - *no system-level hardware coherence traffic*
    - owner can be any node - not just the (local) one adjacent to DRAM
- Remote memory accesses, remote mailbox, remote interrupts:
    - for fast synchronization & processing of distributed (read-only) data
- Remote-page borrowing for memory disaggregation
- Zero-copy remote direct memory transfer (RDMA)
    - sockets over zero-copy RDMA
    - MPI over sockets

# ExaNeSt hardware

- Quad FPGA Board (QFDB) is under development
- 4 ARMv8 cores @1.5GHz per FPGA + Accelerators 64GB RAM per Board
- One NVM per Board for TIER 0
- Intra-QFDB low latency interconnect 230Gb/s High speed serial links for TIER0 (APEnet)
- Inter-board HSS @ 160Gb/s TIER1/2 (APEnet)
- Intra-RACK based on Photonic technology (160 Gb/s)
- FPGA:  more computing capabilities (2500 DSP @ 300MHz)
- Mali-400MP2 GPU (low computing power) 600Mhz




Double sided 8U blade
384 Cores + 96 FPGA + 96 GPUs for 1U cabinet
Photonic Interconnect
BeeGFS multi-tiered Parallel FS

# Resilience


How many hardware failures each day?  Try to guess....

- Impact on interconnect design (routing and scalability)
- Impact on Filesystem (availability and scalability)
- Impact on System Software (availability and scalability)
- Impact HW design (packaging and topology)
- Virtualization for/in HPC

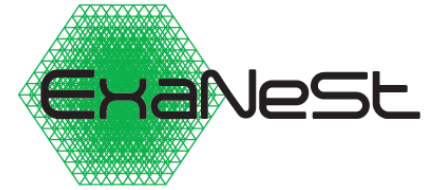
- Impact on Applications: is check-pointing mechanism (terascale era approach) a valuable methodology?
  - Not all A&A applications are using CP
  - CP relies on FS (do you remember the data movement problem?) and its availability
  - System Software must be aware of CP and restart  (queue systems)

# System software

"System software must play a more active role in making decisions about how resources are allocated and managed, and the strategies for managing these resources can be very different for different applications."

- System software is dealing with billion-fold concurrency and associated locality.
- System software will have more responsibilities to deliver performance for applications.
- Lightweight specialized Operating Systems to "simplify" access to memory, network and processors.
- New programming models, runtime environment and computational libraries

# Blueprint for exascale applications




- New algorithms must take into account communication/synchronization - avoiding algorithms that increase the computation/communication ratio (Flops per communicated Bytes
- algorithms that implement a law B/F ration (<0.1)
- algorithms that support simultaneous computation/communication,
- algorithms that vectorize well and have a large volume of functional parallelism.
- algorithms that adaptively respond do load imbalance of billion-threads scale (e.g. dynamic scheduling by DAG) without compromising with spatial locality

# ExaNeSt applications

- Algorithms that use more flops and less bytes → increasing bytes but decreasing flops (task based codes).
- We consider complex and deep memory hierarchies, the heterogeneity of memory latencies, and the efficient use of logical units attached to memory modules.

- Algorithms with more sophisticated scheduling and memory management than heretofore seen. Data-flow-like models, where parallelism is expressed explicitly in DAGs, allow for the scheduling of tasks dynamically, support of massive parallelism and application of common optimization techniques to increase throughput.

- Algorithms that adapt to possibly heterogeneous environments and/or resources, and that are fault-oblivious and error-tolerant.

# Astrophysical codes

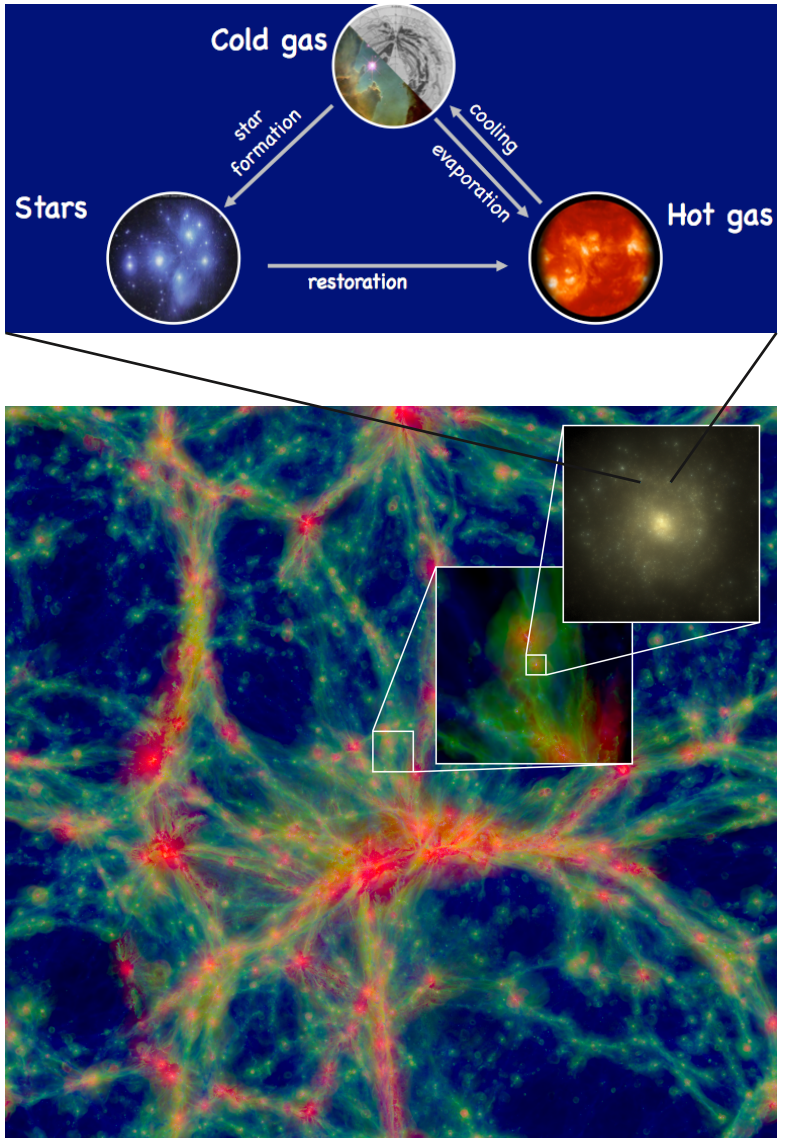
- N-Body Simulations
  - GADGET (http://wwwmpa.mpa-garching.mpg.de/gadget/)
  - SWIFT (http://icc.dur.ac.uk/swift/)
  - ChaNGa (http://www-hpcc.astro.washington.edu/tools/changa.html)
- Semi-analytical model
  - PINOCCHIO (arXiv:1605.04788)

- More applications are going to come: data reduction and analysis (SKA precursors, GAIA)

# Numerical simulations

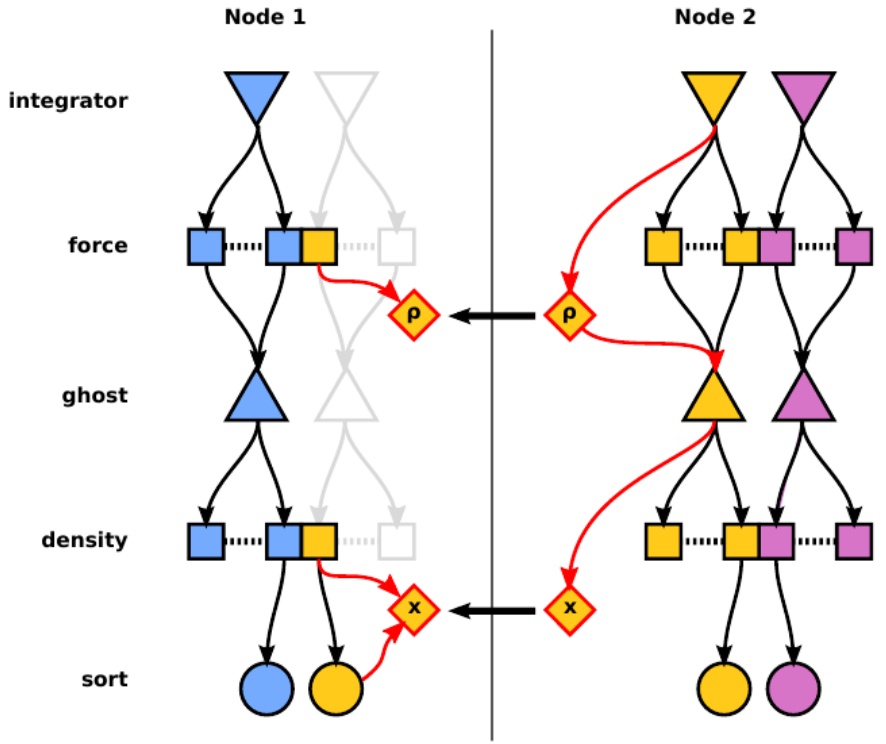
- Gravity: long range interaction, no screening
- Large (spatial and temporal) dynamic ranges:
  - From ~100 Mpc of cosmological environment to sub-pc scale, relevant for astrophysical processes: > 8 decades
- Resolve down to ~100 pc scales and describe the rest through sub-resolution models
- Cross-talk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

- Designed in the “terascale” Era, ….



# GADGET code

- PRACE Unified European Applications Benchmark Suite
- old-fashioned tera-scale parallelism model: low (with respect to million/billion – scale) number of MPI threads execute the same work-flow on a fraction of the system – potential OpenMP threads execute concurrently the same task until all the threads come to end.
- Computation is done as a “monolithic” workflow instead of being split-up in smallest autonomous “tasks”.
- Based on ”blocking” MPI (v2.0) communications and frequent all-to-all communication / synchronization cycles.
- Moving towards a task based approach with an efficient “exascalable” application scheduler.




**SWIFT** tasks approach

# Resilience approach for applications

- Old style (but always useful) Check-pointing


- Hybrid approach: Virtualization and HPC.
  - Schedulers on VMs that migrates in case of errors
  - Schedulers are directing task and they are aware of system faults, so task can ”migrate” on other nodes

- “lockstep” approach
  - More realization of the same simulation running on the same system.

# Conclusions

- Exascale supercomputers require a great investment in terms of research activities: today it does not exist a naïve approach to exascale.
- Exascale is not for everyone.
- Computing at “exascale” level will be possible for applications with low B/F. What will it happens to Big Data computing?
- Applications should think in term of a new paradigm: Task based approach, resilience, data locality, heterogeneous computing (CPUs/GPUs/HW accelerators).
- Deadline: 2020 (hopefully)
- This is a new revolution in the way we are coding: we are not familiar with multi-core and accelerators…

# ExaNeSt”ers” @ INAF

- Stefano Borgani
- Luca Tornatore
- Giuseppe Murante
- David Goz
- Valentina D’Odorico
- Gianluigi Granato

- Credits: S. Borgani, M.Viel, L.Tornatore, G. Murante