Gruppo di Cosmologia numerica di Trieste

Supercomputing: the challenge of the incoming architectures

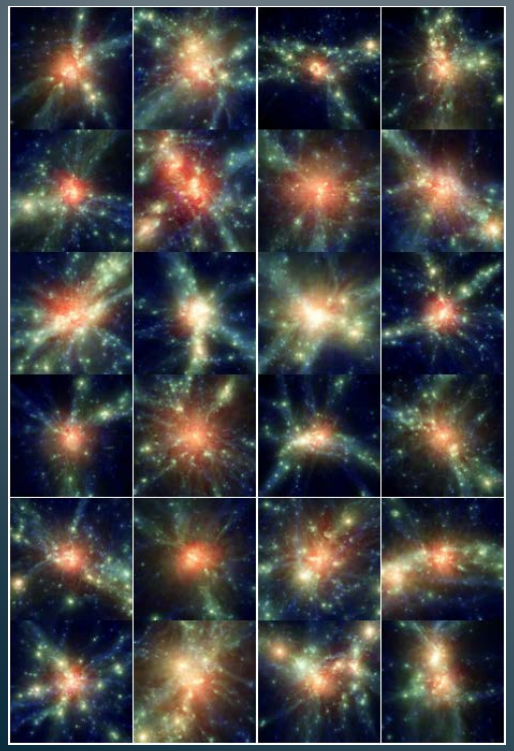
2015

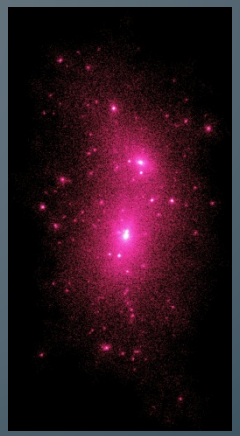
(P. Barai)
V. Biffi
S. Borgani
G. De Lucia
(P. Di Cerbo)
(D. Goz)
D. Fabjan
G. Granato
P. Monaco
C. Mongardi
(E. Munari)
G. Murante
(S. Planelles)
E. Rasia
M. Valentini
M. Viel
[G. Ucci]
F. Villaescusa-Navarro

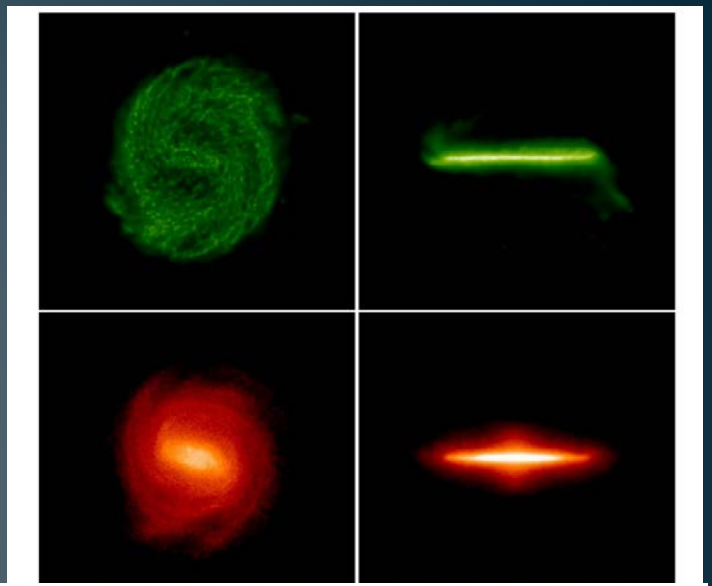
A. Beck (Munich)
K. Dolag (Munich)
M. Gaspari (Bologna)
C. Ragone-Figueroa (Cordoba)
L. Steinborn (Munich)
(L. Tornatore)

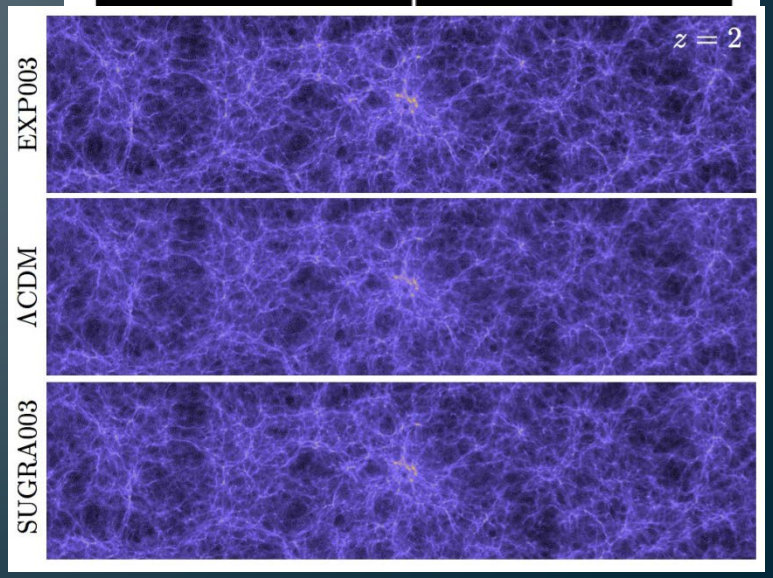
Formazione di strutture cosmiche

- Galassie
- Ammassi di galassie
- Struttura a grande scala

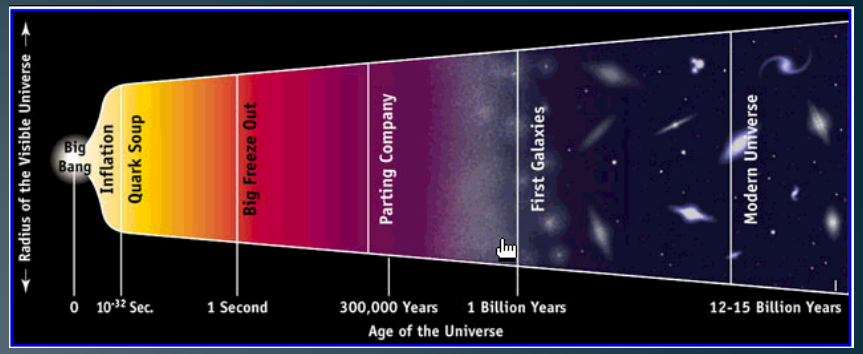








La fisica



- Condizioni iniziali cosmologiche
- Gravita'
- Idrodinamica

- Processi astrofisici sottogriglia:
  - Cooling
  - Star Formation & Feedback
  - UV heating
  - Evoluzione chimica
  - AGN feedback

  - Formazione idrogeno molecolare
  - Trasferimento radiativo
  - ….

CODICE LAGRANGIANO (Gadget-3)

PM+TREECODE

SPH

Risultati piu' recenti: ammassi di galassie

- Modern SPH + nuovo modello AGN
- Riprodotta la dicotomia CC/NCC in simulazioni cosmologiche di singolo oggetto
- Buon accordo con altre proprieta' osservate dell'ICM (metallicita'…)
- Simulazioni di dimensione medio/piccola

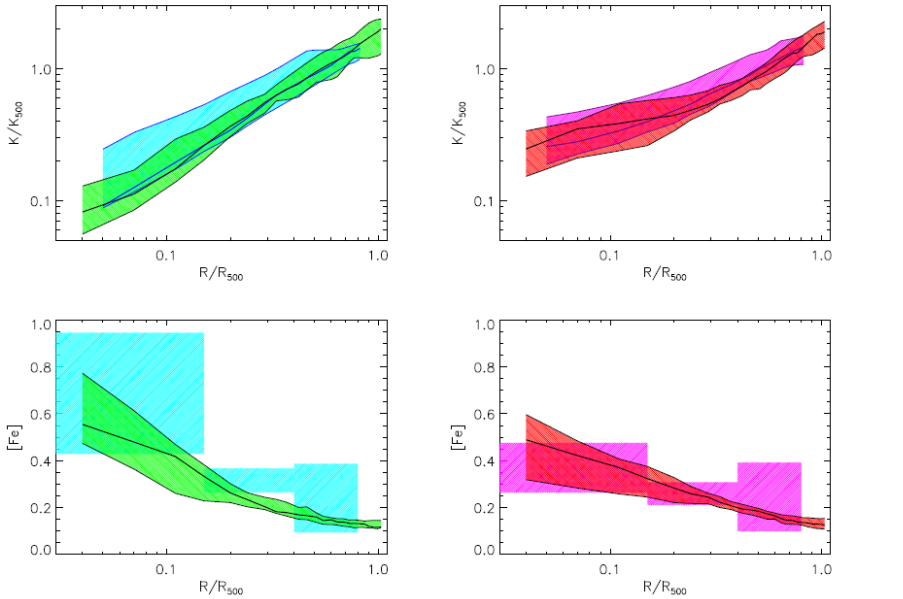


FIG. 1.— Simulated $z = 0$ entropy (top panels) and Iron abundance (bottom panels) profiles compared with observations by Pratt et al. (2010) and by Ettori et al. (2015), respectively. The medians of the simulated CC (left panels) and NCC (right panels) profiles are in black. The shaded regions (in green and red, respectively) delimit the 16^{th} and the 84^{th} percentiles. The observed profiles for CC and NCC are, respectively, in cyan and magenta. Metallicity profiles are expressed in units of the solar abundance as reported by Anders & Grevesse (1989).

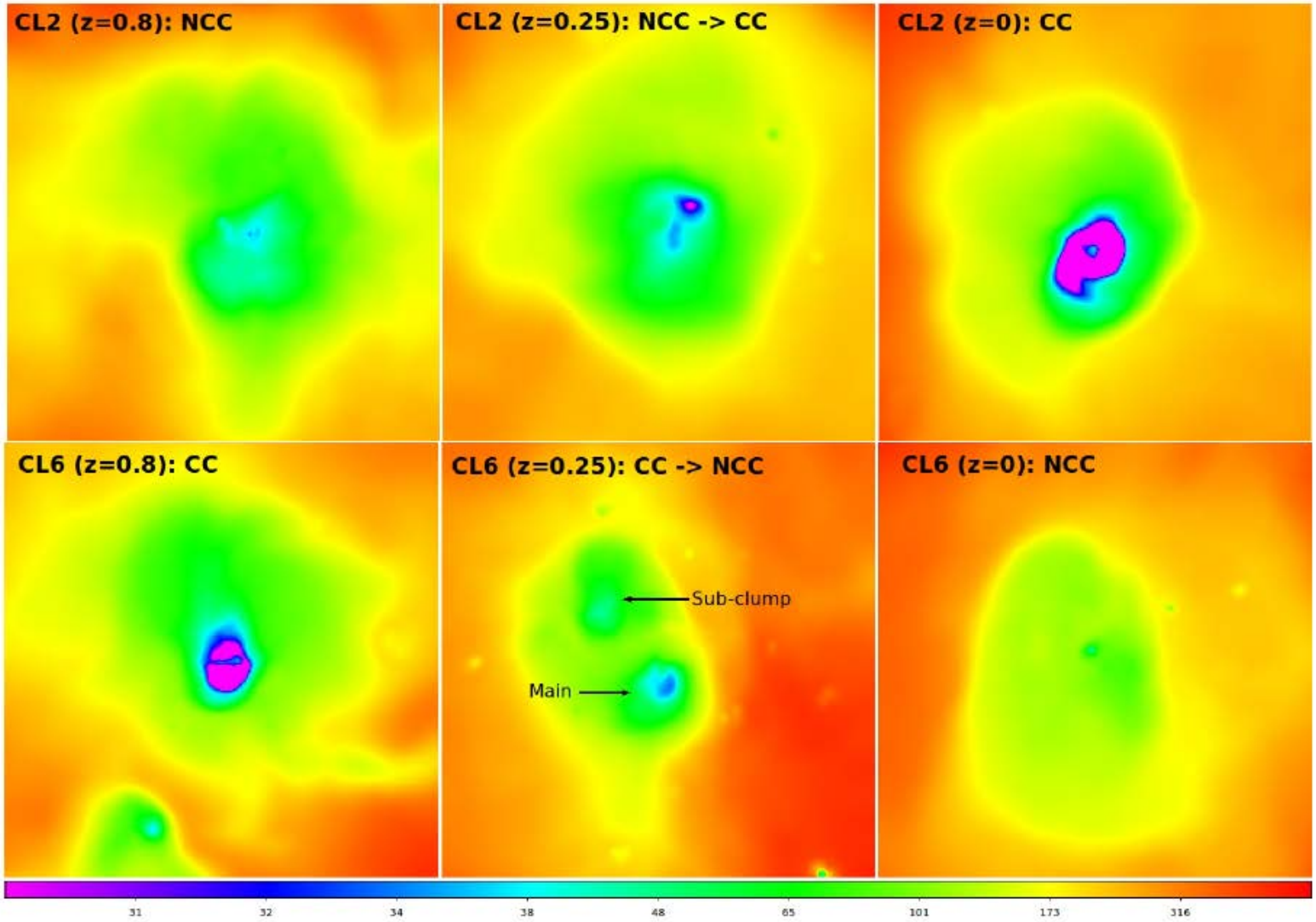
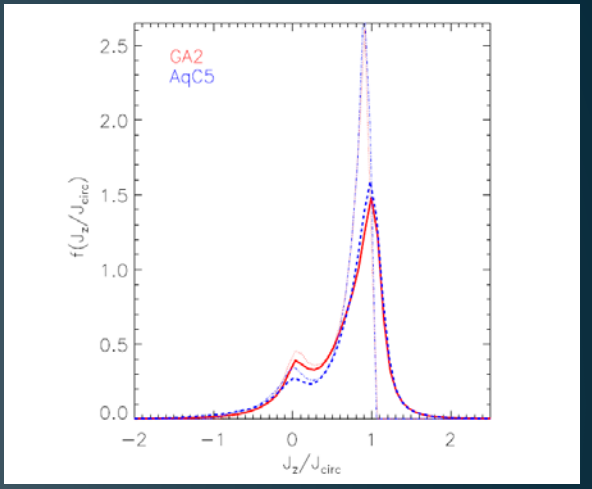
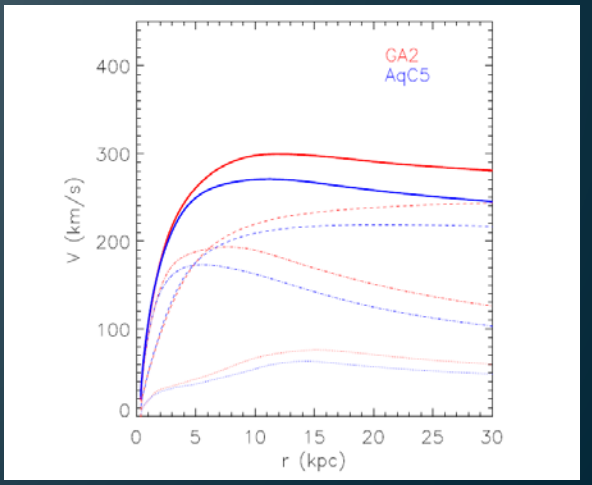


FIG. 2.— Maps of pseudo entropy of two simulated clusters that have masses $M_{500} = 2.4 \times 10^{14} h^{-1} M_{\odot}$(upper panels) and $M_{500} = 7.3 \times 10^{14} h^{-1} M_{\odot}$(lower panels) at $z = 0$. The size of the images is 1Mpc.

Risultati piu' recenti: simulazioni di galassie a disco





Progetti a breve/medio termine (simulazioni)

- Simulazioni di ammassi di galassie ad alta ed altissima risoluzione (Iscra B, DECI)
- Simulazioni cosmologiche di galaxy formation (Iscra B, DECI)
- …CLUES? (appena finito run in bassa risoluzione ☺)

- Simulazione di un ammasso di galassie a risoluzione ancora maggiore, (MUPPI+NewAGN+modern SPH). Prace Class A
- Simulazioni cosmologiche di galaxy formation, ad alto redshift ed altissima risoluzione. Prace Class A

Progetti di sviluppo codice (TS)

- Unione moduli MUPPI/AGN
- Implementazione di vari tipi di FB cinetico da SN
- Implementazione formazione H_2 in MUPPI
- Traccianti passivi
- Feedback cinetico da AGN

- GSPH (?)
- SPH++

- Riscrittura paralelizzazione in paradigma task-based

Struttura di GADGET

Infrastruttura:
parallelizzazione
load balancing
domain decomposition
treebuild/treewalk
acceleratori

0.2 staff
[2 ass. Ricerca?]

Gravita' (PM+Treecode)

nessuno

Idrodinamica (SPH)

0.3 staff

Fisica aggiuntiva (cooling, star formation+FB, AGN, evoluzione chimica…)

2.5 staff
1(+1) postdoc
2 [+1] PhD
(1 laureando)

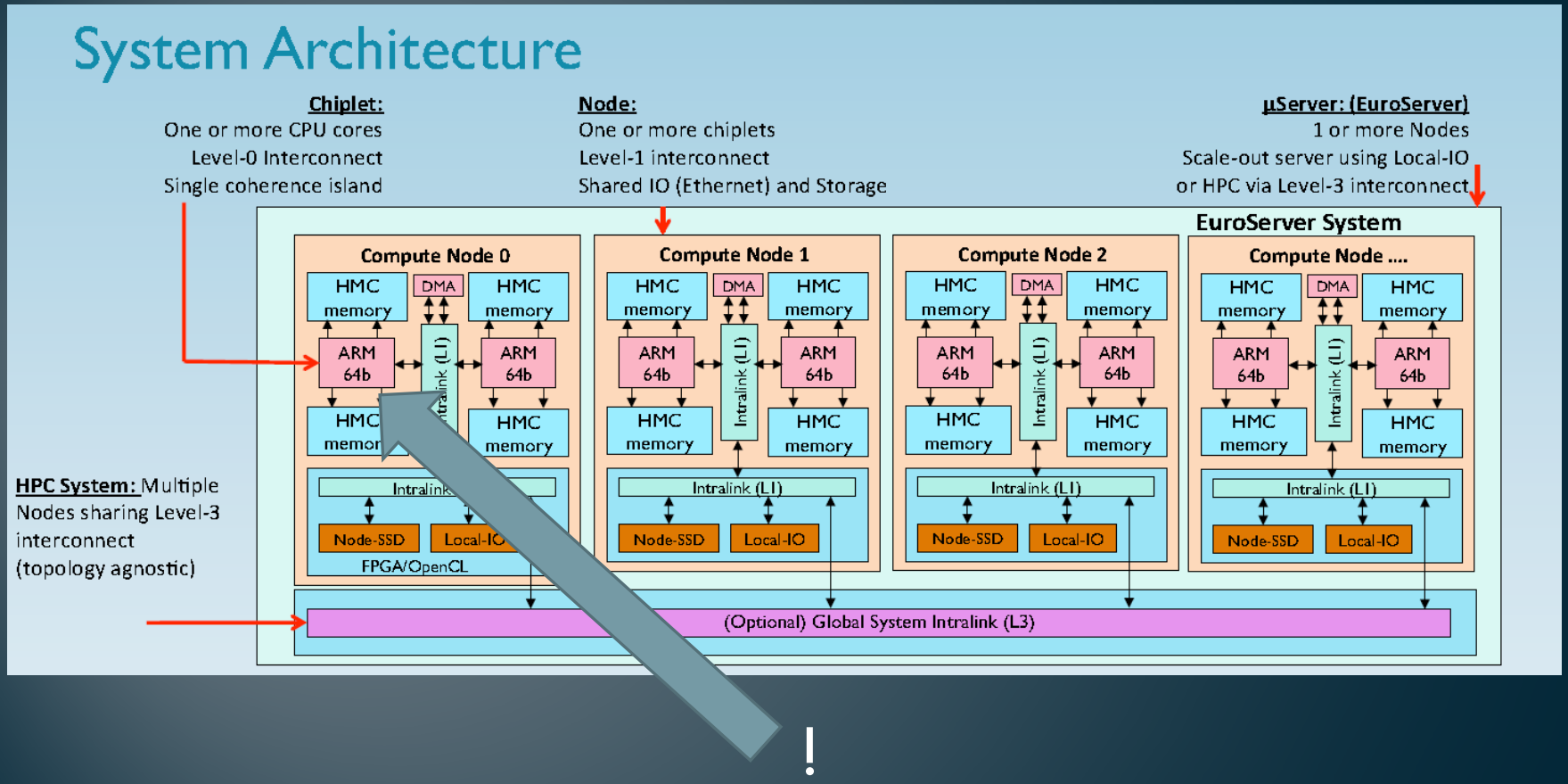
Tentativo di porting su GPU

- EuroHack 2015, Lugano
- Team: Monaco (K. Dolag, M. Petkova, A. Ragagnin) + Trieste (D. Goz, M. Valentini). Supporto esperti INTEL in OpenACC

- Tentativo di portare treewalk ed sph
- Grossi problemi coi compilatori!!
- Treewalk: total failure
- SPH: porting parziale, ma si e' ottenuto uno **slowdown**

Conclusione: serve personale dedicato, occorre riscrittura completa di parti di codice. L'operazione non si fa in una settimana od un mese.

ExaNest



(Presentazione di Giuliano Taffoni)

Gerarchia del sistema Exaflop

Hierarchy	Scale	Performance	DRAM	Storage	Maximum Power
Chiplet (Compute Unit)	Heterogeneous CPU/GPU compute unit	8 CPU 200 GFLOPS	Up to 6x 8GB	virtualized	15 W (16 GB)
Interposer (3D-IC)	4 × Chiplet	32 CPU 800 GFLOPS	64 GB	virtualized	70 W
Compute Node (Shared IO & Acceleration)	2 × Interposer, I/O + OpenCL FPGA	64 CPU 3.5 TFLOPS	128 GB	Host SSD 400-3400 GB	140 W + 20 W for I/O
Compute Element (daughter board PCB)	2 × Nodes	128 CPU 7 TFLOPS	256 GB	6.8 TB	320 W
Mezzanine (mother-board for Elements)	4 × Elements	512 CPU 28 TFLOPS	1 TB	27 TB	1.28 kW + 120 W Interconnect
Blade (deployment unit / hot-swap)	3 × Mezzanine	1536 CPU 84 TFLOPS	3 TB	81 TB	4.2 kW + 0.8 kW cooling
Rack (metal frame)	72 × Blades	110,592 CPU 6 PFLOPS	221 TB	5.8 PB	360 kW + 1 kW TOR switch
Example HPC System	100 × Racks	11 M CPU 600 PFLOPS	22 PB	58 PB	36 MW
ExaScale Level	167 × Racks	1 ExaFLOPS 18.5 M CPU	37 PB	1 ExaByte	60 MW

Livello di parallelismo richiesto!!

Codici e macchine exaflop

- Assolutamente necessario personale specializzato, che porti su queste architetture i codici scientifici
- La figura del tecnologo INAF **DEVE** essere concepita anche come *software engineer*, a supporto dei gruppi di ricerca **TEORICI**
- Difficolta' di reperimento personale di questo tipo (pochi soldi; non facciamo ricerca di punta in informatica; richiediamo competenze medio-alte)

- Lo scienziato sviluppatore deve occuparsi principalmente dei moduli scientifici (e gia' non sara' facile)

Quale codice?

- GADGET3: privato, non task-based, non modulare, 10 anni di lavoro sui moduli di fisica
- CHANGA: Charm++, open source, molti moduli di fisica, necessario porting dei nostri e della nuova implementazione SPH, stabilita'?
- SWIFT: Task-based, open source, moduli di fisica da portare, modern SPH da implementare, performance?
- Problema COSMO-Ts: …altri gruppi?

Nota.

- Siamo entrati nel progetto europeo non perche' contattti come INAF da industrie ma grazie a rapporti costruiti da persone OATs/ eurotech . Via la persona, via il rapporto
- INFN e' contattata come partner
- Se resta fine a se stessa, la nostra partecipazione al progetto e' un'occasione persa

L'incubo e' gia' realta'...

- Simulazioni cosmologiche (leader: Monaco) con range dinamico relativamente limitato scalano fino a 20k cores
  - Buona threadizzazione OpenMP, load balancing facile
- Simulazioni di singolo oggetto (leader: Ts) hanno problemi oltre 1-2k cores
  - OpenMP funziona male, load balancing difficile ed affamato di RAM

- Nessuna simulazione usa acceleratori…..

Tempo di calcolo

- LARGE projects: Prace Class A (Italia: Fermi; si chiede estero)
  - Tier0 (O(10^5) cores)
  - Grants non banali da ottenere
  - Code solitamente lente (Fermi: code ferme)
  - Porting (decente) solitamente non banale
- MEDIUM projects: Iscra, DECI (Italia: Galileo)
  - Tier1 (O(10^4) cores)
  - Code solitamente lente
  - DECI puo' avere problemi di supporto
- SMALL projects: ..... (Ts: convenzione CINECA, galileo)
  - Tier2 (che non abbiamo) (O(10^3) cores)
  - **Necessario per poter partecipare a calls per medium e large**
- DEVELOPMENT: ..... (Ts: server locale, ma: no accel)
  - Tier "2.5" (che non abbiamo); necessario a tutto quanto sopra

Postprocessing e data preservation

CINECA: PICO

PROS:

- Flessibile
- Molto storage
- Risorse immediatamente disponibili
- Sostituisce bene un server di gruppo dedicato ed e' piu' potente
- Facile da accedere per molti utenti

CONS:

- non veramente adatto ad uso risorse da parte di un gruppo di utenti
- EOI (UNICO ACCESSO POSSIBILE ATTUALMENTE) SCADE IN UN ANNO. ***"DATA PRESERVATION"????????????***
- RICHIEDERE UN NUOVO PROGETTO: FARE ISCRA C PER **GALILEO**……

Visualizzazione ed animazioni

- Utile per la scienza in se
- Sempre di effetto in presentazioni scientifiche
- NECESSARIA per l'outreach
- NECESSARIA per il "marketing" (promozione dell'Astrofisica)

- Sopra: movie mio, evolutivo
- Sotto: tirocinio studente polacco, sei settimane, ray tracing

- Anche qui gradito aiuto professionale, anche interno…

Conclusioni

Viviamo in tempi interessanti.

(antica maledizione cinese)

- SE INAF vuole fare calcolo, allora c'e' bisogno di:
  - INFRASTRUTTURA. Centro di calcolo con Tier2 (…in house??)
  - ORGANIZZAZIONE. Per esempio, prevedere investimenti per personale dedicato, NON solo scientifico
  - "LOBBYING" per migliorare il contesto nazionale
- **Attenzione** che calcolo non e' solo HPC, ma anche HTC, data processing….. Dovremmo agire tutti insieme
- se NON vuole.. ce lo dica. La discussione sulla necessita' del calcolo e' comunque stata affrontata ad astrofrontierie.
- Non si puo' dire di voler fare calcolo e non fornire quanto sopra. E' come dire: l'astronomia ottica e' una priorita' ma non vogliamo i telescopi.