THE EINSTEIN TOOLKIT ON SUMA SYSTEMS

Michele Brambilla (INFN Milano Bicocca & Parma Univ)

in collaboration with:

Roberto De Pietri (Parma Univ)

Roberto Alfieri (Parma Univ)

Alessandra Feo (Parma Univ)

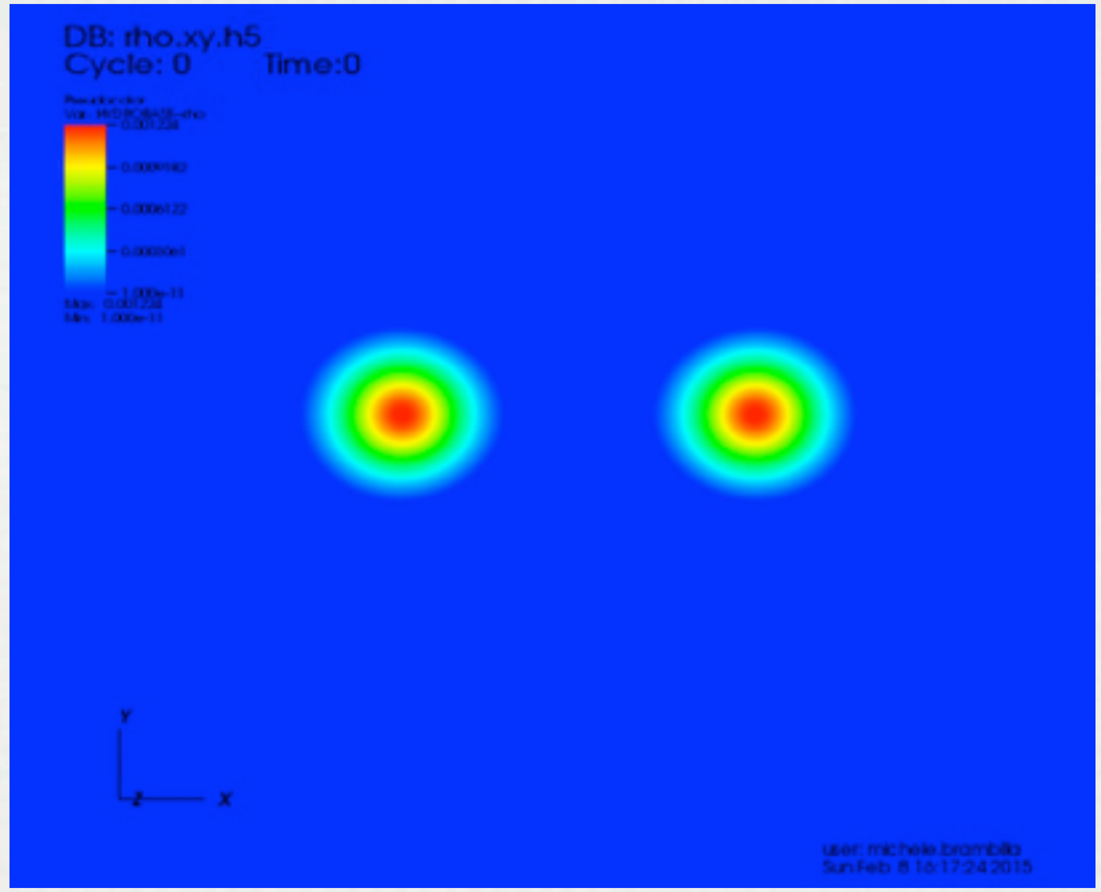
Francesco Maione (Parma Univ)

AT THE BEGINNING.. THE FINAL GOAL

This talk will be about computing, but leT me first spend some words on physics

- ☐ high resolution simulation of inspiral and merger phase of binary neutron stars system
- ☐ most likely source of gravitational waves expected to be observed by the VIRGO experiment
- ☐ strong EM emissions (engine of short gamma ray burst?)

a low resolution example of BNS merger



STATE OF THE ART COMPUTATIONS

PRD 90, 041502(R) (2014)

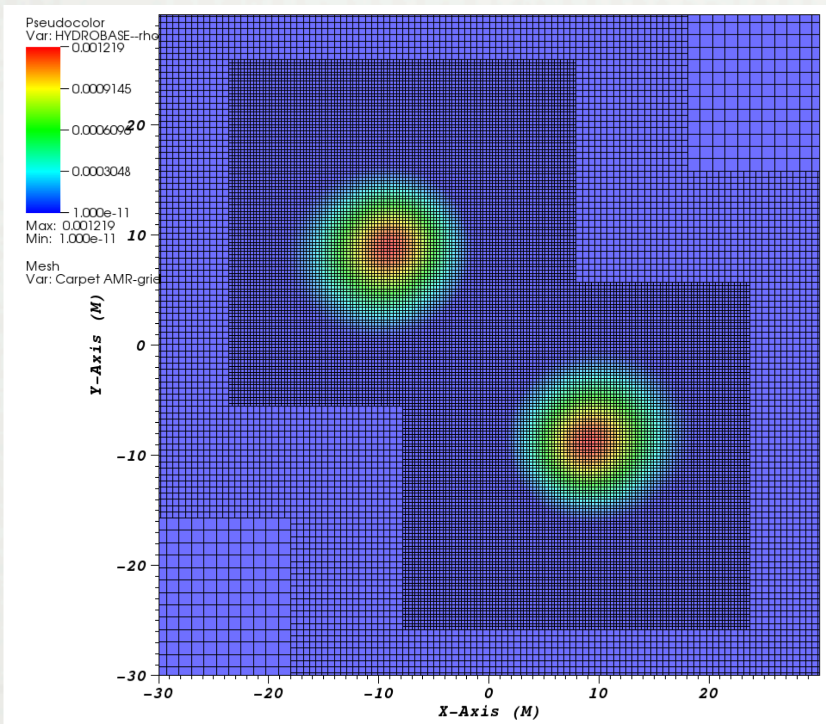
- resolution: 150 to 70 m
- simulated time: ~100ms
- $\log_{10}$(Bmax[G]): 14 to 16
- piecewise polytrope EOS
- performed on K supercomputer, ~10PF

NUMERICAL RELATIVITY

Einstein equations	$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} = 8\pi G T_{\mu\nu}$
Conservation laws	$\nabla_\mu T^{\mu\nu} = 0$ $\nabla_\mu (\rho u^\mu) = 0$
Equation of state	$p = p(\rho, \epsilon)$

- ☐ 6 equations for the metric
- ☐ 6 equation for the extrinsic curvature
- ☐ 1 hamiltonian + 1 momentum constraint
- ☐ 1 gauge condition

the computational challenge we are dealing with: time evolution of a set of PDE on a carTesian grid



- a grid of 1000^3 with 3 time levels and 10 variables per site requires 300 GB of memory
- if the update of each variable requires 50 flop per time step we are dealing with ~1TFlop
- we usually need (at least) 10-20K time steps

- different number of FP variables associated to each grid point
- different number of FP ops for the update of different variables
- different levels of refinement
- memory requirement grows fast increasing resolution (~1/r^3)

THE EINSTEIN TOOLKIT

The Einstein Toolkit is an open source set of tools for simulating and analyzing relativistic astrophysical systems

Einstein Toolkit

- based on Cactus infrastructure
- initial data, vacuum space-time solver, hydrodynamic solver, analysis tools
- ~ 500K lines of code
- currently ~50 sites worldwide
- regular tested releases every ~6 month

Cactus: the underlying computational infrastructure

- general framework for development of portable, modular applications
- programs are split into independent components (thorns)
- thorns are developed independently and should be interchangeable
- support for C,C++, Fortran

SYSTEMS EXPLORED

- ☐ FERMI
- ☐ ZEFIRO
- ☐ EURORA
- ☐ GALILEO

- ☐ **Model:** IBM-BlueGene /Q
- ☐ **Processor Type:** IBM PowerA2, 1.6 GHz
- ☐ **Computing Nodes:** 10.240 with 16 cores each
- ☐ **Computing Cores:** 163.840
- ☐ **RAM:** 16GB / node; 1GB/core
- ☐ **Internal Network:** Network interface with 11 links ->5D Torus
- ☐ **Peak Performance:** 2.1 PFlop/s

SYSTEMS EXPLORED

- ☐ FERMI
- ☐ ZEFIRO
- ☐ EURORA
- ☐ GALILEO

- ☐ **Model:** Linux cluster
- ☐ **Processor Type:** AMD Opteron 6380 2.50 GHz
- ☐ **Computing Nodes:** 128 (16 cores each)
- ☐ **Computing Cores:** 2048
- ☐ **RAM:** 512 GB / node

SYSTEMS EXPLORED

- FERMI
- ZEFIRO
- EURORA
- GALILEO

- **Model:** Eurora Prototype
- **Processor Type:** Intel Xeon (Eight-Core SandyBridge) E5-2658 2.10 GHz, E5-2687W 3.10 GHz
- **Computing Nodes:** 64 (16 cores each)
- **Computing Cores:** 1024
- **RAM:** 16 GB / node
- **Accelerators:** 64 nVIDIA Tesla K20 + 64 Intel Xeon Phi (MIC)

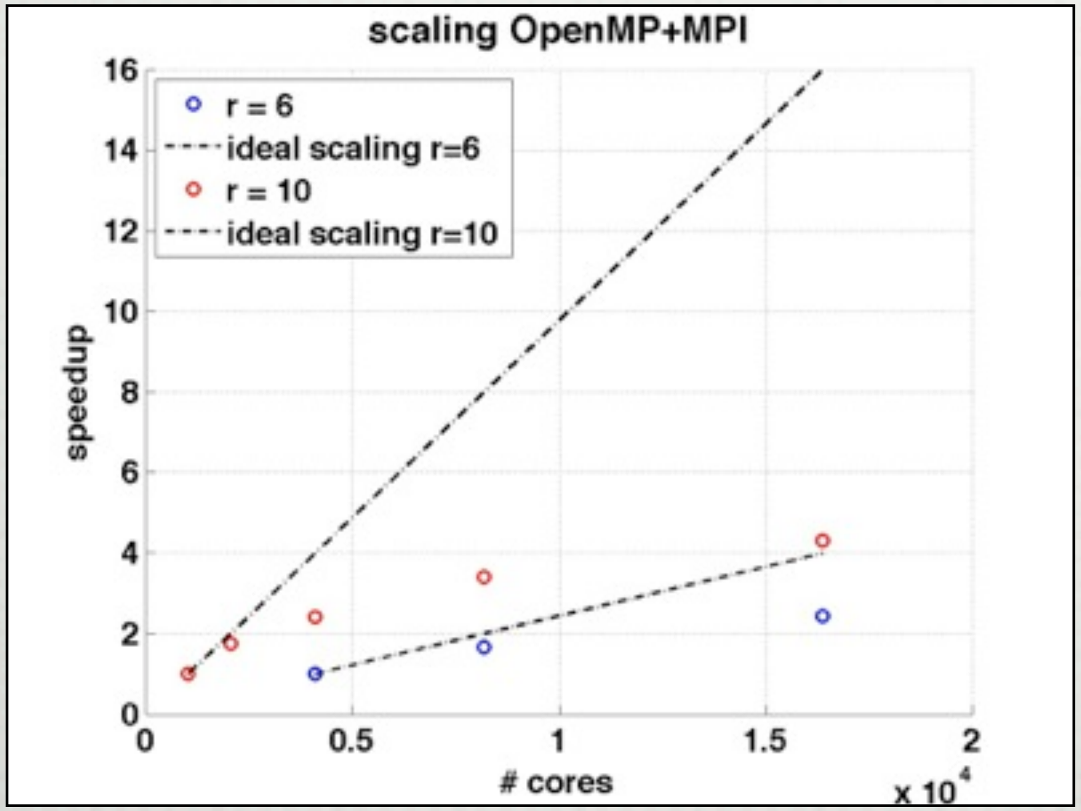
SYSTEMS EXPLORED

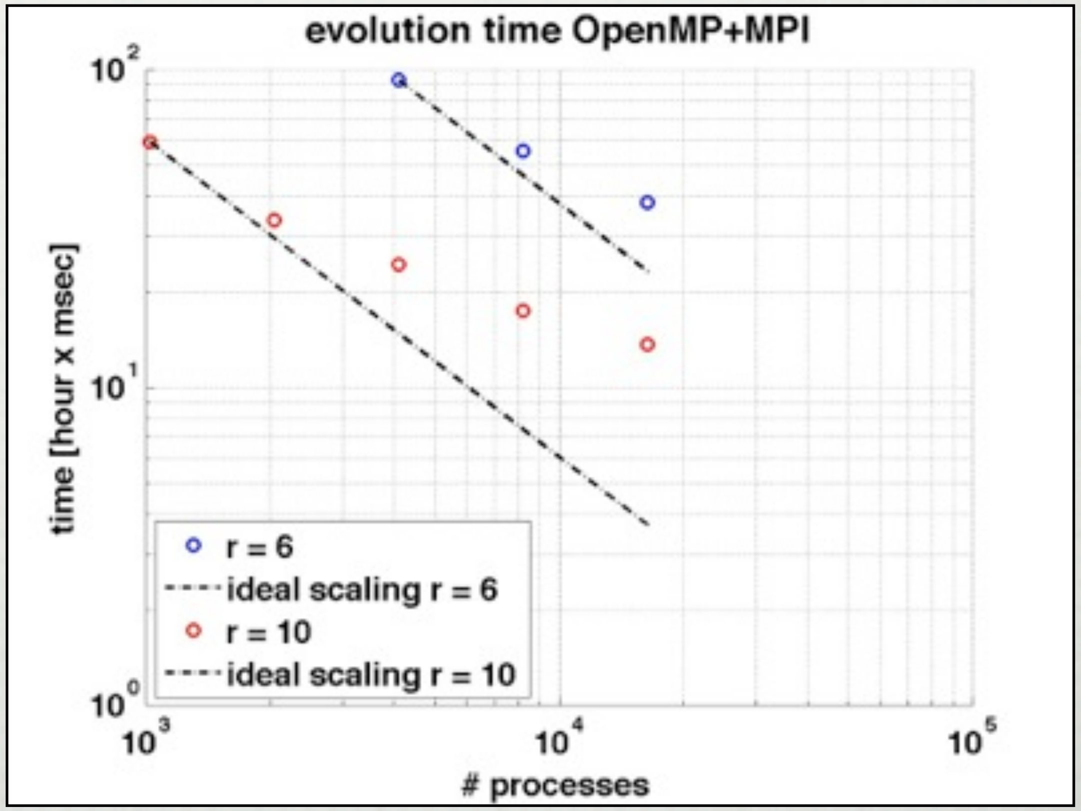
- ☐ FERMI
- ☐ ZEFIRO
- ☐ EURORA
- ☐ GALILEO

- ☐ **Model:** IBM NeXtScale
- ☐ **NODES:** 516
- ☐ **PROCESSORS:** 8-cores Intel Haswell 2.40 GHz (2 per node)
- ☐ **CORES:** 16 cores/node, 8256 cores in total
- ☐ **ACCELERATORS:** 2 Intel Phi 7120p per node on 384 nodes (768 in total)
- ☐ **RAM:** 128 GB/node, 8 GB/core
- ☐ **INTERNAL NETWORK:** Infiniband with 4X QDR switches
- ☐ **PEAK PERFORMANCE:** TO BE DEFINED

FERMI

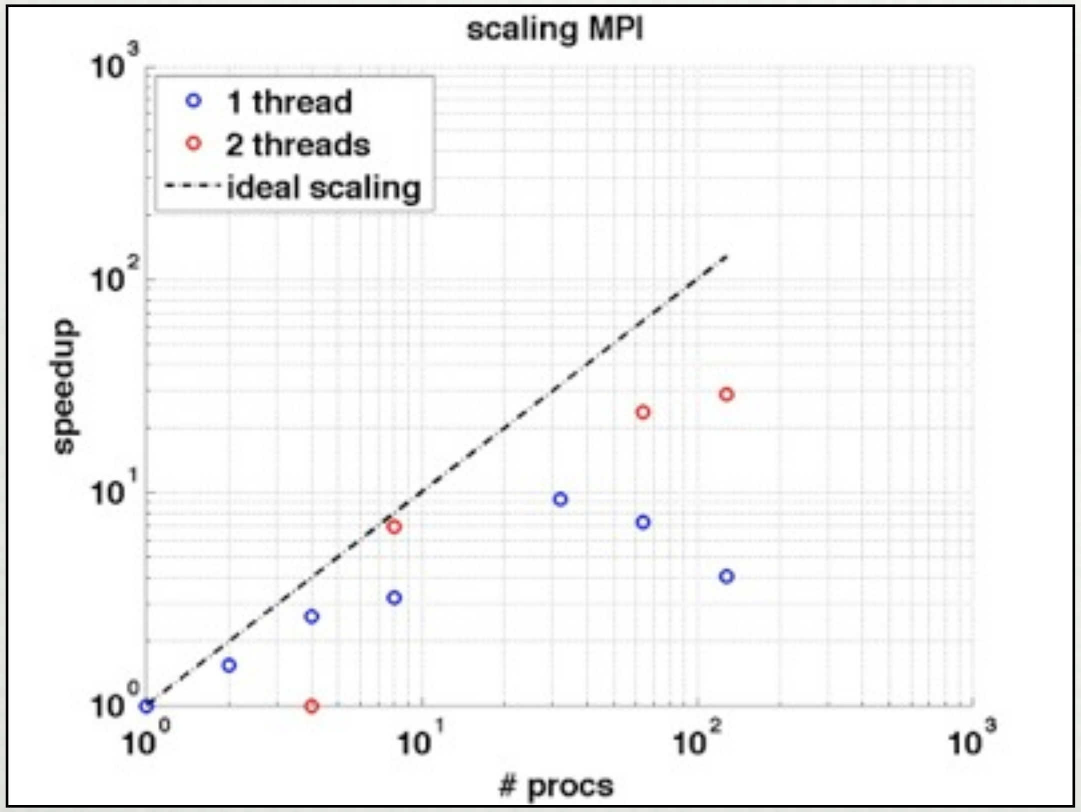
- well known "reference" architecture
- explored (strong) scaling at different resolutions

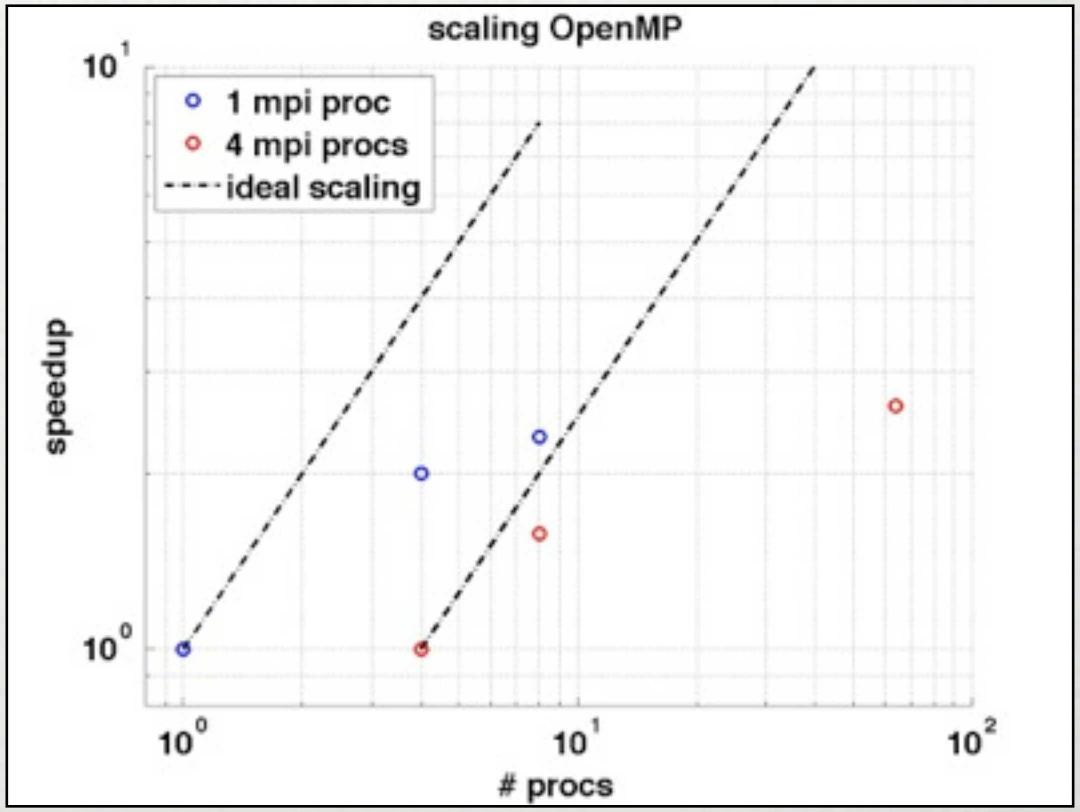




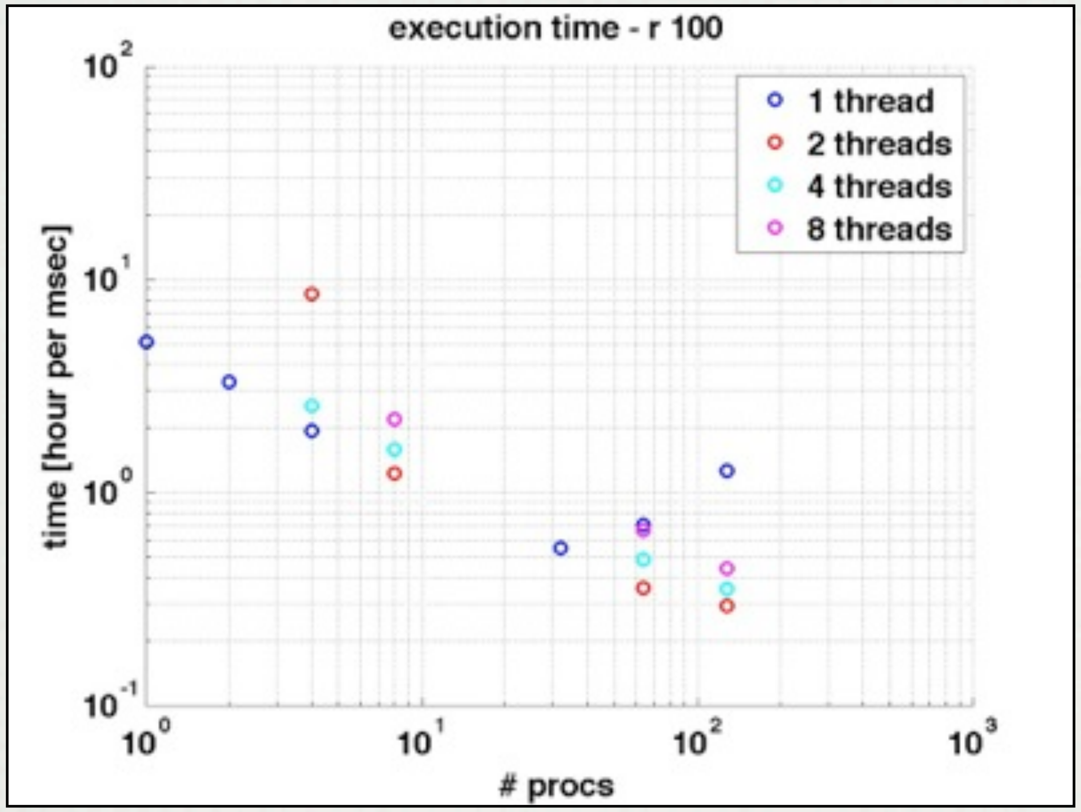
ZEFIRO

- consider r=100
- inspect differences MPI vs OpenMP





- MPI looks scale better than OpenMP (affinity issue?)

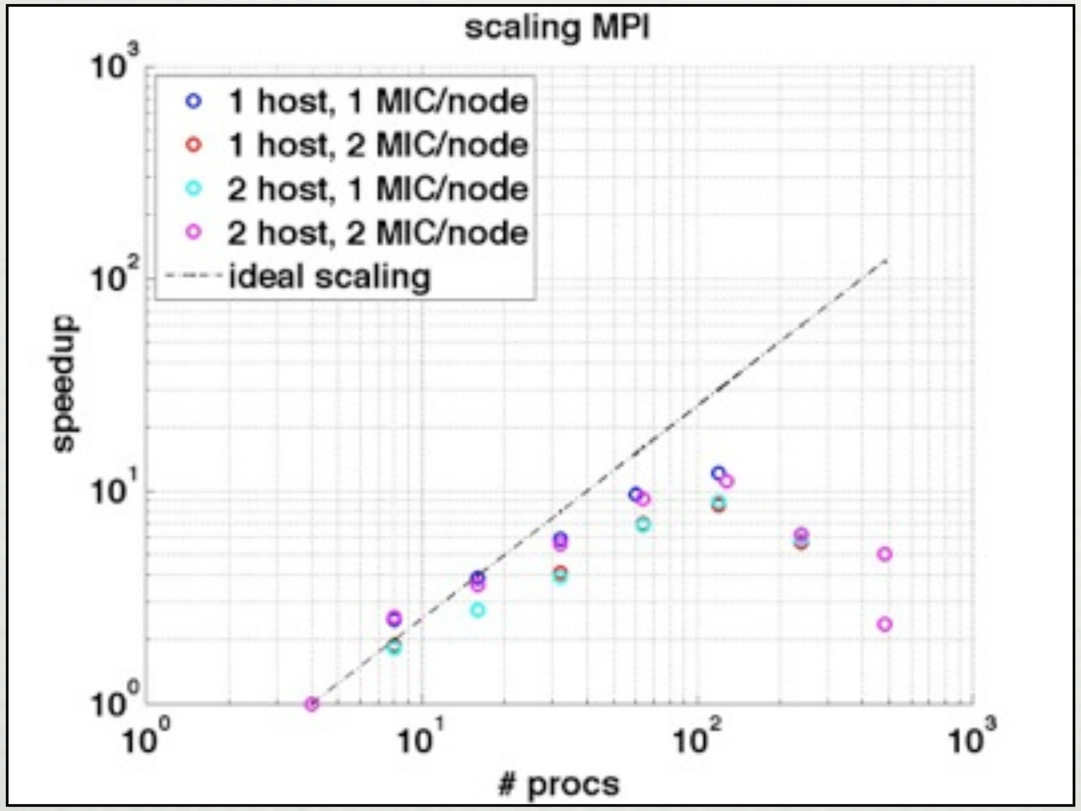


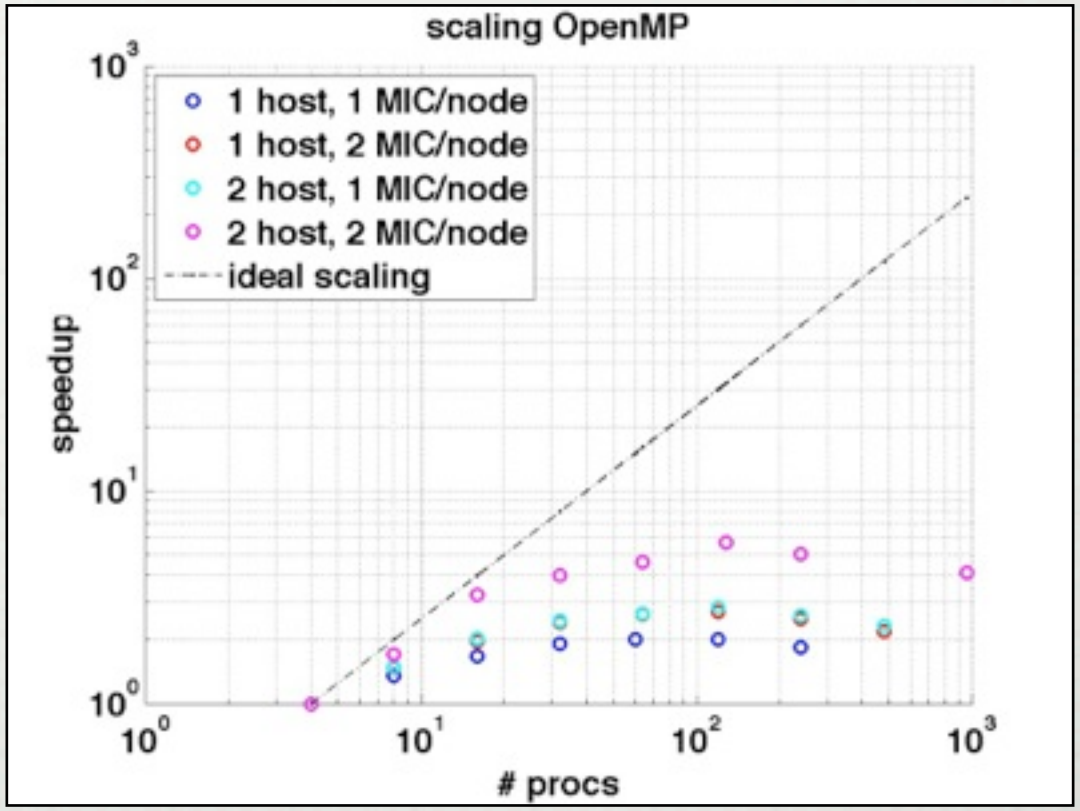
we use all the possible processor on the board: comparison is "fair" w.r.t. cache effects

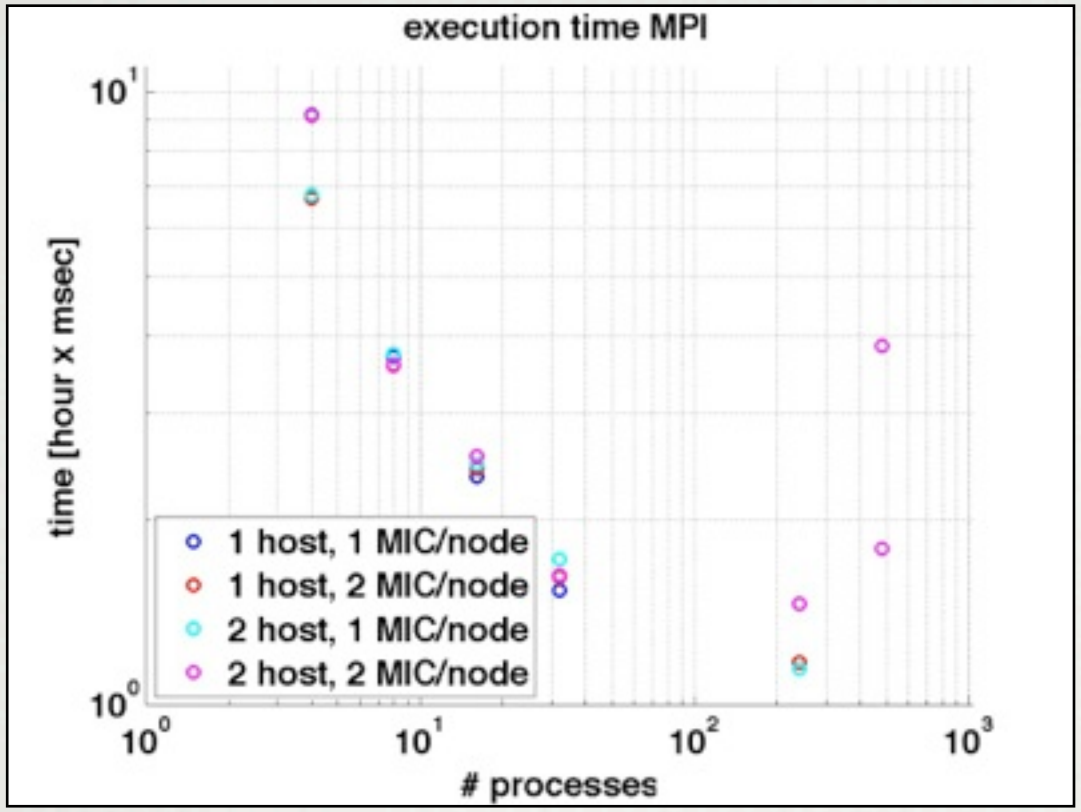
keeping a small number of threads and use MPI parallelization seems to be the best approach

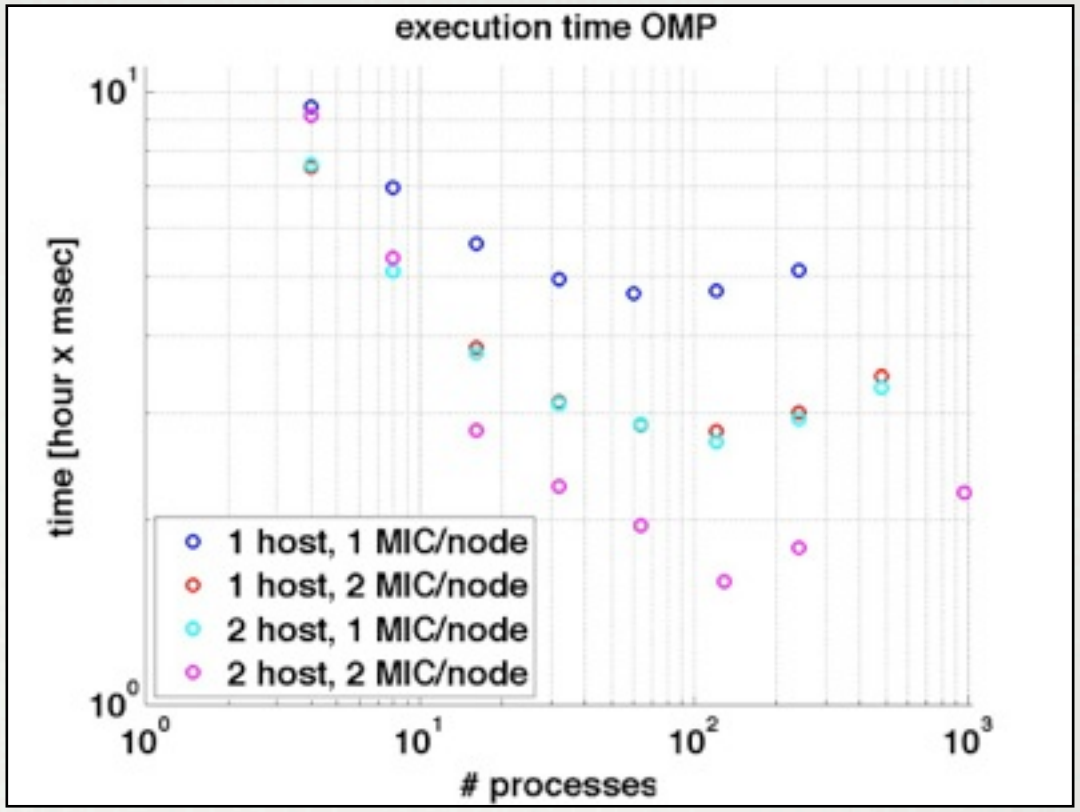
EURORA

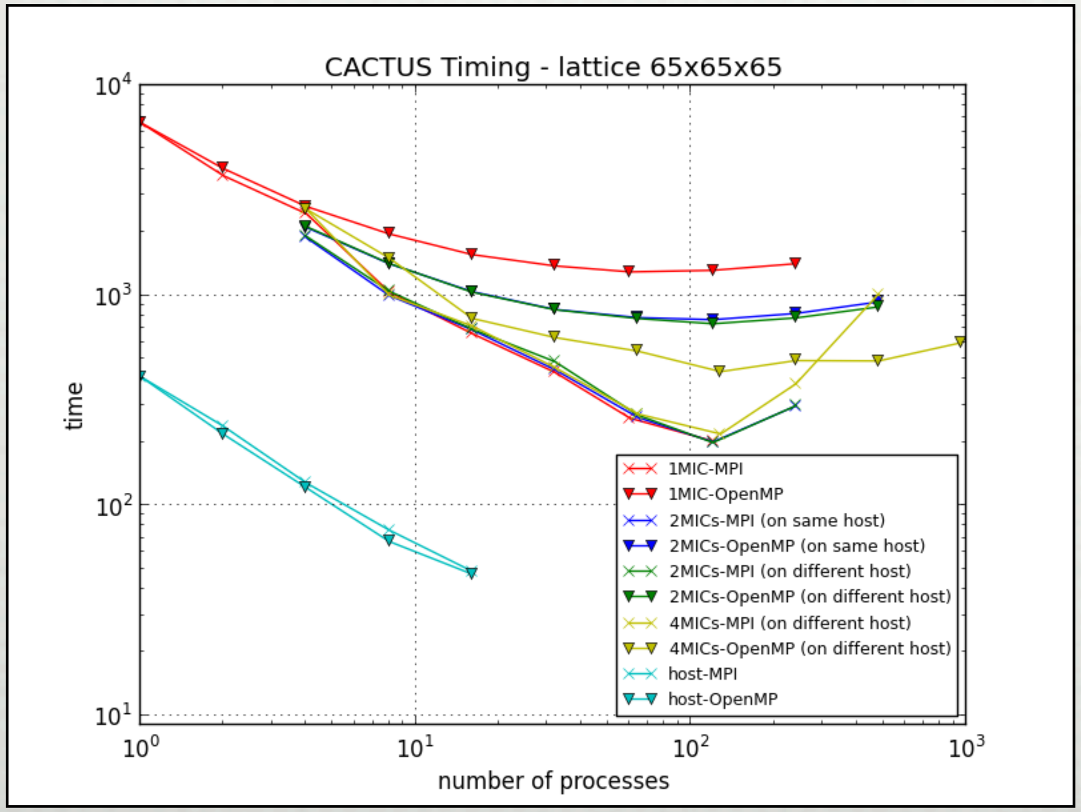
- nodes vs accelerator (MIC)
- MPI vs OpenMP







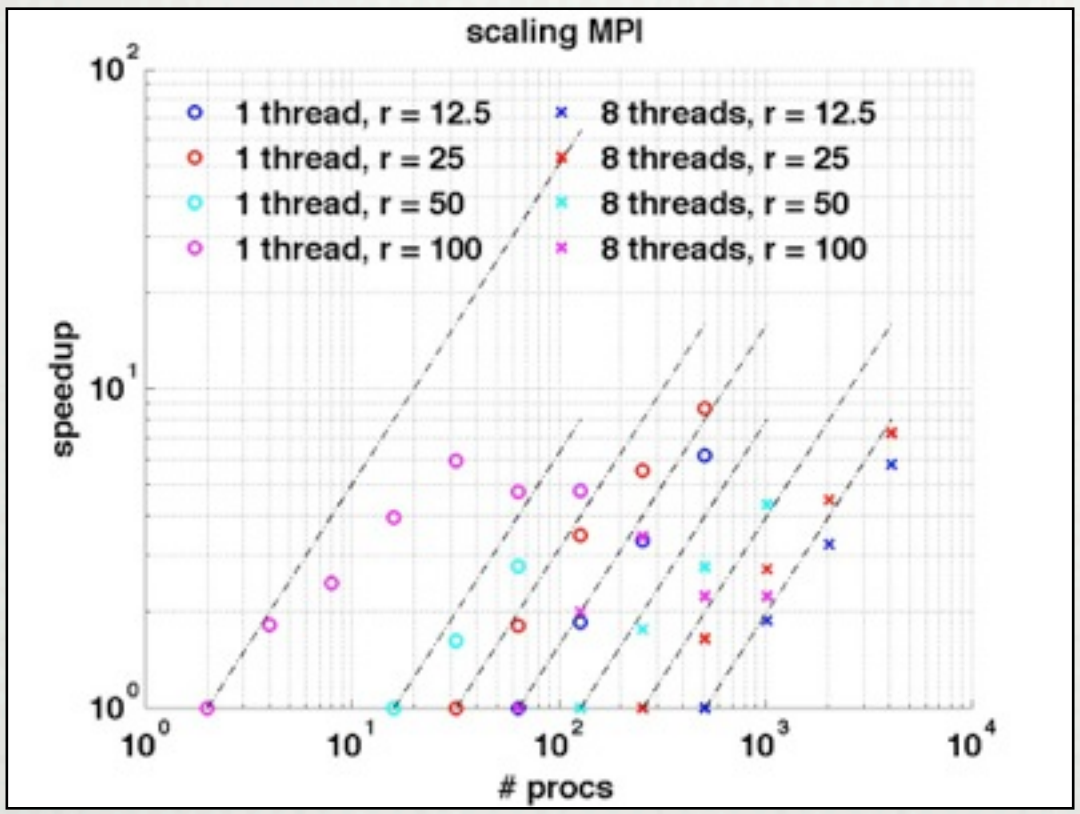


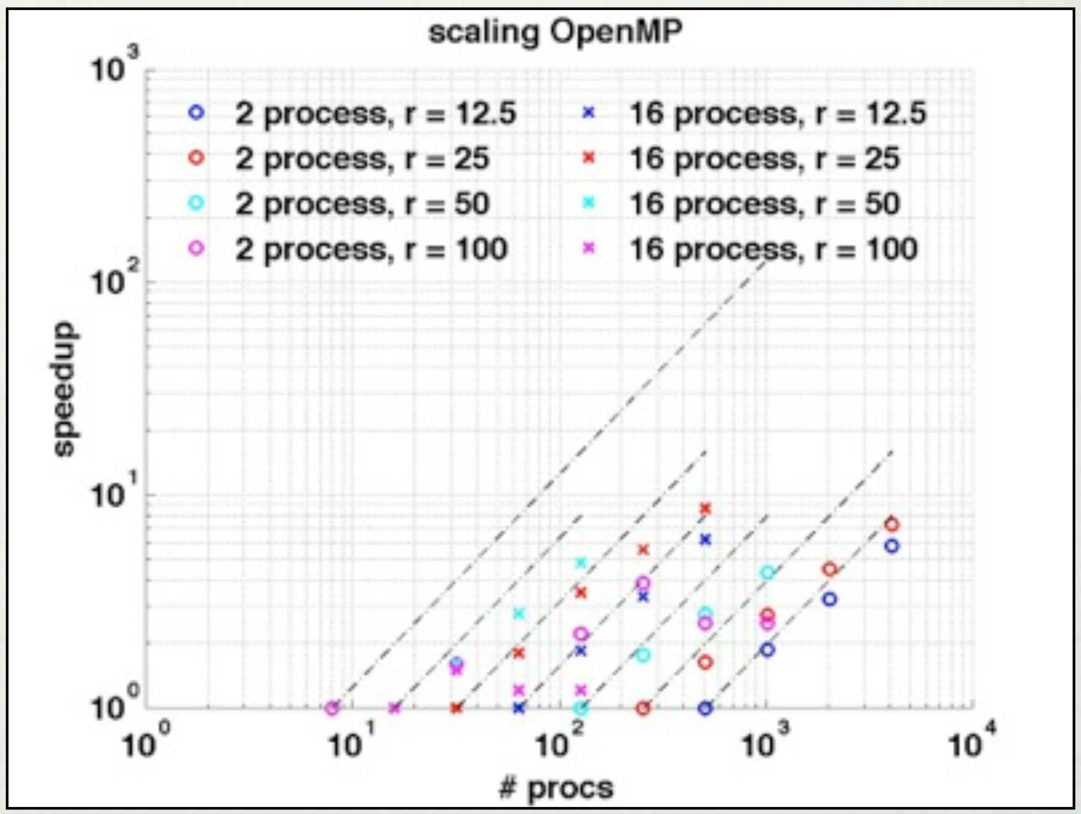


- MIC: ~1 TFlops in double precision (240 processes)
- host: ~240 GFlops in double precision (16 processes)

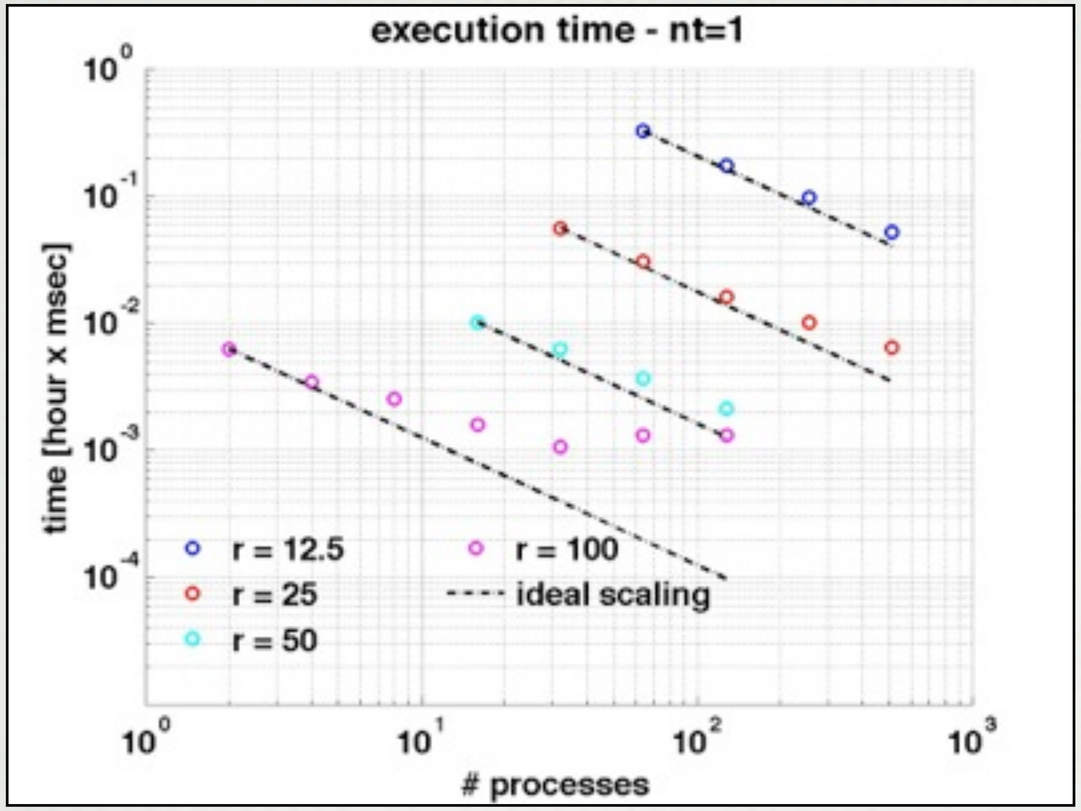
GALIELO

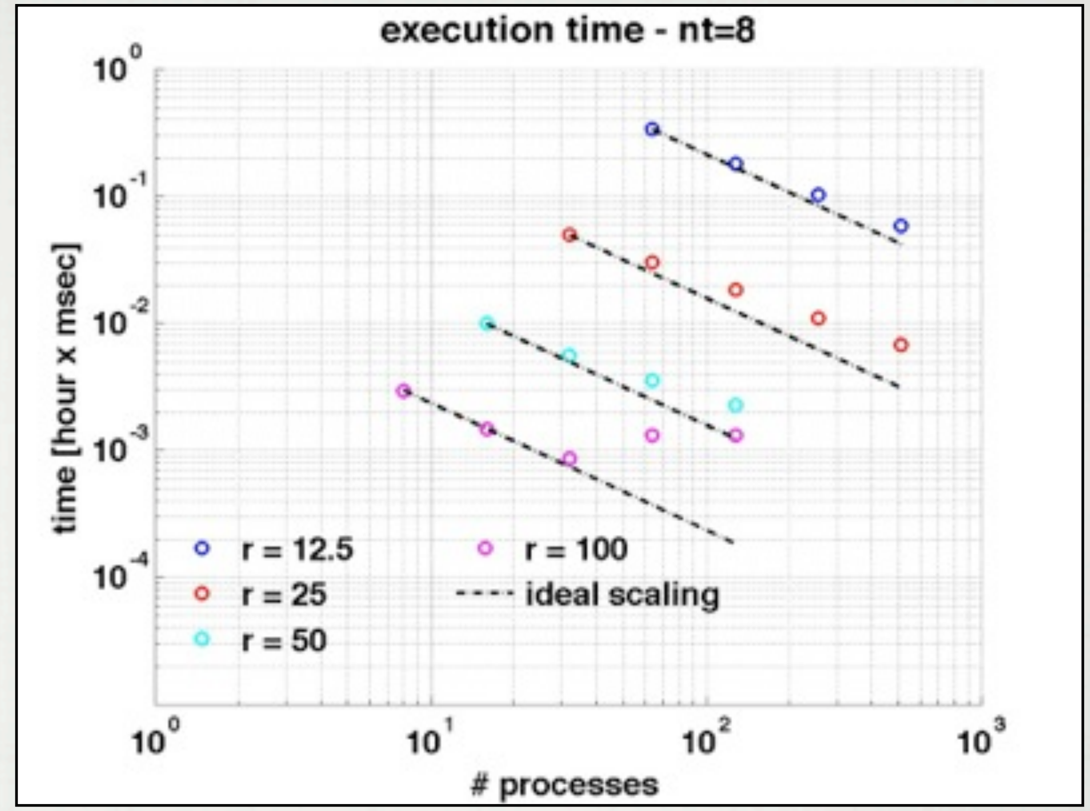
- strong and weak scaling
- inspect differences MPI vs OpenMP

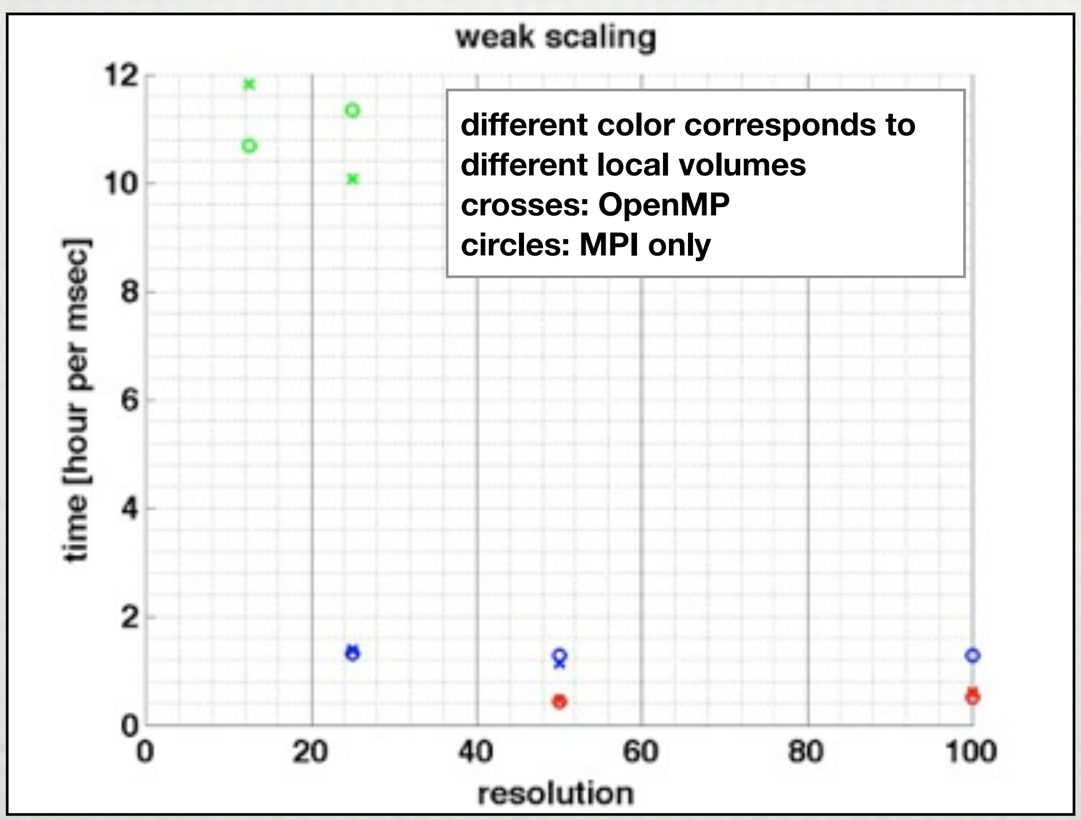


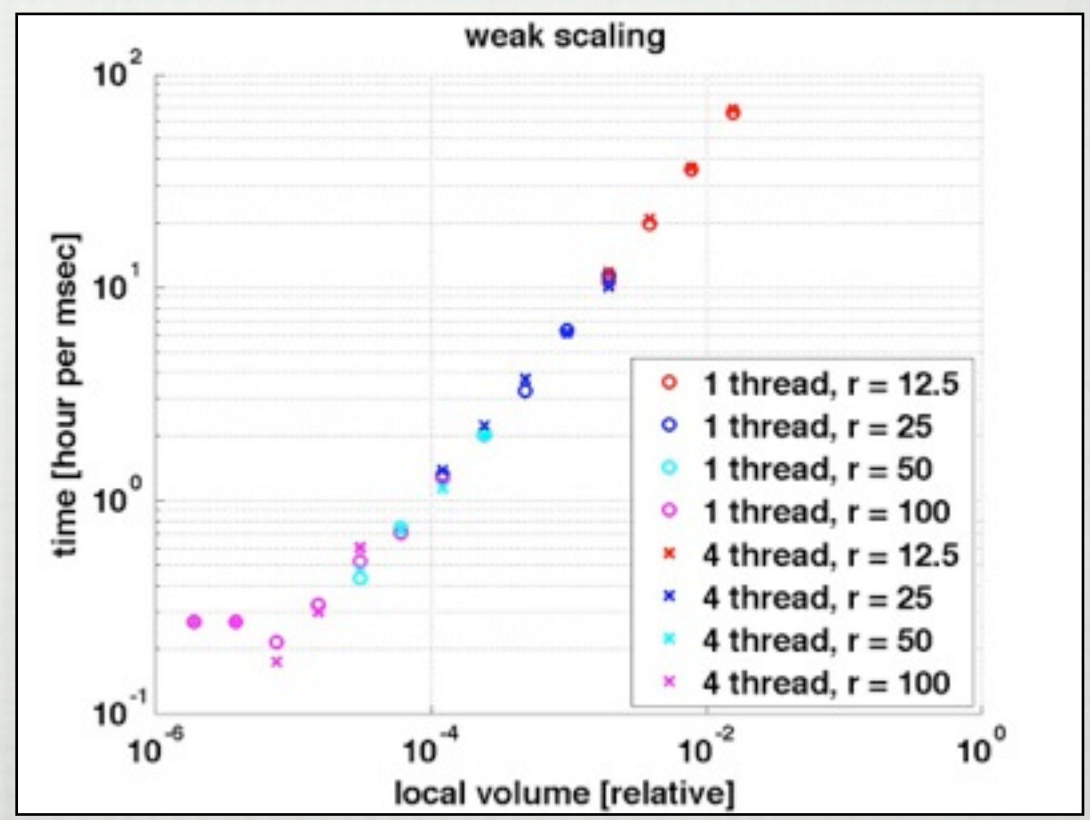


- less sensitive to MPI or OpenMP
- scaling improves increasing volume


execution time - nt=1
time [hour x msec]
# processes
r = 12.5
r = 25
r = 50
r = 100
ideal scaling


execution time - nt=8
time [hour x msec]
# processes
r = 12.5
r = 25
r = 50
r = 100
ideal scaling


weak scaling
time [hour per msec]
resolution
different color corresponds to
different local volumes
crosses: OpenMP
circles: MPI only


weak scaling
time [hour per msec]
local volume [relative]
1 thread, r = 12.5
1 thread, r = 25
1 thread, r = 50
1 thread, r = 100
4 thread, r = 12.5
4 thread, r = 25
4 thread, r = 50
4 thread, r = 100

COMPARISON OF MACHINES

	peak performance/ node	simulated time (msec/hour)	relative performance
fermi*	200 GFlops	500	0.98
zefiro	160 GFlops	369	0.9
galileo	300 Gflops	765	1

*extrapolated to 1 node assuming perfect scaling

CONCLUSIONS

- NUMERICAL RELATIVITY ALLOWS THE STUDY OF THE EXPECTED FORM OF GW SIGNAL. BESIDES THE DETECTION OF GW THIS CAN GIVE HINTS ON THE STELLAR EOS
- SIMULATIONS SCALE WITH (1/RESOLUTION) ^ 4: A RESOLUTION OF 50 OF MERGER OF BNS REQUIRES ~PFLOP
- CURRENT AND FORTHCOMING ARCHITECTURES ARE VIABLE TO SUCH SIMULATIONS
- WEAK SCALING WORKS, CAN STRONG SCALING BE IMPROVED FOR SMP APPROACH?
- COULD THE NEW STANDARD OPENMP 4.0 OFFER A SOLUTION FOR THE OFFLOADING ON ACCELERATORS (MIC, GPU)?

THANKS FOR YOUR ATTENTION