TheoMpi:
a large MPI cluster on the grid for Theoretical Physics

Roberto Alfieri

Parma University & INFN – Italy

Co-authors:

S. Arezzini, A. Ciampa, E. Mazzoni (INFN-PI), A. Gianelle, M. Sgaravatto (INFN-PD),
S. Monforte, G. Andronico (INFN-CT), R. De Pietri, F. Di Renzo (INFN-PR).

Outline

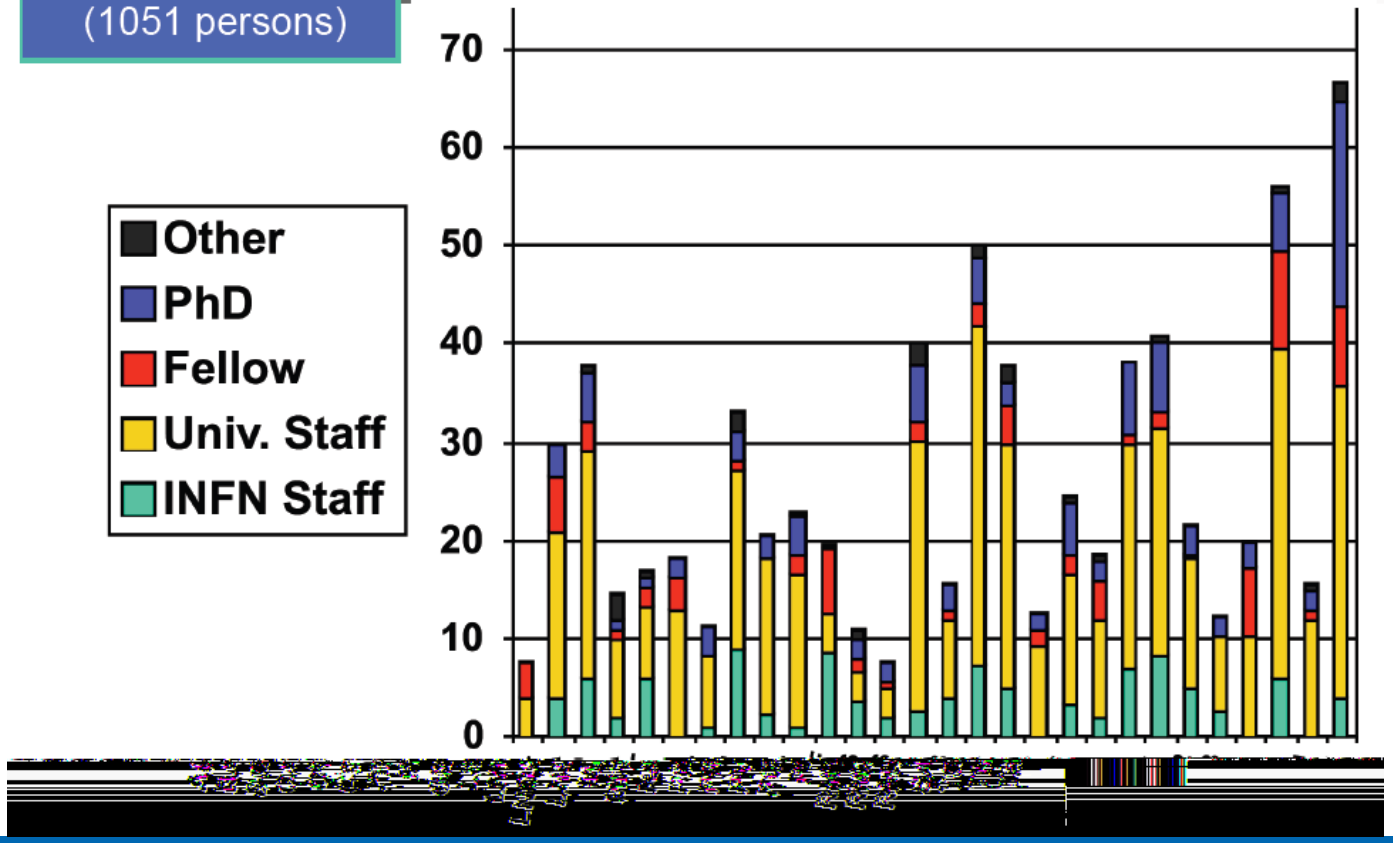
- INFN theoretical physics community and computational needs
- The TheoMpi project
- MPI support in Grid and new “Granularity” attributes
- TheoMpi jobs submission
- Examples of physics parallel applications executed on the cluster
- Conclusions and future works

The Community

FTE
2008: 717.4
(950 persons)
FTE 2009:
721.8
(1051 persons)

STRUCTURE OF CSN4

3 National Labs and
25 Divisions (Sezioni)



Reseach Areas

• String & Field Theory:
QFT, Strings & M-Theory, Gravity, Lattice Gauge Th. and confinement, AdS/CFT;

• Particle Phenomenology:
SM and BSM (Susy, Extra Dim., Composite Higgs), QCD at colliders (MC simulations, finite T and μ), Flavor Physics (and lattice) and EFT for heavy flavors, AdS/CFT and QCD;

• Hadronic & Nuclear Physics:
nuclear structure and reactions (radioactive beams, stability valley and beyond), heavy ion collisions (quark-gluon plasma, saturation, jet quenching, T and μ phase transitions), confined hadronic matter (spin structure of hadrons, exotic spectroscopy, GPD);

• Mathematical Methods:
general relativity (gravitational waves,..), quantum theory (foundations, chaos,..), conformal field theories;

• Astroparticle & Cosmology:
neutrino physics, “dark things” (matter, energy,..), astrophysical radiation sources, astronuclear physics, gravitational waves

• Statistical Field Theory & Applications:
complex & non-eq. systems, spin glasses, applications (quant. biology, turbulence,..).

The activities are organized in 54 research projects denoted **Iniziativa Specifica (IS).**

Computational needs

20% of the computational activity

- lattice simulations (LGT, gravitational physics, turbolence, …)
- typical jobs: 2011 needs O(10 TFlops-year) - 2015: O(PFlops-year)
- computational resources (last 20 years) : APE supercomputer family *(web site)*
- current main resource: apeNEXT *(web site)*

80% of the computational activity

- general purpose (mainly numeric computations)
- typical jobs are serial and parallel, but a lot of them
- computational resources (last 10 years) : PC clusters
- current main resource : **CSN4cluster (TheoMPI project)**

timeline

late 2009: cluster requirements definition and sites proposal evaluation

Febr. 2010: INFN-Pisa project approved

June 2010: cluster in operation for sequential jobs

July 2010: call for scientific proposals

Sept. 2010: projects approved and fair-shares defined

Dec. 2010: cluster in operation for parallel jobs

Access method: via Grid only (both serial and parallel jobs)

Access policy:

- Theophys VO members (~130 up to now) with low priority
- Theophys subgroups (or others) can apply for a granted fairshare to the CSN4 cluster committee

Active fairshare grants:

- 16 fairshares have already been assigned, corresponding to 16 CSN4 IS proposals
- Requests: 130K day*core serial + 250K day*core parallel = 380K day*core
  (availability: 365 K day*core per year)

Details: *http://wiki.infn.it/cn/csn4/calcolo/*

resources

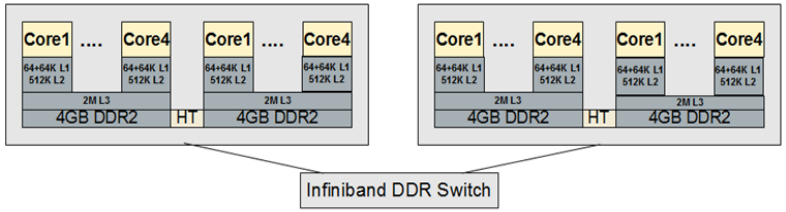
The cluster is installed, configured and maintained by the **INFN-Pisa computing center** *(web site)*

Computing:

128 WNs dual Opteron 8356, 2x4 cores per node, 8GB ram, 1024 cores,10TFlops peak perf.

- SW: Linux x86_64, openMPI

1 CE gridce3.pi.infn.it : Cream-CE, LSF



High Speed Network:

Infiniband DDR

Storage:

1 SE gridsrm.pi.infn.it : STORM

10 TBytes GPFS over Infiniband

queues organization

Direct job submission to Cream-CE in the JDL

```
Requirements =(other.GlueCEInfoHostName == "gridce3.pi.infn.it")
```

gives access to 2 queues:

- **theompi** : parallel job only - runtime 72h - reservation time 8h - Role=parallel required

```
voms-proxy-init –voms theophys:/theophys/<group_name>/Role=parallel
```

- **theoshort**: serial short jobs - runtime 4h - Role=parallel will not be specified

```
voms-proxy-init –voms theophys:/theophys/<group_name>
```

The serial queue allows the exploitation of cores when they are unused by parallel jobs using two scheduling techniques:

- **Slots reservation** ensures a lock on a number of processors
- **Backfill** allows serial Jobs to use reserved parallel Job Slots

MPI support in gLite

It comes from EGEE MPI-WG (2007-2008) recommendations: *http://www.grid.ie/mpi/wiki/*

The support is based on **MpiStart**

MpiStart is a set of scripts which main advantage is the possibility to detect and use site-specific configuration features, like:

- **MPI implementations** (openMPI, MPICH, MPICH2, PACX-MPI, LAM)
- **Batch scheduler** (SGE, PBS, LSF)
- **File distribution** (if the home isn't shared)
- **Workflow control** (user's pre/post execution scripts)

MPI and multi-thread support in EMI

Despite the work of the MPI-WG, MPI in grid was scarcely used in 2009.

In 2009 EGEE-III designated a new MPI-WG
Purpose:
- Investigate why Mpi isn't used and provide new enforced recommentations
- Provide a solution for the support of the upcoming multicore architectures

Recommendation document released in 06/2010:
http://www.grid.ie/mpi/wiki/WorkingGroup

Main Recommendations:

-- MPI-start is confirmed
-- MPI Packages: multiple MPI flavours support in gLite (MPICH JobType deprecated)
-- Shared file-system
-- SSH password-less among WNs
-- Functionality tests (SAM) regularly executed
-- Documentation and training
-- **New JDL TAG is introduced to support multicore architectures**

Semantics and syntax

Attribute	Meaning
CPUNumber=P	Total number of required CPUs
SMPGranularity=C	Minimum number of cores per node
HostNumber=N	Total number of required nodes
WholeNodes=true	Reserve the whole node (all cores)

New JDL Attributes

```
CPUNumber = 64;          # 32 nodes, with 2 CPUs per node
SMPGranularity = 2;      # (SMPsize >=2 )

CPUNumber  = 16;         # 2 nodes, with 8 CPUs per node
HostNumber = 2;          # (SMPsize >=8 )

WholeNodes=true;         #  2 whole nodes with SMPsize>=8
HostNumber=2;
SMPGranularity=8;

WholeNodes=true;         # 1 whole node with SMPsize>=8
SMPGranularity=8;        # (default HostNumber=1)
```

Examples

Preliminary patch

The New JDL attributes proposed by the WG **aren't implemented in gLite yet**

- CE support is coming with Cream-CE 1.7 (EMI-1 release ?)

A **preliminary patch for Cream-CE** has been developed and tested in collaboration with the gLite middleware developers

It comes with a **different syntax but the same semantics**.

Examples:

```
CeRequirements = "wholenodes=\"true\" && hostnumber==2";     # 2 whole nodes
```

```
CPUNumber = 16;                                              # 8 nodes with 2 CPUs per node
CeRequirements = "SMPGranularity==2"
```

The patch has been **installed and is working on the TheoMpi cluster.**

Job types

- **Pure MPI jobs**

via Mpi-Start

- **Multi-thread jobs**

require a single "**Wholenodes**" with C cores and start C threads

- **Hybrid MPI-openMP jobs**

require N "**Wholenodes**" with C cores each and start R<N MPI ranks per node.
Each Mpi rank will spawn C/R threads.

MPI job submission

MPI-start is the submission method recommended by the EGEE MPI-WG.

The Mpi-Start tool handles most of the low-level details of running the MPI job on a particular site.

This example executes 16 MPI ranks (2 whole nodes):

JDL:

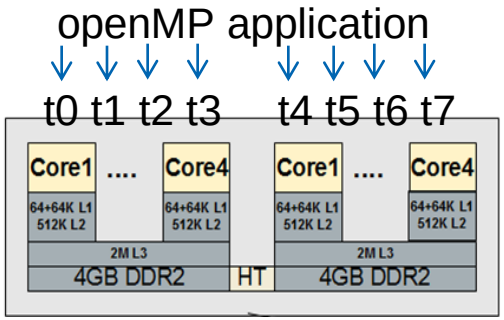
```
Executable = "mpistart-wrapper.sh";
Arguments = “mympi OPENMPI";
InputSandbox = {"mpistart_wrapper.sh","mpi-hooks.sh",“mympi.c"};
Requirements =(other.GlueCEInfoHostName == "gridce3.pi.infn.it");
CeRequirements = "wholenodes==\"true\"  && hostnumber==2";
```

- **mpistart_wrapper.sh** is a standard script provided by the mpi-start package
- **mpi-hooks.sh** includes pre and post execution scripts

multi-thread jobs

Wholenodes attribute allows the submission of **multi-thread jobs**
(or jobs requesting the **exclusive usage of the node memory**)

- This example executes 8 openMP threads on a whole node -



JDL

```
Executable = "openmp.sh";
Requirements =(other.GlueCEInfoHostName == "gridce3.pi.infn.it");
CeRequirements = "wholenodes==\"true\"  && hostnumber==1";
```

openmp.sh

```
export OMP_NUM_THREADS=8
./myomp
```

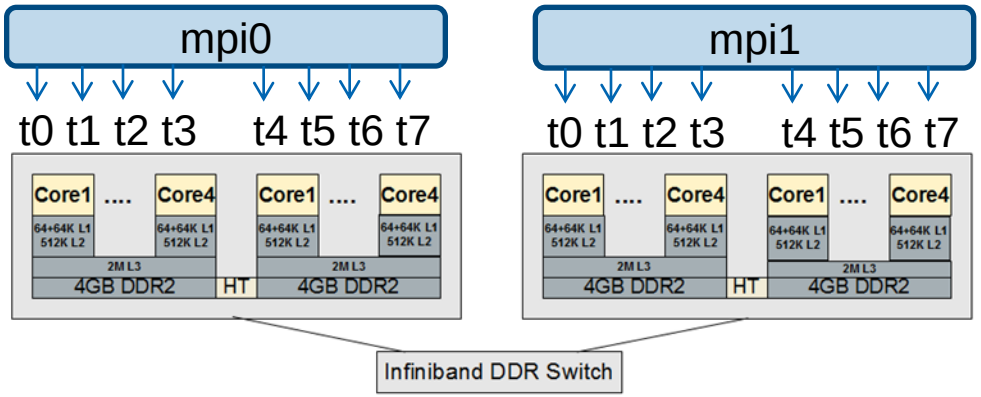
$LSB_HOSTS

```
csn4wn110
csn4wn110
csn4wn110
csn4wn110
csn4wn110
csn4wn110
csn4wn110
csn4wn110
```

hybrid MPI-openMP jobs

Hybrid MPI-openMP programming is not supported yet by mpi-start (EMI-1 ?), but we know execution environment details (SMP arch., MPI flavour, queue manager,..) **so we can directly submit the job.**

This example requires
2 MPI ranks with 8 openMP threads each.



LSB_HOSTS

```
csn4wn110
csn4wn110
csn4wn110
csn4wn110
csn4wn110
csn4wn110
csn4wn110
csn4wn110
csn4wn111
csn4wn111
csn4wn111
csn4wn111
csn4wn111
csn4wn111
csn4wn111
csn4wn111
```

JDL

```
Executable = "mpi_openmp.sh";
Requirements =(other.GlueCEInfoHostName == "gridce3.pi.infn.it");
CeRequirements = "wholenodes==\"true\"  && hostnumber==2";
```

mpi_openmp.sh

from LSB_HOSTS to nodefile.txt

```
echo  $LSB_HOSTS | tr ' ' '\n' | sort -u > nodefile.txt
NP=$[`cat  nodefile.txt  |  wc --lines`]
mpirun -np $NP  -x OMP_NUM_THREADS=8  --hostfile nodefile.txt   mympiomp
```

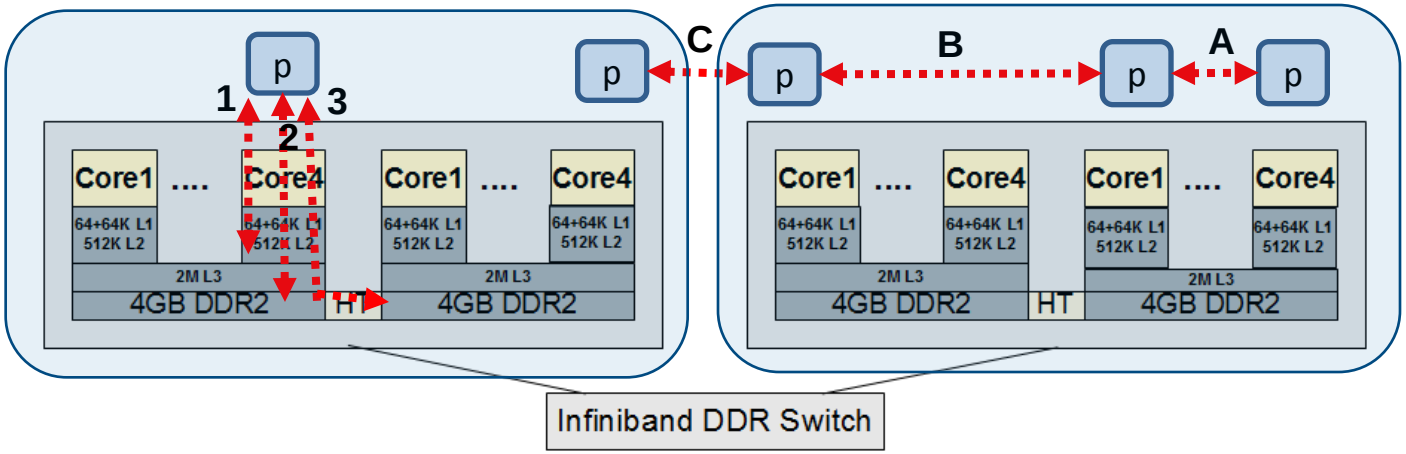
nodefile.txt

```
csn4wn110
csn4wn111
```

memory architecture

In modern multicore processors the memory architecture is **NUMA**

- Cpu/memory **affinity** is the ability to bind a process to a specific CPU/memory bank -



Memory theoretical peak performance:

	Memory Type	Latency	Bandw.
1	L3 cache	≈35 ns	
2	RAM	≈ 50 ns	≈30 GB/s
3	Numa (HT or QPI)	≈ 90 ns	≈10 GB/s

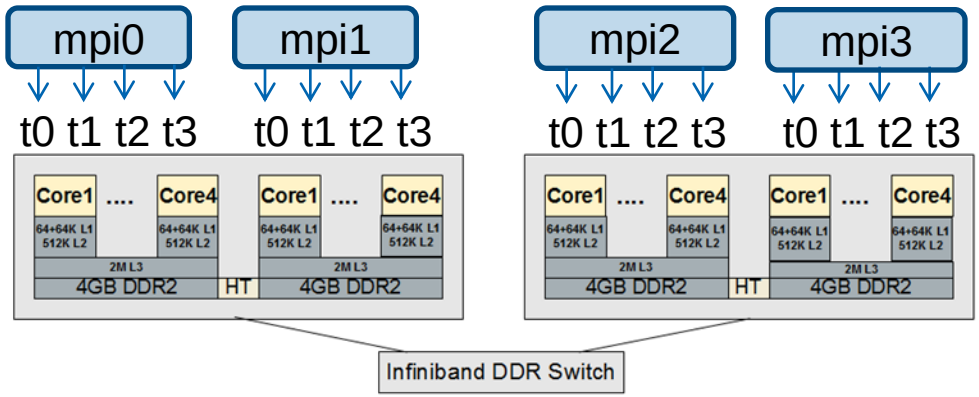
Measured communication performance (using NetPIPE)

	Comm. Type	Latency	Bandw.
A	Shm (intra-socket)	640 ns	14 GBytes/s
B	Shm (inter-socket)	820 ns	12 GBytes/s
C	infiniband	3300 ns	11 GBytes/s

hybrid MPI-openMP jobs with affinity

openMPI supports the **CPU affinity** by means of the Rankfile. CPU affinity is not supported yet by mpi-start (EMI-1 ?), but we can directly submit the job.

This example requires 4 MPI ranks with 4 openMP threads each



nodefile.txt

```
csn4wn110
csn4wn111
```

rankfile.txt

```
rank 0=csn4wn110 slot=0-3
rank 1=csn4wn110 slot=4-7
rank 2=csn4wn111 slot=0-3
rank 3=csn4wn111 slot=4-7
```

JDL

```
Executable = "mpi_openmp.bash";
Requirements =(other.GlueCEInfoHostName == "gridce3.pi.infn.it");
CeRequirements = "wholenodes==\"true\"  && hostnumber==2";
```

mpi_openmp.bash

from LSB_HOSTS to nodefile.txt

From nodefile.txt to rankfile.txt

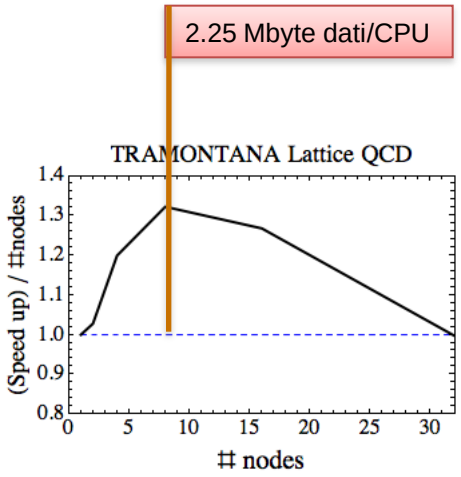
```
echo  $LSB_HOSTS | tr ' ' '\n' | sort -u > nodefile.txt
awk '{print "rank " i++ "=" $1 " slot=0-3" "\n" "rank " i++"="$1 " slot=4-7"}' nodefile.txt > rankfile.txt
RN=$[`cat rankfile.txt  |  wc --lines`]
mpirun -np $RN -x OMP_NUM_THREADS=4 --hostfile nodefile.txt -rankfile rankfile.txt  mpiomp_exec
```

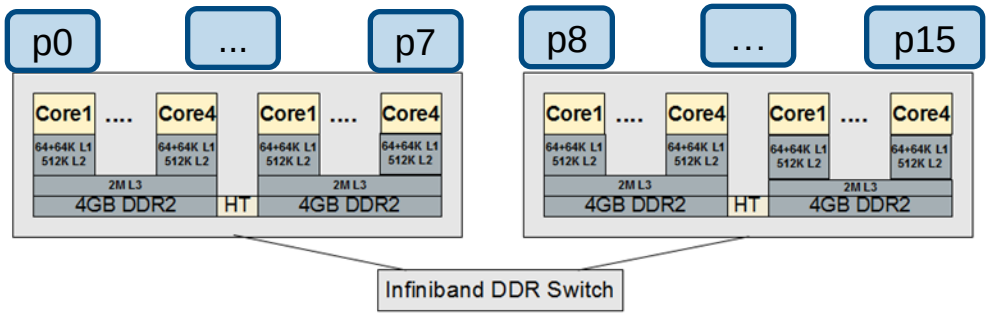
A real MPI application: "Choma library"

Hybrid-Montecarlo simulation of the Pure Gauge SU(3) on a 32x32x32x8 lattice (2000 sweep) using the publicly available **USQCD collaboration** "Chroma" library (http://usqcd.jlab.org/usqcd-docs/chroma/).

- **Pure MPI code**
- Total memory occupation of the grid **~36MBytes**
- Importance of **memory affinity** - when all the data are not in cache -
- Cache effect - efficiency >1 -

Np	**8** (1x8)	**16** (2x8)	**32** (4x8)	**64** (8x8)	**128** (16x8)
Non-ranked	295 min	146 min	62 min	27 min	14 min
Ranked	287 min	139 min	59 min	27 min	14 min

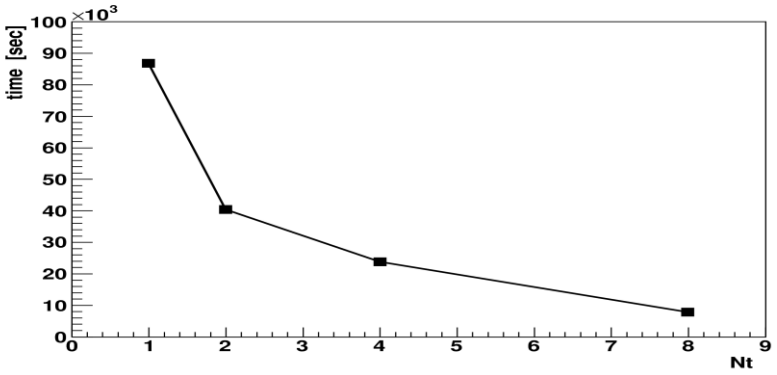




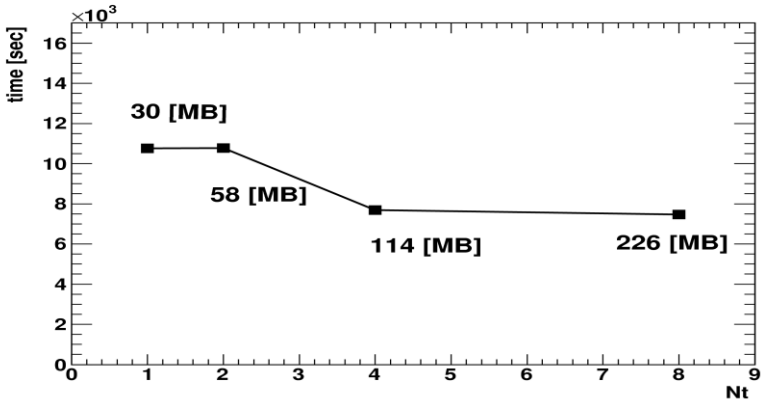
Thanks to A Feo (Turin Univ.)

"Numerical Stochastic Perturbation Theory"

Numerical Stochastic Perturbation Theory is introduced to perform high order perturbative calculations in lattice gauge theory, and numerically integrate the differential equations of Stochastic Perturbation Theory.
The code has been written in Parma, and it supports **openMP parallelization**.



Keep the volume fixed (16x16^3 ~ 226 MB) and increase number of threads.
The expected behaviour is: time ~ 1/Nt



Increase volume with number of threads according V = (Nt x 2) x 16^3.
The expected behaviour is: time = const

Thanks to M. Brambilla and M. Hasegawa (Parma Univ.)

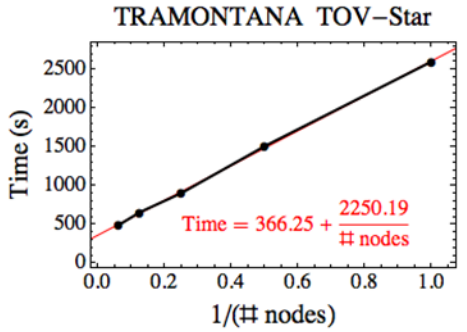
A real hybrid application: "Einstein Toolkit"

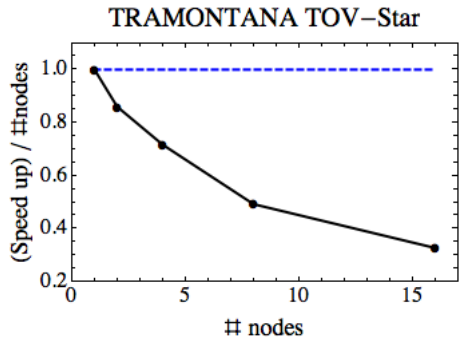
Evolution of a stable general relativistic TOV-Star using the **Einstein Toolkit consortium** codes (http://einsteintoolkit.org/).

Hydro-dynamical Simulation of a perfect fluid coupled to the full Einstein Equations (dynamical space-time) on a 3-dimensional grid with 5-level of refinement spanning an octant of radius 177 km with a maximum resolution within the star of 370 m.

- **Very complex code** (more than 100 developers over 20 years using F90,F77,C++,C)
- **Hybrid MPI-openMP** (not fully parallelized)
- Total memory occupation of the grid **~8GByte**.

#node	Np=8x#	Np=4x#	Np=2x#	Np=#	Np=2x#
	Nt=1	Nt=2	Nt=4	Nt=8	Nt=4 (rank)
1	2291.90	2934.21	3126.73	3360.96	**2608.08**
2	1438.72	1619.83	1797.30	2061.55	**1516.04**
4	1007.71	993.79	1007.71	1268.79	**909.36**
6	767.45	783.07	694.31	927.35	**745.63**
8	663.03	638.81	694.31	753.79	**661.37**
16	461.85	448.77	484.20	552.89	**497.78**





Thanks to R. De Pietri (Parma Univ.)

Conclusions and future work

The **TheoMPI** parallel cluster has been successfully installed and configured.

Grid is the unique access method.

Special features (waiting for support in EMI-1 release):

1) The new **"Granularity" attributes** are working with a temporary patch and syntax.
2) **Hybrid programming** (including openMP and CPU affinity) support is realized with direct job submission (without Mpi-Start).

Important **applications in Theoretical Physics**, using different **parallel approaches,** have been successfully executed on the cluster.

This model **will be extended to other MPI sites** in Theophys (and other interested VOs).

Thank you

for your attention!