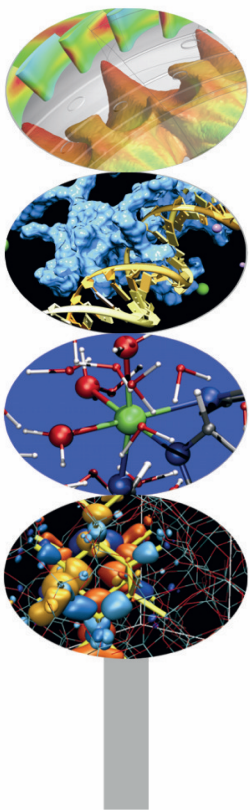
CINECA SCAI
SuperComputing Applications and Innovation

Cineca COURSES High Performance Computing 2015

HPC enabling of OpenFOAM® for CFD applications



HPC facilities: overview and how to use

Ivan Spisso, Giorgio Amati

25-27 March 2015, Casalecchio di Reno, BOLOGNA.

SuperComputing Applications and Innovation Department, CINECA

PRACE

CINECA

Table of Contents

What is CINECA

- Cineca is a non profit Consortium, made up of 70 Italian universities, 4 Italian Research Institutions and the Italian Ministry of Education.
- Today it is the largest Italian computing centre, one of the PRACE Tier-0 hosting site, one of the most important worldwide.
- With more 700 employees, it operates in the technological transfer sector through high performance scientific computing, the management and development of networks and web based services, and the development of complex information systems for treating large amounts of data.
- Cineca has three locations: Bologna, Milan and Rome

Mission: Cineca offers support to the research activities of the scientific community through supercomputing and its applications

SCAI department

- **SCAI**(**S**uper**C**omputing **A**pplications and **I**nnovation) is the High Performance Computing department of CINECA, the largest computing centre in Italy and one of the largest in Europe.
- The mission of SCAI is to accelerate the scientific discovery by providing high performance computing resources, data management and storage systems, tools and HPC services, and expertise at large
- aiming to develop and promote technical and scientific services related to high-performance computing for the Italian and European research community.
- CINECA enables world-class scientific research by operating and supporting leading-edge supercomputing technologies and by managing a state-of-the-art and effective environment for the different scientific communities.
- The SCAI staff offers support and consultancy in HPC tools and techniques and in several scientific domains, such as physics, particle physics, material sciences, chemistry, fluid dynamics

Table of Contents

HPC @ CINECA at glance

Cineca is currently one of the Large Scale Facilities in Europe and it is a PRACE Tier-0 hosting site.

http://www.hpc.cineca.it/content/hardware

1. FERMI (Tier-0): It is a IBM BG/Q supercomputer, classified among the most powerful supercomputers in the Top500 List: rank 7th in June 2012. On June 2012 it was ranked 11st in the Green500's energy-efficient supercomputers list.
   It will be replaced by a new Tier-0 system at the end of the 2015
2. GALILEO (Tier-1): it is a IBM NeXtScale cluster accelerated with Intel Phi's and (GPUs): in full production February the 2nd, 2015. It will be upgraded shortly with a bunch ($\simeq 100$) Nvidia K80 Gpu. It will reach around 100th position in the next top500 list.
3. PICO: BigData infrastructure has been recently acquired (Nov 2014) devoted to "Big Analytics".

FERMI

High-end system, devoted for extremely scalable applications

- IBM power2@1.6 Ghz
- 10240 computing nodes, 16 core each (163,840 total)
- 16 GB or RAM per computing node, 163 TB of Total RAM
- Proprietary Network (5D torus)
- Peak performance 2PFlops

GALILEO

x-86 based system for production of medium scalability applications

- Intel Xeon E5-2630 v3 @2.4 GHz (a.k.a Haswell)
- 516 computing nodes, 16 core each (8,256 total)
- 128 GB of RAM per computing node, 66 TB of Total RAM
- Infiniband QDR ($\simeq 40 Gb/s$)
- 768 Intel Phi 7120p (2 per node)
- Nvidia K80 ($O(100)$)
- 8 nodes devoted to login/visualization
- Peak performance 1.2 PFlops
- $\simeq$ 480 GFlops single node LINPACK (only CPU) sustained performance

PICO

Processing system for large volumes of data

- Intel Xeon E5 2670 v2 @2.5 GHz (a.k.a Ivy Bridge)
- 66 computing nodes, 20 core each (1320 total)
- 128 GB or RAM per computing node, 8.3 TB of Total RAM
- Infiniband FDR ($\simeq 56 Gb/s$)
- 4 nodes devoted to login/visualization
- Peak performance $\simeq 40 Tflops$
- $\simeq$ 400 GFlops single node LINPACK (only CPU) sustained performance

Storage

- Each system has:
  - a */home* area
  - a */scratch* area
  - a */work* area (project-based)
- a common */gss* area ($\simeq 3PB$)
  - shared by all login nodes
  - shared by all PICO computing nodes
  - Tape subsystem (LTFS, up to $\simeq 12PB$)

Table of Contents

How to use the resources

User Portal: http://www.hpc.cineca.it/content/users

It is organized in sections:

- Getting started: you have to register yourself to the UserDB portal to get a valid login, once you have an active project (see next talk) you can access to CINECA facilities
- Get in touch: register to the mailing list to be upgraded about the status of the machine
- Help desk:superc@cineca.it for any problem/help/support
- Documentation: User Guide, FAQ et al....

Table of Contents

OpenFOAM @ CINECA

Experience:

- OpenFOAM installed and tested on our clusters: GALILEO, FERMI, PICO
- Used in Several Academic project: 37 ISCRA + 10 LISA + 1 PRACE (under evaluation)
- OpenFOAM in FORTISSIMO for the Enabling Manufacturing SMEs to benefit from HPC and Digital Simulation.
  - Cloud-based Computational Fluid Dynamics Simulation in collaboration with Konigsegg, ICON, CINECA and NTUA. DES solvers for Drag and Lift prediction of supercars.
  - Shape Optimization under Uncertainty through HPC Clouds in collaboration with Optimad Eng, University of Strathclyde and Automobili Lamborghini (OF + Dakota)
- Support for industrial development: 14 projects.
- Analysis of *Current Bottlenecks in the Scalability of OpenFOAM on Massively Parallel Clusters* of OpenFOAM in the framework of PRACE (M. Culpo White Paper)

OpenFOAM Installation 1/2

OpenFOAM can be installed for many users (network installation) or for a single user (local installation):

- Network installation: This installation is suitable when a group of people is supposed to use OpenFOAM, and when not everyone want to learn how to install and compile it. All users will use exactly the same (base) installation.
  Pro: A single installation for each version of OpenFOAM, maintained by the CINECA UserSupport.
  Cons: limited to major release and most common used tools (`swak4foam`, `pyfoam`, `dakota`).
- Local installation: This is the most common way of installing OpenFOAM. The installation will be located in `HOME/OpenFOAM/OpenFOAM-3.x.y`.
  Pro: Each user will 'owns' his proper installation and may update it any time. For info on installation
  Cons: Requires extra disk space for several users with their own installations (minor issue), and all users have to know how to install OpenFOAM and the Third-Party products (major issue)

OpenFOAM Installation 2/2

CINECA policies:

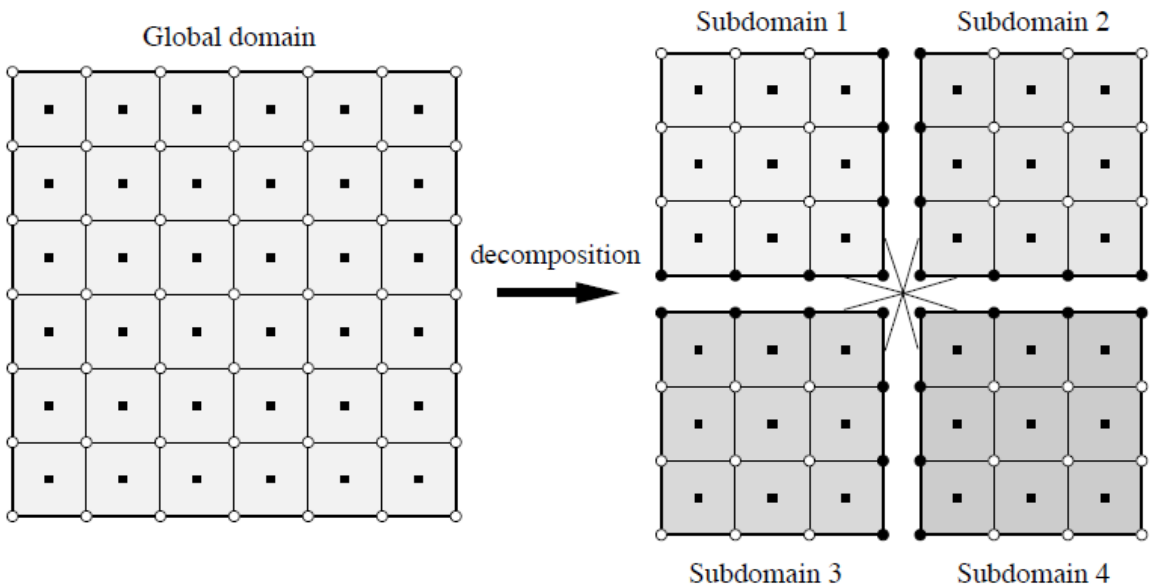
- Network installation only major 2.3.0, 2.4.0, . . . 2.n.0 by default.
- Minor installation 2.3.1, 2.3.2, . . . 2.3.4 upon request.
- `profile base` ⇒ last two majors + 1 minor.
- `profile advanced` ⇒ other versions
- Git and .x only local installation

Network Installation status on HPC platforms:

- FERMI ⇒ version 2.1.1 (no upgrade possible due to issues with bgq-gnu/4.7.2 compiler)
- GALILEO version 2.3.0 + swak4foam/0.3.1 + pyfoam/0.6.4 + (dakota 6.1 on-going)
- PICO only local installation up to now

Running in parallel

- The method of parallel computing used by `OpenFOAM` is based on the standard Message Passing Interface (MPI) using the strategy of domain decomposition.
- The geometry and the associated fields are broken into pieces and allocated to separate processors for solution.
- A convenient interface, `Pstream`, is used to plug any Message Passing Interface (MPI) library into OpenFOAM. It is a light wrapper around the selected MPI Interface



Performance and bottlenecks

An analysis has been done in the framework of PRACE 1IP to study the current bottlenecks in the scalability of OpenFOAM on Massively parallel clusters.

- OpenFOAM scales reasonably well up to thousands of cores, upper limit $\sim$ 1,000 cores.
- An in-depth profiling identified the calls to the `MPI_AllReduce` function in the linear algebra as core libraries as the main communication bottleneck
- A sub-optimal performance on-core is due the sparse matrices storage format that does not employ any cache blocking.

M. Culpo, Current Bottlenecks in the Scalability of OpenFOAM on Massively Parallel Clusters,

PRACE White Paper, available on-line at www.prace-ri.eu

Actual bottlenecks

Missing for a full enabling on Tier-0 Architecture:

- Improve the parallelism paradigm, to be able to scale from the actual $\sim$ 1,000 procs to at leat one order of magnitude ($\sim$ 10,000 or 100,000 procs).
- improve the I/O, which is a bottleneck for big simulation. For example LES/DNS with hundreds of procs that requires very often saving on disk.

Table of Contents

Best Practices

Tune your application on HPC enviroment

- strong scaling $\Longrightarrow$ how the solution time varies with the number of processors for a fixed total problem size
- The performance results vary depending on different parameters including the nature of the tests, the solver chosen, the number of cells per processors, the class of cluster used, choice of MPI distributions, etc
- Rule of the thumb: tests cases up to tenths of millions of cells scales well with orders of thousands of cores
- try to reduce the communication time when running your test-cases
- increase (as much as possible) the number of cells per processors to find out the optimal number for your application in the selected cluster
- check the memory usage. Now you have a lot of memory per node, try to use it