# The I/O benchmarking methodology of EoCoE project

S. Lührs[1], A. Funel[2]
M. Haefele[3], F. Ambrosino[2], G. Bracco[2], M. Celino[2], A. Colavincenzo[4], S. Giusepponi[2], G. Guarnieri[2], M. Gusso[2], G. Ponti[2]

1- JÜLICH SUPERCOMPUTING CENTRE– GERMANY
s.luehrs@fz-juelich.de

2 - ENEA – ITALY
{fiorenzo.ambrosino, giovanni.bracco, massimo.celino, agostino.funel}@enea.it
{simone.giusepponi, guido.guarnieri, michele.gusso, giovanni.ponti}@enea.it

3 - MAISON DE LA SIMULATION – FRANCE
matthieu.haefele@maisondelasimulation.fr

4 - KELYON S.R.L. – ITALY
antonio.colavincenzo@kelyon.it

## The EoCoE project

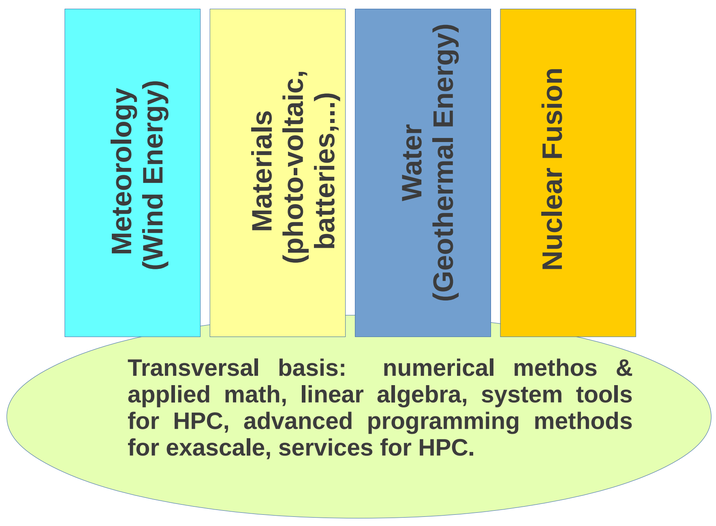
The Energy oriented Centre of Excellence in computing applications (EoCoE) [1] is one of the 8 centres of excellence in computing applications established within the Horizon 2020 programme of the European Commission. The primary goal of EoCoE is to foster and accelerate the European transition to a reliable and low carbon energy supply by giving high qualified HPC support to research on renewable energy sources. EoCoE project aims to create a new community of HPC experts and scientists working together to achieve advances on renewable energies. The project started on October 1st 2015 and will run until September 30th 2018.

## Structure of the project

EoCoE project is composed of four pillars: a) Meteorology; b) Materials modelling to design efficient low cost devices for energy generation and storage; c) Water; d) Nuclear Fusion; and a transversal basis. The objective of the transversal basis is to overcome bottlenecks in application codes from the pillars.

It develops cutting-edge mathematical and numerical methods, and benchmarking tools to optimize application performances on many available HPC platforms. The transversal basis gives support in the following fields:

- Numerical Methods & Applied Mathematics
- Linear Algebra
- System Tools for HPC
- Advanced Programming Methods for Exascale
- Tools and Services for HPC




## Consortium

EoCoE is structured around a central Franco-German hub coordinating a pan-European network, gathering a total of eight countries and twenty-two partners. Coordinator: Maison de la Simulation (France) [2].

http://www.eocoe.eu/about-us

## Services

EoCoE provides services to academy and industry

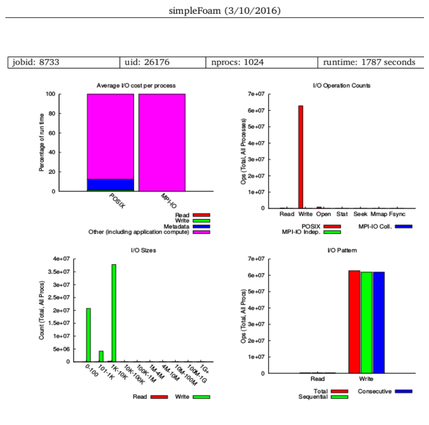
- code auditing, optimization and enhancements
- HPC training and consultancy.

One of the issues common to all activities of the transversal basis is I/O benchmarking. Performing efficient I/O for very large datasets on current supercomputers is already challenging and will become more challenging for the next supercomputer generations. We present the methodology adopted by EoCoE to detect and remove I/O bottlenecks.

## I/O benchmarking of codes

Efficient I/O is crucial for a fast execution of a code. It could happen that a code run fast on a HPC platform and slow on another even if the two systems have the same microprocessors. Often the bottleneck is due to poor I/O performance. There are many causes which may degrade I/O performance: inefficient I/O strategy, inappropriate size of I/O buffers, large number of I/O calls, poor file system performance etc. It is very important to use benchmarks to get I/O behaviour in a reproducible way to identify bottlenecks. The adopted methodology is outlined below:

- find I/O pattern at runtime (Darshan) [3]
- simulate the I/O pattern on many HPC systems to find bottlenecks (IOR) [4]
- remove bottlenecks:
  - code changes, optimized libraries
  - code refactoring and integration of optimized libraries




File Count Summary
(estimated by I/O access offsets)

| type | number of files | avg. size | max size |
|---|---|---|---|
| total opened | 829446 | 334K | 1.4M |
| read-only files | 10246 | 191K | 1.4M |
| write-only files | 819200 | 336K | 1.1M |
| read/write files | 0 | 0 | 0 |
| created files | 819200 | 336K | 1.1M |

Average I/O per process

| | Cumulative time spent in I/O functions (seconds) | Amount of I/O (MB) |
|---|---|---|
| Independent reads | 2.210158 | 1.894863 |
| Independent writes | 22.705621 | 261.994389 |
| Independent metadata | 196.660971 | N/A |
| Shared reads | 0.000123 | 0.000008 |
| Shared writes | 0.000000 | 0.000000 |
| Shared metadata | 0.029269 | N/A |

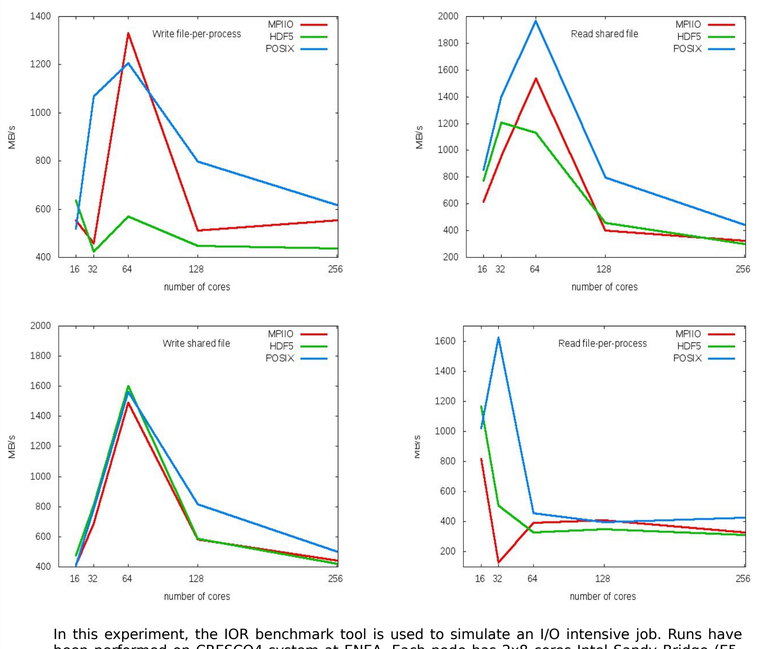
Data Transfer Per Filesystem

| File System | Write | | Read | |
|---|---|---|---|---|
| | MiB | Ratio | MiB | Ratio |
| /gporq1.1M | 268282.25418 | 1.00000 | 1940.34817 | 1.00000 |

An example of I/O pattern captured by Darshan for a CFD (external aerodynamics) simulation on CRESCO4 [5] supercomputer at ENEA [6]. CRESCO4 system has ~300 nodes each of which with 2x8 cores Intel Sandy Bridge (E5-2670 2.6 GHz) 64 GB RAM. The simulation uses 1024 cores and a GPFS file system with 6 I/O servers over IB 4xQDR (40 Gbps) network. The run execution time is ~1420 s. This example shows that the huge number of I/O accesses (~830000) and the small I/O size (~KB) are hints for a bottleneck.

## Scalability analysis of parallel I/O libraries by means of simulations

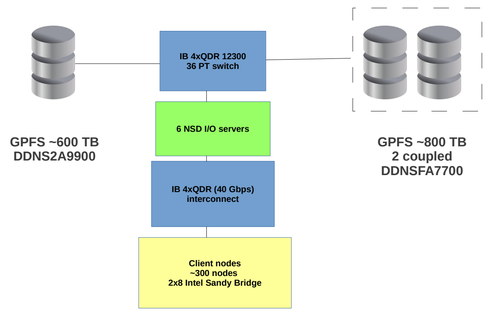
The study of I/O scalability of common used libraries (MPI, HDF5, PnetCDF, POSIX) provides hints to users on how to setup their codes. Sometimes, the best number of tasks for I/O management and/or the I/O strategy (master-slave, single shared file, one file per task etc.) can only be found experimentally.

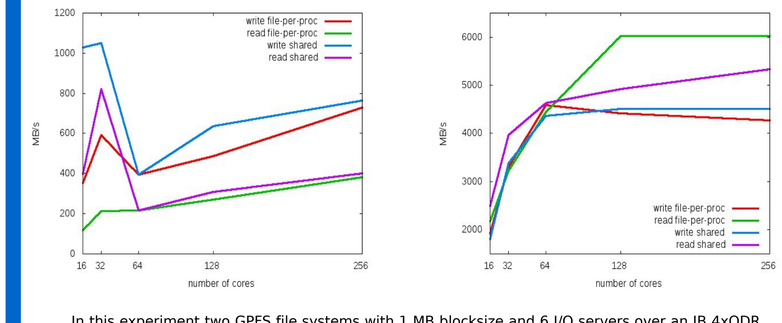



In this experiment, the IOR benchmark tool is used to simulate an I/O intensive job. Runs have been performed on CRESCO4 system at ENEA. Each node has 2x8 cores Intel Sandy Bridge (E5-2670 2.6 GHz) 64 GB RAM. It has been used a GPFS file system with 6 I/O servers over IB 4xQDR (40 Gbps) network. Each core is assigned an I/O task. Results show that for the three tested interfaces (MPIIO, HDF5, POSIX) the best configuration is with 64 tasks reading/writing a common shared file.

## I/O performance of disk storage systems

The purpose of this experiment is to measure the efficiency of two different disk storage systems under the same heavy I/O workflow by using IOR benchmark. The efficiency of a system is obtained by measuring how much the I/O performance differs from its maximum peak. Efficiency gives an idea on how a storage technology is evolving.




In this experiment two GPFS file systems with 1 MB blocksize and 6 I/O servers over an IB 4xQDR (40 Gbps) are used to access two storage systems: (A) a DDNS2A9900 of ~600 TB and (B) two coupled DDNSFA7700 hosting ~800 TB. The maximun available I/O throughput is ~6 GB/s and ~18 GB/s for (A) and (B) respectively. Each I/O client node has 16 cores Intel Sandy Bridge (E5-2670 2.6 GHz). To simulate a heavy workload each core executes an I/O task which reads/writes 1 GB and in this situation the 16 tasks on each client share the bandwidth of the network card. Results show that in the case of full load the efficiency is ~8% for system (A) and ~30% for (B).

[1]http://www.eocoe.eu/ [2]http://www.maisondelasimulation.fr/ [3]http://www.mcs.anl.gov/research/projects/darshan/ [4]https://github.com/LLNL/ior [5]http://www.cresco.enea.it/ [6]http://www.enea.it/it [7]http://www.fz-juelich.de/ias/jsc/EN/Home/home_node.html [8]http://www.man.poznan.pl/online/en/