Parallel simulations produce large amount of data. The traditional approach consists in writing these data to disk to later on read them back from disk and analyse them. But this approach is becoming prohibitively slow. An alternative consists in analyzing the data online, often called in situ, as soon as provided by the simulation, before to write the data to disk. Thus the data are analyzed in parallel and reduced in size before to go to disk, enabling to significantly reduce the disk performance bottleneck. The DataMove team has addressed this issue through different approaches leading to the FlowVR and Melissa frameworks. The data produced by the simulation can also be seen as a stream of data: each process of the simulation (a parallel simulation can run thousands of processes) produces a new results at each time step (a numerical simulation can classically computes thousands of time steps). The DataMove team has also investigated how stream processing approaches developed for Big Data processing (Flink in particular) could be used in this context. These different works provide DataMove with a sound base for investigating an even more challenging data analysis process called Data Assimilation. This is the topic to be investigated through this PhD.

Data assimilation consists in using observation data (measures acquired from other scientific instruments or network of sensors) to adjust the internal parameters of the running simulation. The goal is to better control the course of the running simulation to get higher quality results as well as to limit the required number of simulation executions, thus saving CPU energy and scientist time. Data assimilation requires to analyse the outputs from the simulation, them to combine these results with the observation data to compute adjustment values for the internal simulation parameters. This process takes place periodically while the simulation is running. Often the correlation between the outputs and the internal parameters is a complex inverse problem requiring advanced algorithms. Today solutions rely on optimization methods like gradient descent to minimise some objective function.

Because the simulation outputs are analyzed to compute and retro-propagate parameter adjustments on the running simulation, a careful assembly and parallelization of the different steps are required. Modularity ans separation of
concerns is also critical to keep this complex workflow modular, flexible and amenable to modifications of its individual components without having to master all the different components. But today’s solutions fail to reach this scalability and modularity level and rely on a single monolithic application.

If this workflow was not challenging enough, data assimilation can also require to handle not one single simulation execution but several simulation runs in parallel (often called an ensemble run). When working with a single simulation, the data assimilation process is called variational, while ensemble run approaches are called statistical. Variational approaches intend to minimize an error function on a single simulation. Statistical approaches look at minimizing a variance. They require to run several simulations with different parameter sets enabling to provide a more complete information about the uncertainties associated with the results.

Main activities

The goal of this PhD work is to investigate solutions to push data assimilation capabilities to novel scalability and performance levels, using thousands of compute cores if necessary. We will cover aspects like infrastructure (leveraging the one like Melissa and FlowVR developed at DataMove as well as other frameworks like Flink), fault tolerance, parallelization, exploration of novel data assimilation processes based on unsupervised learning algorithms for instance.

To achieve this goal we will work in close with scientists from other domains, learning specialists, as well application domains experts in particular from Hydrology or Molecular Dynamics. They will provide us the main use cases we will work with (simulation codes, observational data and data assimilation processes).

Skills

We are looking for enthusiastic, motivated and creative candidates, capable of developing novel innovative approaches. The PhD candidate will have a master in computer science with expertise in one or several of the following domains: parallel, distributed computing, numerical simulation, large scale data analytics, optimization technics. Good programming skills and a Linux expertise are required, with a taste for leading experiments on large supercomputers. The candidate will have opportunities to travel in France and abroad to present its results, in particular at some of the major international conferences of the domain. A good speaking and writing English level is sufficient in the working environment, but some bases or at least the will to learn French will greatly help for the every day life. The candidate will integrate the DataMove team and benefit from its friendly work environment. Grenoble valley is known to offer a good quality of life in the middle of the French Alps.

Bibliography

Benefits package
- Subsidised catering service
- Partially-reimbursed public transport
- Social security
- Paid leave
- Flexible working hours

Remuneration

Monthly salary after taxes: around 1596,05€ for 1st and 2nd year. 1678,99€ for 3rd year. (medical insurance included).