In-situ Analysis and Visualization of Ensemble Data Sets: the case of Fluid Dynamics Simulations in Energy Production

Alejandro Ribés, EDF R&D

Context

Computational Fluid Dynamics (CFD) is a fundamental step for the study and optimization of electricity production. Indeed, current power plants use water as a mean of convective heat transfer. Consequently, the simulation, analysis and visualization of fluid dynamics phenomena are of great importance for the energy industry. *Electricite de France* (EDF), being one of the biggest electricity producer in Europe, has developed for the past 15 years an open source CFD code named *Code_Saturne*, allowing for the solution of very large models [1]. EDF has several supercomputers that regularly run this code in order to perform analysis involving large amounts of data.

Motivation

Numerical simulations are producing an increasingly larger volume of data to be analyzed and visualized; this is a general tendency in the world of numerical simulation. Furthermore, a new trend appears in this context: the parametric studies. When a parametric study is conducted, the simulation is ran N times with one or several varying parameters. The result of such a family of runs is called *ensemble data set*, and each individual run is called a *member*. Ensembles are very large because the already large amount of data produced in a single simulation is multiplied by N. Ensembles are also multidimensional, multivariate and multivalued [2]. The main challenges in using ensembles stem from the size and complexity of the data.

Parametric studies are becoming increasingly popular in industrial environments. EDF R&D has already developed software infrastructure to perform them, such as OpenTurns [3] and the *Parametric* module of SALOME plate-form [4]. Among others, a sensitivity study of a parameter or a simple visual comparison (of different views of a varying parameter) is usually performed. However, solvers such as Code_Saturne are currently being optimized for performing numerical simulations on Petascale systems [1], this implies that massive ensembles will be generated in the future. Code_Saturne has already been reported to deal with meshes of 1,000 million cells; for this case, we can obtain an estimate ensemble size with the hypothesis that the solver is (using the nomenclature of [2]):

- Multidimensional, we consider a 3D mesh with 1,000 million cells and 200 time steps
- Multivariate, we consider calculating 10 result vector fields (with 3 components per field element)
- Multivalued, we consider a simple parametric study with N=100 runs

In this example, and for a coding in 2 bytes, we obtain an ensemble size s = 1,000,000,000 x 200 x 10 x 3 x 100 x 2 = 1,200 To! Indeed, we obtain 12 To per simulation run. It could be argued that such 1,000 million cells mesh parametric study has never been performed. If we simply consider a mesh 100 times smaller, 10 millions cells, we realize that this case is, at the moment, easily found in industrial environments; in this case N=100 runs would generate an ensemble of 12 To. Thus, the a posteriori traversal of huge ensembles to perform simple summary statistics would be, currently, extremely time-consuming or just not possible. This is the main reason that motivated some preliminary work, at EDF R&D, on in-situ analysis of ensembles.

The challenge

One of the current challenges of the visualization community is how to deal with the multivalued nature of the ensembles, [2], [5], [6]. A solution consists in eliminating part of the complexity by choosing one variable (eliminate multivariavility) and one point in the simulation (eliminate spatial multidimensionality). This solution leads to a reduced dataset where only time and the multiple values of a parameter are kept. This is indeed a very common solution in industrial parametric studies. In Figure 1 we show 600 Monte Carlo simulations of temperature in function of time in a thermal-hydraulical transient parametrical simulation, which represented a large-break loss of primary coolant accident in a power plant. Indeed the application scientist can study the lost of temperature by exploring this set of curves but they are focusing in a specific location in the simulation domain and in a specific variable.

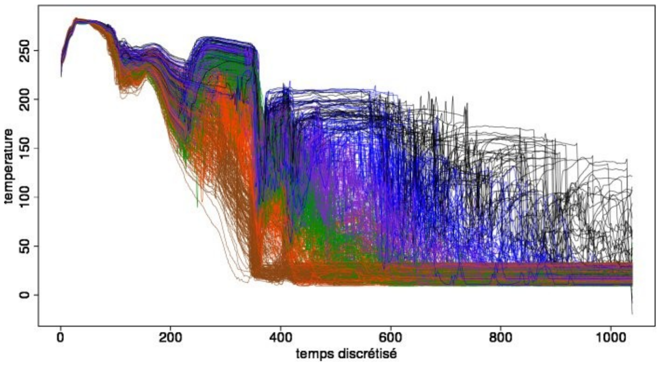

Figure 1. 600 Monte Carlo simulations of thermal-hydraulical transients.

Studying the complete ensemble is a complex task; this is why the output of our work should be a result of an interaction between HPC (High Performance Computers) engineers, visualization experts and CFD scientist. Indeed, we have performed some preliminary work where an in-situ approach is used. This allows exploring the ensemble: N Code_Saturne simulations are ran at the same time in a supercomputer, and they communicate in order to calculate summary statistics. The resulting statistics are defined over all dimensions (space and time) and for all variables! Furthermore, the summary statistics are available when the parametric study is finished.

An important point not to be forgotten is that ensembles are expensive to store. An in-situ approach can choose not storing the ensemble but just the extracted pertinent information that will be later analyzed. In this case, the storage needed for the in-situ outputs can be of the same order of a single run instead of N times a run.

The challenge will be to propose generic HPC mechanisms to perform in-situ statistical analysis of ensemble data.

Bibliography :

[1] Y. Fournier, J. Bonelle, C. Moulinec, Z. Shang, A.G. Sunderland and J.C. Uribe, Optimizing Code_Saturne computations on Petascale systems, Computers & Fluids, Vol. 45, Issue 1, June 2011, pp. 103–108.

[2] A.L. Love, A. Pang, D.L. Kao. "Visualizing spatial multivalue data", IEEE Computer Graphics and Applications, Vol. 25, no. 3 (2005): 69-79.

[3] EDF, EADS France and PHIMECA Engineering. “OpenTRUNS.” Internet: http://www.openturns.org, [May. 20, 2014].

[4] EDF R&D. “Salome.” Internet: http://www.salome-platform.org/, [May. 20, 2014].

[5] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. Moorhead. Noodles: A tool for visualization of numerical weather model ensemble uncertainty. IEEE Transactions on Visualization and Computer Graphics, 16(6):1421–1430, 2010.

[6] K. Potter, A. Wilson, V. Pascucci, D. Williams, C. Doutriaux, P.-T. Bremer, and C. Johnson. Ensemble-vis: A framework for the statistical visualization of ensemble data. In Data Mining Workshops, 2009. ICDMW ’09. IEEE International Conference on, pages 233–240, 2009.