The last few years have been a fertile ground for the development of many scientific and data-intensive applications in all fields of science and industry. These applications provide an indispensable mean of understanding and solving complex problems through simulation and data analysis. As large-scale systems evolve towards post-Petascale computing to accommodate applications increasing demands for computational capabilities, many new challenges need to be faced, among which fault tolerance is a crucial one. With failure rates predicted in the order of tens of minutes for the exascale era and applications running for extended periods of time over a large number of nodes, an assumption about complete reliability is highly unrealistic. Because processes from scientific applications are, in general, highly coupled, even more pressure is put on the fault tolerance protocol since a failure to one of the processes could eventually lead to the crash of the entire application.

The research I have been involved in during the last four years focuses on offering ways of reducing the overhead induced by fault tolerance strategies, by combining them with failure avoidance methods. Failure avoidance is based on a prediction model that detects fault occurrences ahead of time and allows preventive measures to be taken, such as task migration or checkpointing the application. My key observation is that errors are often predicted by changes in the frequency or regularity of various events. For this purpose, I have investigated the linkage between signal processing concepts and data mining techniques in the context of failure analysis for large-scale systems. By shaping the normal and faulty behaviour of each event, and of the whole system, I was able to propose appropriate models and methods for descriptive and forecasting purposes. I have made multiple experiments on different production HPC systems, from Argonne’s Blue Gene systems, to NCSA’s Mercury and Blue Waters and Tokyo Institute of Technology‘s Tsubame2. The preliminary results show that conventional signal processing techniques can create clear markers for changes in events behavior. Moreover, machine learning techniques become much more efficient when applied to the derived markers, rather than to the original signal. Consequently, I have worked with the developer of FTI (multi level checkpointing strategy) and created the first hybrid fault tolerance implementation that combines a proactive with a preventive checkpoint strategy based on the signal analysis predictor. Our method improves the performance of classical fault tolerance techniques when dealing with failures in Petascale systems and the results show the potential of using such an approach on future Exascale systems.