Title

Model-based System Identification Cloud (MbSIC) for Large-scale Data Analytics

Authors

Hisashi Miyashita[*], Yoichi Hatsutori[*], Junya Shimizu[*], and Yoshiyuki Yamada

(*) IBM Research – Tokyo {himi, yoichi, jshimizu}@jp.ibm.com,

(+) Department of Physics, Kyoto University yamada@amesh.org

Abstract

In modern research experiments and field test engineering, we need to cope with large-scale observations and complex models to understand and utilize the results especially in areas such as astronomy, remote sensing, automotive companies, and natural resource industries.

In the state of the art when analyzing large volumes of data, there are two approaches: (a) developing analytics program by using general-purpose computing platforms such as Hadoop, Giraph, or hBase or (b) using general-purpose data analytics packages such as SPSS and ILOG. However, we face challenging issues in data analytics development and efficiency since Approach (a) requires deep programming skills for optimized performance and many scientists do not understand the details of such programming, which may lead to erroneous results, while Approach (b) often lacks the flexibility to represent the models that scientists would like to use in their experiments, such as motion equations, state functions, mechanical models, or engineering details.

Our approach seeks to address these problems with higher level languages, “systems model”, rather than using only implementation-specific programming language such as C, C++, and Java, for scientists and engineers to use in communicating with each other in describing their target problems. Throughout the project lifecycle of the mission requirements, operations, and maintenance, we create and make use of such models not only for data analytics purpose. For this vision, we propose two kinds of technologies: (1) by leveraging the model compiler technologies and cloud computing technologies, we will create a Model-based System Identification Cloud that accepts models of users’ target problems as inputs, compiles them into optimized code, and then execute them on highly scalable cloud computing platforms based on MapReduce and GPGPUs; and (2) a novel noise estimation technology for extracting and identifying colored noise, which is hidden within white Gaussian noise, so that apparently noisy observations can provide statistically relevant results with least-squares methods, thereby effectively bridging the gaps between the observations and models used for the identification.

We are developing this Model-based System Identification Cloud (MbSIC) to apply these technologies to specialized satellite-based astrometry problems. The systems models will represent how stars are observed from satellites orbiting the earth and allow us to estimate the distances, positions, and proper motions of stars by using a nonlinear least-squares method. We executed the analysis with Hadoop and GPGPUs and analyzed the parameters for randomly distributed stars with orbits and attitudes of a simulated satellite.

We believe MbSIC can be applied to many large-scale data analytics problems that use least-squares methods, which include remote sensing, automatic identification system, and asset management, and it will provide precise and efficient model-based systems simulations and analyses by helping us to construct precise plant models from the observations.